H-STATISTIC WITH WINSORIZED MODIFIED ONE-STEP *M*-ESTIMATOR AS CENTRAL TENDENCY MEASURE

ONG GIE XAO

MASTER OF SCIENCE (STATISTICS)
UNIVERSITI UTARA MALAYSIA
2017

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)

ONG GIE XAO

calon untuk Ijazah — MASTER
(candidate for the degree of)

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

"H STATISTIC WITH WINSORIZED MODIFIED ONE - STEP M-ESTIMATOR AS 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 : ***10 Oktober, 2016.***
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:
October 10, 2016.

Pengerusi Viva: *(Chairman for VIVA)*	Assoc. Prof. Dr. Suzilah Ismail	Tandatangan *(Signature)*
Pemeriksa Luar: *(External Examiner)*	Dr. Muzirah Musa	Tandatangan *(Signature)*
Pemeriksa Dalam: *(Internal Examiner)*	Dr. Shamshuritawati Sharif	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. Suhaida Abdullah	Tandatangan *(Signature)*

Tarikh:
(Date) ***October 10, 2016***

Permission to Use

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

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

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

Abstrak

Ujian *t*-dua sampel bebas dan *ANOVA* adalah kaedah klasik yang masing-masing digunakan secara meluas untuk menguji kesamaan dua kumpulan dan lebih daripada dua kumpulan. Walau bagaimanapun, kaedah berparameter ini mudah dipengaruhi oleh ketidak kenormalan, lebih ketara lagi apabila wujud varians yang heterogen dan saiz sampel yang tidak seimbang. Sebagaimana yang diketahui umum, pelanggaran dalam andaian ujian ini akan menyebabkan peningkatan dalam Ralat jenis I dan kemorosotan dalam kuasa ujian. Kaedah tidak berparameter seperti Mann-Whitney dan Kruskal-Wallis adalah merupakan alternatif kepada kaedah berparameter, namun, kehilangan maklumat berlaku disebabkan oleh data berpangkat. Bagi meringankan masalah ini, kaedah teguh boleh digunakan sebagai alternatif lain. Salah satu daripada kaedah tersebut adalah *H*-statistik. Apabila digunakan dengan penganggar *M*-satu langkah terubahsuai (*MOM*), statistik ujian ini (*MOM-H*) dapat menghasilkan kawalan Ralat jenis I yang baik walaupun dalam keadaan saiz sampel yang kecil, tetapi tidak konsisten pada beberapa keadaan yang dikaji. Tambahan pula, kuasa ujian adalah rendah yang berkemungkinan disebabkan oleh proses pangkasan data. Dalam kajian ini, *MOM* diwinsor (*WMOM*) bagi mengekalkan saiz sampel asal data. *H*-statistik apabila digabungkan dengan *WMOM* sebagai sukatan kecenderungan memusat (*WMOM-H*) telah menunjukkan kawalan Ralat jenis I yang lebih baik berbanding dengan *MOM-H* terutamanya di bawah rekabentuk seimbang walaupun dalam apa saja bentuk taburan. Ia juga menunjukkan prestasi yang baik di bawah taburan yang amat pencong dan berhujung berat bagi rekabentuk yang tidak seimbang. Di samping itu, *WMOM-H* juga mampu menjana kuasa yang lebih baik berbanding dengan *MOM-H* dan *ANOVA* di bawah kebanyakan keadaan yang dikaji. *WMOM-H* juga didapati dapat mengawal Ralat jenis I dengan lebih baik tanpa nilai liberal (>0.075) berbanding dengan kaedah berparameter (*t*-dua sampel bebas dan *ANOVA*) dan tidak berparameter (Mann-Whitney dan Kruskal-Wallis). Secara umum, kajian ini menunjukkan bahawa proses winsor (*WMOM*) boleh meningkatkan prestasi *H*-statistik dari segi kawalan Ralat jenis I dan meningkatkan kuasa ujian.

Kata kunci: Winsor, Ralat jenis I, Kuasa Ujian, Kaedah Teguh, *H-s*tatistik

Abstract

Two-sample independent *t*-test and *ANOVA* are classical procedures which are widely used to test the equality of two groups and more than two groups respectively. However, these parametric procedures are easily affected by non-normality, becoming more obvious when heterogeneity of variances and unbalanced group sizes exist. It is well known that the violation in the assumption of the tests will lead to inflation in Type I error rate and decreasing in the power of test. Nonparametric procedures like Mann-Whitney and Kruskal-Wallis may be the alternative to the parametric procedures, however, loss of information occur due to the ranking data. In mitigating these problems, robust procedures can be used as the other alternative. One of the procedures is *H*-statistic. When used with modified one-step *M*-estimator (*MOM*), the test statistic (*MOM-H*) produces good control of Type I error rate even under small sample size but inconsistent under certain conditions investigated. Furthermore, power of test is low which might be due to the trimming process. In this study, *MOM* was winsorized (*WMOM*) to retain the original sample size. The *H*-statistic when combines with *WMOM* as the central tendency measure (*WMOM-H*) shows better control of Type I error rate as compared to *MOM-H* especially under balanced design regardless of the shape of distributions. It also performs well under highly skewed and heavy tailed distribution for unbalanced design. On top of that, *WMOM-H* also generates better power value, as compared to *MOM-H* and *ANOVA* under most of the conditions investigated. *WMOM-H* also has better control of Type I error rates with no liberal value (>0.075) compared to the parametric (*t*-test and *ANOVA*) and nonparametric (Mann-Whitney and Kruskal-Wallis) procedures. In general, this study demonstrates that winsorization process (*WMOM*) is able to improve the performance of *H*-statistic in terms of controlling Type I error rate and increasing power of test.

Keywords: Winsorization, Type I error rate, Statistical Test Power, Robust Statistics, *H*-statistic

Acknowledgment

Firstly, I would like to express my highest appreciation to Universiti Utara Malaysia for providing me a chance to pursue my postgraduate degree in Master of Science (Statistics). I had been gained lots of precious experiences along this journey in many aspects, helped me to understand my strength and weakness which needed to continuously polished and improved.

I would like to express my eternal gratitude to my supervisor Prof. Dr. Sharipah Soaad Syed Yahaya for her patience guidance, encouragement and continuous supports on my research. She has been most helpful throughout my entire study. I also would like to express my gratitude to my co-supervisor, Dr. Suhaida Abdullah who also provides guidance and advice on my works. I could not have imagined having better advisors and supporters for my master study in University Utara Malaysia. Besides, I also would like to extend my gratitude to Prof. Dr. Abdul Rahman for his help in the programming used in the study.

Last but not least, an honourable mention goes to my families, course mates and friends for their spiritually supports, encouragements and understanding throughout my study periods. Without any helps and encouragements from the particular that I mention above, I would face with lots of challenges on completed my research.

Table of Contents

List of Tables

List of Figures

List of Appendices

APPENDIX A

List of Abbreviations

ANOVA	Analysis of Variance
H-statistic	Robust test to measure the equality of central tendency Measure
MOM	Modified One-step M-estimator
WMOM	Winsorized Modified One-Step M-estimator
SAS	Statistical Analysis Software
SAS/IML	Statistical Analysis Software/ Interactive Matrix Language

CHAPTER ONE

INTRODUCTION

1.1 Background

In recent years, procedures for testing the equality of central tendency (location) measures or locating group effects has been studied and improved. The main purpose of this continuous improvement is to get a procedure that can perform well in controlling Type I error rate, simultaneously increasing power to detect the effects. It is well known that distribution of data and the variance among treatment groups are one of main concern for parametric procedures such as *t*-test and analysis of variance (*ANOVA*). In order to use these procedures, assumptions such that the data must be normally distributed and the variances must be homogeneous have to be fulfilled. Any deviation from these two assumptions will cause Type I error rate to be inflated and depressed in power rate (Keselman, Algina, Lix, Wilcox, & Deering, 2008; Syed Yahaya, 2005; Syed Yahaya, Othman, & Keselman, 2006). As a consequence, the null hypothesis will be falsely rejected and the effect of the procedures will go undetected. In real world, data that we get can hardly fulfill the assumptions needed by the parametric procedures.

Conventionally, nonparametric procedures such as Mann-Whitney and Kruskal-Wallis are the common alternatives when data fail to fulfill the assumptions of parametric procedures. However, the nonparametric procedures are more appropriate for weak measurement scale data and larger sample size is needed to reject a false hypothesis due to low power as compared to parametric procedures (Md Yusof, Abdullah, & Syed Yahaya, 2012a). Moreover, lesser information could be captured when using nonparametric procedures because the procedures are based on ranking instead of the data measurement (Siegel, 1957). Taking into consideration the weaknesses of the parametric and nonparametric procedures, this study will embark on the search of a suitable procedure that can overcome the weaknesses of the aforementioned procedures. Existing in between the parametric and nonparametric procedures are the robust procedures which are not only flexible to assumptions, but are also powerful. These procedures also work well regardless of the size of the data, by controlling Type I error although the sample size is small. One of the robust location measures is modified one-step M-estimator (MOM), which provide good control of Type I error rate even in small sample size (Wilcox & Keselman, 2003b). Thus, robust procedures are the better alternatives when dealing with small data.

1.2 Robust Statistics

As there are weaknesses on both parametric and nonparametric procedure, another alternative which is gaining acceptance is robust statistics. According to Wilcox (1997, 2012), robust statistics can have good control in Type I error rate and maintain the power rate although the data set is non-normal or even heteroscedasticity exist. Theory of robustness was being developed by Huber (1964) and Hampel (1968) in the 1960's as a solution to overcome the weakness in statistical procedure. In robust approach, there are also assumptions on the data distribution but the assumption does not always need to be fulfilled. Robust statistics can withstand the violation of parametric assumptions, by which it will perform as well as possible if the assumption is met, however, will not perform worse even if the assumption is slightly violated (Syed Yahaya, 2005). As mentioned by Huber (1981), robustness signifies insensitivity to small deviation from parametric procedure assumptions and robust procedure usually adopts what might be called an “applied parametric viewpoint”, such that for a parametric procedure, it is hopefully to have a good approximation to the true underlying situation, but we cannot and do not assume that it is exactly correct.

There are several robust procedures that have been proven to be able to deal with non-normal distributed data and variance heterogeneity. Welch test is capable in handling the problem of variance heterogeneity (Welch, 1951). Othman, Keselman, Padmanabhan, Wilcox, and Fradette (2004) demonstrated that when H-statistic is made robust by replacing its location measure with modified one-step M-estimator (*MOM*), the proposed procedure known as *MOM-H* showed good controlled in Type I error rate. MOM is a central tendency measure that apply trimming approach which eliminate the value of the tails from a set of data through the trimming criterion and it is an approach that able to deal with non-normality. Apart from trimming, another approach in dealing with non-normality is winsorizing, whereby through this approach, the original sample size is preserved by replaceing the tail of the data, rather than eliminate them. In section 1.3, 1.4 and 1.5, trimming, winsorizing and *MOM-H* will be discussed further.

1.3 Trimming

In the robust development process, trimming is being recommended as one of the approaches to deal with non-normality (Wilcox & Keselman, 2003a). One of the estimators generated from the process is trimmed mean. Trimmed mean is the average value of the remaining data after the trimming process (data on the left and the right tails being eliminated based on the trimming criteria). Generally, there are two types of trimming; symmetric and asymmetric trimming. Symmetric trimming is a classical trimming procedure that data on both side of the tails are equally trimmed based on a predetermined percentage value. However, there is an issue regarding the percentage of trimming when using this approach. Different researchers have proposed different percentages for trimming. Rosenberger and Gasko (1983), proposed 20% in order to have a relatively small standard error and 25%when working with small sample sizes. On the other side, Wilcox and Keselman (2003b) had shown that 25% and 50% (medians) trimmed mean or trimmed mean with at least 25% trimming have good control over Type I error rate but these might fail to achieve satisfactory power rate. Wilcox (2003) then suggested 20% trimming to achieve better Type I error rate and power rate.

Nevertheless, the symmetric trimming approach might lead to unnecessary trimming. Due to the predetermined amount of trimming, the data will be trimmed even if the distribution is normal, thus causing unnecessary loss of information. Symmetric trimming is also deemed to be not suitable for skewed distribution as the data should be trimmed more on the skewed tail as compared to the opposite. Hogg (1974) then proposed asymmetric or adaptive trimmed mean such that trimming can be done based on the distribution of data. When the approach was applied on hinge estimator (Reed & Stark, 1996) as the location measure for Welch test, the Type I error rate for the test was found to be well controlled (Keselman, Wilcox, Lix, Algina, & Fradette, 2007). However, adaptive trimming approach also requires a predetermined percentage of trimming as in the usual trimming approach. Another trimming approach known as automatic trimming was introduced by Wilcox and Keselman

(2002). Unlike the aforementioned trimming approaches, in automatic trimming, data will be trimmed based on the shape of the distribution via a trimming criterion.

Even though trimming is known to one of the best approach in reducing the effect of outliers, the removal of certain values in the calculation of the estimators is seemed to be the major weakness of this approach. Realizing this problem, Charles P. Winsor (1895 – 1951), a biostatistician, proposed a procedure to compensate the loss due to the trimming effect known as "winsorize" (Dixon, 1960).

1.4 Winsorizing

Other than trimming, winsorizing is another approach to deal with non-normal distribution. It is a procedure that being used to reduce the impact of outlier by limiting the extreme value in a data set. When calculating a mean, the result is always easily to being dominated by the tails. Thus, winsorization could help to reduce the effect. Winsorization is a strategy that gives more attention around the center rather than weighted in tails of a set of data which will lead to bias (Wilcox, 1997, 2012). The calculation of winsorized mean follows the steps as in trimmed mean, but the data that are supposed to be trimmed and discarded will be replaced with the highest and lowest end of the remaining data respectively (Tukey & McLaughlin, 1963). Thus, winsorizing will maintain the original sample size.

Winsorized Mean (*WM*) are the central tendency measure that applied winsorizing approach, the data winsorized symmetrically on both left and right tail of the data according to percentage been set. According to Dixon (1960) and Rivest (1994), *WM* provide better results in their study when measure under normally and skewed distributed data. Anyway, *WM* winsorized the data symmetrically regardless of distribution, thus there might have been removed certain information and lead to loss of important information throughout the winsorization process.

Adaptive Winsorized Mean (*AWM*) is the central tendency measure that able to deal with the problem of loss of important information in *WM*. *AWM* perform winsozring process according to the distribution, whether it is symmetrically or asymmetrically distributed and percentage assigned to the left and right tail depends on the shape of distributions. From the study of Ahmad Mahir and Al-Khazaleh (2009), they discovered that adaptive winsorized mean performed consistently better compared to other procedures. It is effectively being used to estimate the missing value in a time series data compared to other procedures likes averaging the whole data sets and naïve models. Another central tendency measure which adopts the winsorized approach is winsorized *MOM* (*WMOM*). This estimator has been proven to perform well in controlling false alarm (Type I error) and achieves better probability of detection (test power) in multivariate statistical process control (Haddad, Syed Yahaya & Alfaro, 2012).

The study of Ahmad Mahir and Al-Khazaleh (2009) found that compared to traditional central tendency measure, *AWM* has produced desirable result. Meanwhile, Haddad et al. (2012) shows that winsorized *MOM* performs better compared to the traditional and some other existing robust estimators in multivariate aspects. Different estimators perform differently based on the procedures used as shown by various robust statistics researchers such as Syed Yahaya (2005) and Md Yusof et al (2011). Choosing the right estimators to be used in certain procedures can promise a fruitful result. Some estimators work perfectly with certain procedures and sometimes could be otherwise.

1.5 *MOM-H* Statistic

MOM-H is a procedure with combination of central tendency measure, *MOM*, with *H*-statistic. Modified one-step *M*-estimator (*MOM*) (Wilcox & Keselman, 2003a) is a central tendency measure that was enhanced from one-step *M*-estimator (Staudte & Sheather, 1990). *MOM* applies trimming which is done automatically based on outlier detection criteria. *MOM* has good control of Type I error rate and achieves satisfactory power rate under both normal distribution and small sample size (Wilcox & Keselman, 2003a). On the other hand, *H*-statistic was originally introduced by Schrader and Hettmansperger (1980) which is readily adaptable with any central tendency measure. According to Wilcox (1997, 2012), this procedure gives reasonably good results when using *M*-estimator.

With the positive comments for both *MOM* and *H*-statistic, Othman et al., (2004) applied *MOM* on *H*-statistic in their work and observed that Type I error rate can be controlled at nominal level. However, further study by Syed Yahaya (2005) found that *MOM-H* produce low power rate although it able to control the Type I error rate at the nominal level.

1.6 Problem Statement

Even though *MOM-H* showed good result on Type I error rate, but it produced low power (Othman et al., 2004; Syed Yahaya, 2005). This might be due to the loss of information after the trimming process. As mentioned in Section 1.5, *MOM* applied trimming which is done automatically based on outlier detection criteria. The extreme values that identified by the criterion in left and right tails are eliminated and the remaining value is averaged to estimate *MOM*. Throughout this process, the sample size of the data has been reduced and is smaller than the original sample size. Sample size reduced indicate that there is loss of information when the extreme value is eliminated and the severity is depending on how much values are discarded. The loss of information for a data set with heavily skewed and heavy tails distribution will be more severe compared to the data set with slightly skewed distribution. According to Cohen (1992b) and Murphy, Myors and Wolach (2008), sample size is one of the criteria that will impact the power of statistical test. The trimming process of *MOM* which reduce the total sample size rather than retain the original has deviated from the desirable criteria of statistical test power and thus definitely will produce lower power rate accordingly.

In this study, we replace the trimming process in *MOM* with winsorization and use the winsorized *MOM* estimator as the location measure for the *H*-statistic which is denoted as *WMOM-H*. By applying *WMOM-H*, the data are trimmed and replaced accordingly based on the shape of the distribution. There seems to be two advantages using this approach. First, the trimming prior to winsorizing trims data accordingly based on the shape of the distribution. Therefore, more data will be trimmed from the skewed tail compared to other. Second advantage is that the loss of data due to trimming could be reduced when the process of winsorizing takes place because all the trimmed data will be replaced with certain values. Thus, there will be no changes in the sample size and *WMOM-H* may be able to improve the power rate of the test.

1.7 Objective of Study

The main objective of this research is to develop a new robust procedure denoted as *WMOM-H* statistic. This procedure is expected to be able to control Type I error rate better than *MOM-H* and also improve its power rate to a satisfactory level. In achieving this objective, we need to accomplish several tasks as follows:

i. To develop a new robust procedure for testing groups known as *WMOM-H*

ii. To investigate on the performance of the new proposed procedure (*WMOM-H*) in terms of Type I error rate and power rate.

iii. To compare the robustness of the proposed new statistical procedure against the parametric, nonparametric and *MOM-H* procedure under various conditions.

iv. To investigate the ability of the new proposed procedures on real data in medical manufacturing.

1.8 Significant of Study

The accomplishment of this research will benefit the researchers especially in experimental science because this achievement will bring group comparisons methodology to another higher level. For instance, some Research and Development (R&D) section in manufacturing industry might not be able to have large sample size for product change verification due to constraint of cost (high built product cost or high material cost) or time (long cycle time process). Thus, the proposed procedure is desirable as an alternative since the sample size is small and assumptions of normal distributed data with homogeneity variance are more likely to be violated.

Moreover, the proposed procedure has an important advantage which it able to preserve the original sample size. In manufacturing industry, the common practice to deal with extreme value is to remove the data from the original set after confirming that it is an outlier through technical justification. An outlier test (Grubb's test) is performed or boxplot is plotted to confirm the existing of outlier. This practice has eventually reduced the collected sample size from their original quantity and the impact is more serious if the original sample size collected is very small. Thus, the proposed procedure which is able to deal wisely with the extreme value while preserve the original sample size is a potential remedy for the measurement with small same size.

In general, the new methodology from this study can let the researchers to have some freedom in performing their data analysis without worrying about violation of assumptions, be it normality or variance homogeneity.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

Classical procedures such as Student's two-sample *t*-tes*t* and *ANOVA* are widely used to test the equality of groups. However, according to Lix and Keselman (1998), both of these procedures are easily affected by any deviation from normality, more obvious when heterogeneity of variances and group sizes exist. It is well known that the violation in the assumption of these procedures will lead to inflation in Type I error rate and depression in the power rate of the tests (Mendes & Yigit, 2012; Md Yusof, Abdullah, Syed Yahaya, & Othman, 2012b; Md Yusof, Abdullah, Syed Yahaya, & Othman, 2011; Keselman et al.,2008; Syed Yahaya et al., 2006). These liberal values of Type I error rate will subsequently result in spurious rejections of the null hypotheses while low power rate will result in differences going undetected; substantially, leading to misinterpretation of the result (Erceg-Hurn & Mirosevich, 2008). These days, *ANOVA* test is still being employed even the homogeneity of variance assumption is violated (Kulinskaya, Staudte, & Gao, 2003). However, it is well established that classical *ANOVA* is not robust enough when the assumptions are violated (Wilcox, Charlin, & Thompson, 1986). Even a slight deviation from normality will produce quite an impact on the power (Sawilowsky & Blair, 1992; Wilcox, 1995). These similar effects also occur in *t*-test (Wilcox, 1995).

As mentioned in the previous chapter, nonparametric procedure might be a better choice when the assumptions of the parametric procedures are violated, however, it is well known that nonparametric procedure is less powerful and larger sample size is needed in order to get credible results (Md Yusof et al., 2012a). For such reason, robust statistical procedures can be the prime choice to overcome the issues regarding violation of assumptions. Development of robust statistics focuses on parametric procedures, but we are not entirely convinced that the assumptions are always fulfilled (Syed Yahaya, 2005). Robust statistics being proven to be able to control the Type I error rate at the nominal level, while simultaneously produce reasonable statistical test power even under non-normal distribution and unequal variance data (Wilcox, 1997, 2012; Keselman, Wilcox, Othman, & Fradette, 2002; Keselman, Wilcox, Othman, & Fradette, 2004; Othman et al., 2004; Syed Yahaya, Othman, & Keselman, 2004a; Syed Yahaya, Othman, & Keselman, 2004b).

One of the approaches in dealing with non-normal data is trimming (Wilcox & Keselman, 2003a). According to Wilcox, Keselman, Muska and Cribbie (2000), trimming can have good control of Type I error rate and increase power rate. Conventional trimming procedure trimmed the tail of the distribution by priori fixed symmetric percentage to both tails. However, this methodology did not investigate the necessary of trimmed process on the either tails of data (Keselman et al., 2007). Table 2.1 displays some of the percentages proposed for symmetric trimmed means.

Nevertheless, symmetric trimming approach might lead to unnecessary loss of information due to the predetermined trimming percentage. Based on this approach, data will be trimmed according to the priori determined percentage regardless of distributional shape. Take for example a normally distributed set of data. When applying this approach, a predetermined percentage of data will be trimmed even though no trimming is needed in this case.

Table 2.1

Symmetric trimming percentage proposed by researcher

Researcher	Percentage (%)	Comments
Rosenberger and Gasko (1983)	20%	Have a relatively small standard error.
	25%	Work well with small sample sizes.
Wilcox and Keselman (2003b)	25%	Have good control over Type I error rate but might fail to achieve satisfactory power rate.
	50%	
Wilcox (2003)	20%	Achieve better Type I error rate and power rate

Moreover, this approach also not suitable for skewed distribution data because the data should be trimmed more on skewed tail rather than the opposite tail. Hence, to deal with these disadvantages, Hogg (1974) proposed another approach namely asymmetric or adaptive trimmed mean which trimming is based on the distribution of data. This approach is found to be able to control Type I error rate and achieve higher power rate when applied on hinge estimator (Reed & Stark, 1996) as the location measure for Welch test (Keselman et al., 2007). However, this approach also needs a predetermined percentage value, which seems unreasonable when the distribution is normal.

Another procedure which needs no predetermined percentage value, but based on empirically determined trimming is modified one-step *M*-estimator (*MOM*). This approach is able to deal with the problems in symmetric trimming approach (Wilcox, 1997, 2012). *MOM* has been proposed by Wilcox and Keselman (2003b) as central tendency measure in testing for treatment effects.

By using *MOM* on *H*-statistic (Schrader & Hettsmansperger, 1980, Othman et al., 2004) proposed a procedure known as *MOM-H*. The combination of *H*-statistic with *MOM* showed good control of Type I error rate (close to the nominal level); however, it also shows great variability across conditions. Other than that, the statistical power of this procedure is also low across conditions (Othman et al., 2004; Syed Yahaya, 2005).

Winsorization is another approach to make a statistic more robust. It shares the same process as trimming, but instead of trimmed the extreme values, this procedure however, substitutes the extreme portions by the remaining highest and lowest end of the data respectively (Tukey & McLaughlin, 1963; Dixon & Tukey, 1968). Winsorized mean is the trimmed mean that goes through the winsorizing process which preserved the original sample size and overcome the drawback of information loss due to trimming. According to Wilcox (1997, 2012), wisorization focus around the centre of a set of data rather than the tails that which might leads to bias. This measure is able to control Type I error rate under normal and skewed distributions (Dixon, 1960; Rivest, 1994). However, the usual winsorization process which follows the predetermined percentage of symmetric trimming, winsorized the data symmetrically even if the distribution of data is skewed. Furthermore, the number of data to be winsorized is based on the predetermined percentage. To deal with the problems, Ahmad Mahir and Al-Khazaleh (2009) proposed the adaptive winsorized mean and found that it performs consistently better compared to the usual winsorized mean. Anyway, adaptive winsorized mean still depends on the predetermined percentage for left and right tails respectively. On the other hand, Haddad et al. (2012) proposed to winsorize *MOM* which winsorized data asymmetrically according to its winsorizing criterion rather than predetermined percentage. The estimator which was used as the central measure for Hotelling T^2 chart performed so well in controlling false alarm rates (Type I error) regardless of the conditions and achieved desirable probability of detection (power).

Before we further discuss on the selected procedures and central tendency measures, we firstly review on the Type I error rate and power rate of a statistical test that used as the performance measurement.

2.2 Type I error rate

Type I error, α is defined as the probability of incorrectly rejecting a true null hypothesis, thus, it should be a relatively small value to avoid false rejection. The null hypothesis and alternative hypothesis for testing the equality of central tendency measures is given as

$$H_0: \theta_1 = \theta_{\,2} = \ldots = \theta_i \ ,$$

$$H_1: At\ least\ \ one\ \theta_i\ is\ different\ from\ the\ others$$

where θ_i is the central tendency parameter for $F_i : i = 1,2, \ldots, i$, and F_i is the distribution for group i. Generally, Type I error can be explained as making incorrect decision by falsely rejecting the H_0, which in fact H_0 is true. The significant level, α

is the likelihood of the risk taken on Type I error. For example, if α is 0.05, there is 5% probability that a true null hypothesis might be rejected.

In robust statistics, the robustness is the ability of the procedure to control Type I error rate closed to the significant level (nominal level), α and is stable over a range of distributions even if there is some assumption violation (Syed Yahaya, 2005).

According to Bradley (1978), a procedure is considered as robust when empirical Type I error, $\hat{\alpha}$ falls between $0.5\alpha \leq \hat{\alpha} \leq 1.5\alpha$. Thus, for a nominal level of α = 0.05, the Type I error rate should fall between 0.025 and 0.075. In empirical study, Type I errors rates above 0.075 and below 0.025 are considered as liberal and conservative respectively. However, based on Guo and Luh (2000), a test is considered as robust if it's empirical Type I error rate does not exceed 0.075 for the 5% level of significance used. Regardless of any robust criterion we adopt, a procedure that can produce a Type I error rate closest to nominal level will be the best procedure (Md Yusof et al., 2012b).

Type I error rate of independent sample *t*-test is negatively affected by extremely skewed distribution (Sawilowsky & Blair, 1992). Apart from this, variance heterogeneity, unbalanced design, and even the pairings of unbalanced sample sizes with the unbalanced group variances will also give effect on Type I error rate control. Combination of larger variance with smaller sample size showed disruption in the control of the Type I error rate (Spector, 1993; Mendes & Yigit, 2012; Md Yusof et al., 2012b; Md Yusof et al., 2011; Keselman et al., 2008; Syed Yahaya et al., 2006). According to Mendes and Akkartal (2010), the statistical power of a test also decreases when there is heterogeneity in the variances and worsens when the variance ratio among treatment group increases or the sample size is small. Besides, small departure of distribution from normality will also lead to reduce of statistical test power (Wilcox, 1998; Erceg-Hurn & Mirosevich, 2008). A small departure from normality can reduce the power rate of *t*-test from 0.96 to 0.28 (Wilcox, 1998).

2.3 Power of a Statistical Test

Power is being defined as 1 - β, where β is the Type II error probability. According to Cohen (1992b), power of a statistical test is the probability of correctly rejecting a false null hypothesis, which is the probability that the test will conclude that the phenomenon exists. On the other hand, Type II error is the probability of failing to reject the false null hypothesis. Power of a test is important as it will determine how good a test in detecting an effect. Low power will cause the result of a test to be inconclusive. Yet, in most of the work related to robust statistics, power analysis which is also known as, robustness of efficiency, is always continued to be ignored whilst robustness of validity, referring to the analysis on Type I error rate is of more concern (Syed Yahaya, 2005).

A methodological study of Clark-Carter (1997) found that if power of statistical test was not taken into account by researchers, they will encounter high risk of Type II error. Cohen (1988) stated that neglecting power analysis will lead to the slow moving of the methodological advance. Researcher may conclude the proposed procedure is good enough based on the results of Type I error rate which in fact the test with low statistical test power been reduce likelihood to rejecting the incorrect H_0. Meanwhile, neglecting statistical test power will decrease the detection of interest effects and impose negative effect on the ability of researchers to establish statistical consensus through replication (Cohen, 1988).

According to Cohen (1992b) and Murphy, Myors and Wolach (2008), power of a statistical test relies on three criteria, which are the significance criterion, sample size and effect size. Apart from the criteria, homogeneity of variances, population distribution and statistical procedure can also have effect on power rate. Increasing the deviation from the assumption of normality will give lower power rate, more significantly lower power rate when the sample size or effect size is small (Wilcox 1998; Erceg-Hurn & Mirosevich, 2008; Mendes & Akkartal, 2010)

Cohen (1992a, 1992b) proposed 0.80 as the desired power level for general used. According to Murphy et al. (2008), the minimum accepted value of statistical test power is greater than 0.50. The value smaller than 0.5, indicates that the test is more likely to be insignificant because it is unlikely to reject null hypothesis.

2.3.1 The Significant Level

The significance level, α is the probability of incorrectly rejecting the true null hypothesis. This criterion is also known as Type I error. Using α = 0.01 will result in lower power rate compared to α = 0.05 (Alan, Phyllis & John, 2008). In other words, the more conservative the significance level, the lower the power rate. As mentioned by Cohen (1988), directionality of the significance criterion also gives some impact to the power of a statistical test such that the resulting test will be more powerful if the direction is specified. A significant criterion is determined based on the effect size, sample size and desirable level of power (Murphy et al., 2008).

2.3.2 The Sample Size

Sample size is always being a concern for the reliability and credibility of a sample results. The larger the sample size, the greater the reliability and credibility of the results, thus, the greater the probability of detecting a non-null state of affairs (Syed Yahaya, 2005). According to Cohen (1988) and Murphy et al. (2008), the statistical test power will be increasing accordingly when there is an increasing in the sample size.

2.3.3 The Effect Size

According to Murphy et al. (2008), the measure of effect sizes provides a standardized index of the actual treatment impact on the dependent variable. The null hypothesis always means that the effect size is zero. In Cohen (1988), “Effect size” is “the degree to which the phenomenon is present in the population” or “the degree to which the null hypothesis is false”**.** Basically, the measures of effect size can be categorized into three which are; small, medium, and large, depending on the area of research. The value of effect size is arbitrary, but, Cohen (1988) has the conventional definitions of effect size as given in Table 2.2.

Table 2.2

Conventional effect size values by Cohen (1988)

	Number of Groups	
Effect size	2 Groups	> 2 Groups
Small	0.20	0.10
Medium	0.50	0.25
Large	0.80	0.40

2.4 The Estimators

This section delineate the estimators used in this study, started with the central tendency measure, *MOM*, followed by the trimming criterion and the scale estimator used in the criterion. The discussion then continued with the main central tendency measure used in this study, that is winsorized *MOM*.

2.4.1 Modified One-step *M*-estimator (*MOM*)

Modified one-step *M*-estimator (*MOM*) (Wilcox & Keselman, 2003b) is a central tendency measure that was modified from one-step *M*-estimator (Staudte & Sheather, 1990). One-step *M*-estimator is an approach that can help to solve the trimming problem. Based on a trimming criterion, one step *M*-estimator empirically determines whether an observation should be trimmed or not. The formula for estimate one-step *M*-estimator (Wilcox, 1997, 2012) given by

$$\hat{\theta}_j = \frac{1.28\left(MAD_{n_j}\right)(i_2 - i_1) + \sum_{i=i_1+1}^{n_j-i_2} Y_{(i)j}}{n_j - i_1 - i_2} \tag{2.1}$$

$Y_{(i)j}$ = the i^{th} ordered observations in group j.

n_j = Number of observations for group j.

Let $\widehat{M}_j$be the median for group j.

$MAD_{n_j} = MAD_j/0.6745$

$MAD_j = Median\{|Y_{1j} - \widehat{M}_j|, |Y_{2j} - \widehat{M}_j|, \cdots, |Y_{nj} - \widehat{M}_j|\}$

i_1= Number of observations Y_{ij} such that ($Y_{ij} - \widehat{M}_j$)< -1.28(MAD_{n_j})

i_2= Number of observations Y_{ij} such that ($Y_{ij} - \widehat{M}_j$) > 1.28(MAD_{n_j})

However, there is a disadvantage in one-step *M*-estimator such that it fails to perform under small sample size (Wilcox & Keselman, 2003b). For such reason, Wilcox and Keselman (2003b) then modified the estimator as in Equation 2.2 and name it as Modified One-step *M*-estimator (*MOM*) which competes well with trimmed means based estimators in terms of both power and control over the probability of Type I error even with small sample sizes. *MOM* is the average values of observations after the elimination of outliers (if any). Unlike trimmed mean which has the problem of lower breakdown point, this estimator has highest breakdown point of 0.5. Apart from low breakdown point, there is a difficulty to determine the best trimming percentage in trimmed mean since the procedure just trimmed the observations based on the priori set percentage without considering the shape of the distribution. *MOM* estimator is defined as

$$\hat{\theta}_M = \sum_{i=i_1+1}^{n_j-i_2} \frac{Y_{(i)j}}{n_j - i_1 - i_2} \tag{2.2}$$

$Y_{(i)j}$ = the i^{th} ordered observations in group *j*.

n_j= Number of observations for group *j*.

i_1 = Number of observations Y_{ij} such that $(Y_{ij} - \widehat{M}_j)$< -2.24(MAD_{n_j})

i_2 = Number of observations Y_{ij} such that $(Y_{ij} - \widehat{M}_j)$ > 2.24(MAD_{n_j})

2.4.2 Rescaling *MAD*

If the observations are randomly sampled from a normal distribution, *MAD* estimates $Z_{0.75}\,\sigma$ instead of estimate σ, the standard deviation. The 0.75 quartile of the standard normal distribution ($Z_{0.75}$) which is approximately equal to 0.6745 (Wilcox, 1997, 2012). Typically, *MAD* is being rescaled to estimate *σ* when sampling from a normal distribution. So, Wilcox (1997, 2012) suggested using MAD_n, defined as

$$MAD_n = \frac{MAD}{0.6745} \tag{2.3}$$

In general, *MAD* does not estimate *σ* when distributions are non-normal.

2.4.3 Criterion for Choosing the Sample Values

The following Equation 2.4 and Equation 2.5 is used to determine the number of extreme observations in each group *j*,

i_1 = Number of observations Y_{ij} such that $(Y_{ij} - \widehat{M}_j)$< $-K(MAD_{n_j})$ (2.4)

i_2 = Number of observations Y_{ij} such that $(Y_{ij} - \widehat{M}_j) > K(MAD_{n_j})$ (2.5)

i_1 is the number of extreme observations in the left tail, and i_2 is the number of extreme observations in the right tail. For a sample with no extreme value, wherein i_1 = i_2 = 0, *MOM* is equals to the mean for the group.

In *MOM*, the default constant *K* for the criterion is always 1.28 (unless being specified), which is 0.9 quartile of standard normal distribution (Wilcox, 1997, 2012). By using simulations with 10,000 replications, it was found that with *K* = 2.24, the standard error of the sample mean divided by the standard error of $\hat{\theta}$ is approximately 0.9 for $n_1 = n_2 = n_3 = n_4 = n_5 = 20$**,** while for small sample such as *n* = 10 and 15, the ratio is 0.88 (Wilcox & Keselman, 2003b). The value was adjusted in Othman et al. (2004) such that *K* = 2.24 for the purpose of having a reasonably small standard error when sampling from a normal distribution. The criterion for choosing sample values basically is a special case of the general method suggested by Rousseeuw and Croux (1993). As a result, the observation flagged to be eliminated is

$$(Y_{ij} - \widehat{M}_j) > 2.24(MAD_{n_j}) \text{ or } (Y_{ij} - \widehat{M}_j) < -2.24(MAD_{n_j}) \quad (2.6)$$

The advantage of using the criterion based on Median and MAD_n is the resulting finite sample breakdown point of $\hat{\theta}$ is 0.5 (Wilcox & Keselman, 2003b). Wilcox, Keselman and Kowalchuck (1998) in testing group equality suggested that substituting robust measures of central tendency and a corresponding robust measure of scale could obtain test statistic that would not loss in power rate even under the influence of non-normality.

2.4.4 *MAD*$_n$

$$MAD_n = b\ med_i\{|x_i - med_j x_j|\} \tag{2.7}$$

Equation 2.7 is the formula for median absolute deviation about the median, which is a robust scale estimator used in the trimming criterion of *MOM*. The function of the constant b is to keep the estimator consistent at normal model. When observations are drawn randomly from a normal distribution with b = 1, it estimates 0.6745σ (referred to section 2.4.1). Typically, *MAD*$_n$ is rescaled with b = 1.4826. *MAD*$_n$ has been identified by Huber (1981) as the most useful ancillary estimate of scale due to its high breakdown point.

2.4.5 Winsorized Modified One-step *M*-estimator

Winsorized Modified One-step *M*-estimator is a central tendency measure that is modified from the original *MOM* by replacing the trimming procedure with wisorizing procedure (Haddad et al., 2012). According to Haddad et al. (2012), winsoried *MOM* is given by;

$$\bar{W}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} W_{ij} \tag{2.8}$$

where

W_{ij}= the i^{th}ordered observations in group *j* (after replacement of extreme value)

n_j = Number of observations for group *j*.

On the other hand, the construct of winsorized sample is given by;

$$W_{ij} = \begin{cases} X_{(i_1+1)j}, & if\ X_{ij} \le X_{(i_1+1)j} \\ X_{ij}, & if\ X_{(i_1+1)j} < X_{ij} < X_{(n_j-i_2)j} \\ X_{(n_j-i_2)j}, & if\ X_{ij} \ge X_{(n_j-i_2)j} \end{cases} \tag{2.9}$$

X_{ij}= the i^{th}ordered observations in group *j* (before replacement of extreme value)

i_1 = Total number of smaller outlier in the data

i_2 = Total number of larger outlier in the data

Thus, $X_{ij} \le X_{(i_1+1)j}$ and $X_{ij} \ge X_{(n_j-i_2)j}$ are the equations to determine the extreme value in the given data set. After those extreme values being replaced, value of winsorized *MOM*, $\bar{W}_{ij}$, will be estimated by average the entire new data (data with extreme values being replaced).

2.5 *MOM-H* Statistic

The test statistic used in this study is *H*-statistic, originally proposed by Schrader and Hettmansperger (1980) which is readily adaptable to any central tendency measure. *H*-statistic is defined as

$$H = \frac{1}{N}\sum_{j=1}^{J} n_j \left(\hat{\theta}_j - \hat{\theta}_.\right)^2 \tag{2.10}$$

$N = \sum_j n_j$

$\hat{\theta}_. = \sum_j \hat{\theta}_j \,/J$

Keselman et al. (2002) and Othman et al. (2004) modified the *H* test by replacing $\hat{\theta}_j$ in equation 2.10 with modified one-step *M*-estimator, *MOM* ($\hat{\theta}_M$), known as *MOM-H* to test the measures of “’typical’ scores across treatment groups. The null and alternative hypothesis is as shown below.

$$H_0: \theta_{M1} = \theta_{M2} = \ldots = \theta_{Mj}$$

$$H_1: At\ least\ \ one\ \theta_{Mj}\ different\ from\ the\ others$$

MOM-H was proven to be able to control Type I error rate by Keselman et al. (2004), Othman et al., (2004) and Syed Yahaya et al. (2004a, 2004b). However, the Type I error rate does not consistently close to the nominal level (α = 0.05) across different study conditions. Moreover, the statistical test power of this procedure is also low (Othman et al., 2004; Syed Yahaya, 2005). As the sampling distribution of *MOM-H* is unknown, bootstrap method is often recommended. This method has been applied by Babu, Padmanabhan and Puri, (1999), Othman et al., (2004) and Syed Yahaya (2005) in their works on robust procedures for comparing groups. Under moderate sample size, this method has slightly better approximation compared to normal approximation theory (Babu et. al., 1999). Due to the goodness of bootstrap method, this study employed the method to test the hypothesis.

CHAPTER THREE

METHODOLOGY

3.1 Introduction

The main focus of this study is the *H*-statistic with *MOM* (*MOM-H*) and winsorized *MOM* (*WMOM-H*) as the central tendency measures. *MOM* is a form of adaptive trimmed means. Unlike the usual trimmed mean by which the amount of trimming is priori determined and trimming is done symmetrically, the trimming amount in *MOM* is empirically determined and the trimming amount on the left and right tail is not necessarily equal (symmetric). The study on *MOM-H* which was conducted by Syed Yahaya (2005) proved that the Type I error rate of the test for most of the conditions were in control. Nevertheless, the procedure was unable to achieve a desired power level. In order to rectify this problem, *MOM* is winsorized in this study. Winsorization is based on the same trimming procedure conducted on *MOM*, but the trimmed values are replaced with the highest and lowest end of the remaining data.

This study designed to cover various conditions which could highlight the strength and weakness of the procedure. For the purpose of comparison, the conditions proposed followed previous study of Syed Yahaya (2005) which includes the completely randomized design for two and four groups of small samples, types of distribution, heterogeneity of variance and pairing of the variance with group size. Based on the stated design, the study on Type I error rate and power rate of the test were conducted.

3.2 Procedures in the Study

This study covers two procedures for comparing groups known as *MOM-H* and *WMOM-H* procedures. These procedures originated from *H*-statistic, and the acronyms *MOM-H* and *WMOM-H* are based on the central tendency measures *MOM* and *WMOM* used in the test statistic respectively as shown in the figure 3.1.

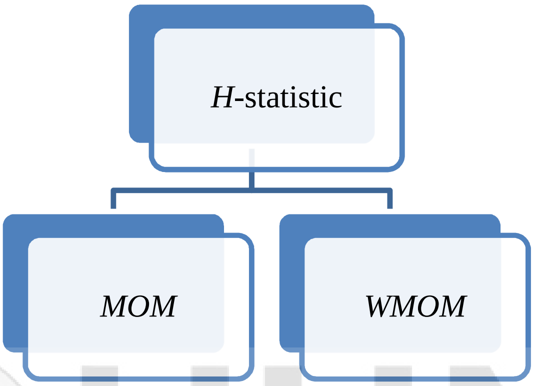

Figure 3.1. H-statistic with the central tendency measures

3.2.1 *H*-Statistic with *MOM* (*MOM-H*)

MOM is the location (central tendency) estimator which uses robust scale estimator, *MAD*$_n$ in its trimming criterion. Wilcox and Keselman (2003b) defined *MOM* as

$$\hat{\theta}_M = \sum_{i=i_1+1}^{n_j-i_2} \frac{Y_{(i)j}}{n_j - i_1 - i_2} \tag{3.1}$$

where

$Y_{(i)j}$= the i^{th}ordered observations in group *j*.

n_j = Number of observations for group *j*.

i_1 = Number of observations Y_{ij} such that (Y_{ij} –$\widehat{M}_j$) < -2.24(MAD_{nj})

i_2 = Number of observations Y_{ij} such that (Y_{ij}–$\widehat{M}_j$) > 2.24(MAD_{nj})

As per the equation above, (Y_{ij}–$\widehat{M}_j$) < -2.24(MAD_{nj}) and (Y_{ij} –$\widehat{M}_j$) > 2.24(MAD_{nj}) are to determine the extreme values in a given data set. After discarding the extreme values, *MOM* ($\hat{\theta}_M$) is estimated by taking the average of the remaining observations. The process then proceeds with the computation of the *H*-statistic such that,

$$H = \frac{1}{N}\sum_{j=1}^{J} n_j \left(\hat{\theta}_M - \hat{\theta}_.\right)^2 \qquad (3.2)$$

$$N = \sum_j n_j$$

$$\hat{\theta}_. = \sum_j \hat{\theta}_M / J$$

where,

J = the number of groups

3.2.2 *H*-Statistic with Winsorized *MOM* (*WMOM-H*)

In winsorized *MOM*, the trimmed observations are replaced by the highest and the lowest values of the remaining data. Thus winsorized *MOM* is defined as below,

$$\hat{\theta}_W = \sum_i^{n_j} \frac{Y_{new(i)j}}{n_j} \qquad (3.3)$$

where

$Y_{new(i)j}$= the i^{th}ordered observations in group j (after replacement of trimmed values)

n_j = Number of observations for group j.

$$Y_{new(i)j} = \begin{cases} Y_{(i_1+1)j}\,, & if\ (Y_{ij} - \widehat{M}_j) < -2.24\ (MAD_{nj}) \\ Y_{(i)j}\,, & if\ -2.24\ (MAD_{nj}) \leq (Y_{ij} - \widehat{M}_j) \leq 2.24\ (MAD_{nj}) \\ Y_{(n_j-i_2)j}, & if\ (Y_{ij} - \widehat{M}_j) > 2.24\ (MAD_{nj}) \end{cases}$$

i_1 = Number of observations Y_{ij} such that (Y_{ij} –$\widehat{M}_j$) < -2.24(MAD_{nj})

i_2 = Number of observations Y_{ij}such that (Y_{ij}–$\widehat{M}_j$) > 2.24(MAD_{nj})

(Y_{ij} –$\widehat{M}_j$) < -2.24(MAD_{nj}) and (Y_{ij} –$\widehat{M}_j$) >2.24(MAD_{nj}) are the formulas to determine the extreme values in the given data set. After replacing those extreme values, the *WMOM* value, $\hat{\theta}_W$, is estimated by averaging the entire new data, and then followed by the calculation of the *H*-statistic which is similar to *MOM-H* (Equation 3.2).

3.3 Variables Manipulated

The main focus of this study is on the robustness of the proposed procedure against the violation of normality and variance homogeneity assumptions. To check on the performance of the procedures, a few variables that are common to the cause of the problems have been manipulated to create conditions that could help in identifying the strength and weakness of the procedure in testing the equality of groups. Those variables are number of groups, balance and unbalance sample sizes, type of distributions, variance heterogeneity, and nature of parings which further discussed in section 3.3.1 to section 3.3.5.

3.3.1 Number of Groups

The procedures in this study will study for comparing two (*J* = 2) and more than two groups. For more than two groups case, a four groups design (*J* = 4) was chosen as this number of groups was proven to perform better in terms of Type I error rate and power rate when tested on the traditional *F*-test (Wilcox, 1994).

3.3.2 Balanced and Unbalanced Sample Sizes

The inequality in the number of observations among groups is another matter of concern in this study as this situation could inflate Type I error rate (Snedecor& Cochran, 1980; Yang, Li, & Guo, 2006; Wilcox, 2003). To check on the impact of the inequality of the sample size (number of observations) on Type I error rate and power rate of the procedures, cases for balanced as well as unbalanced sample sizes were considered.

For the purpose of comparison, the total number of observations for two and four groups follows those suggested by Syed Yahaya (2005) such that *N* = 40 and *N* = 80 respectively. For two groups case, the settings are such that n_1 = n_2 = 20 for balanced design while for the unbalanced, n_1 = 15 and n_2= 25. In the case of four groups, the distribution of sample sizes for balanced design is n_1 = n_2= n_3 = n_4 = 20 and for the unbalanced design, n_1 = 10,n_2 = 15, n_3 = 25 and n_4 = 30.

3.3.3 Types of Distributions

In the independent sample *t*-test (for 2 groups design) and *ANOVA* (for more than 2 groups design), normality is one of the criteria needed to be fulfilled in order to get accurate results. Any departure from normality (skewed or heavier tail) will lead to poor performance in Type I error rate (Bradley, 1968). Moreover, according to Sawilowsky and Blair (1992) and Wilcox (1995), there is negative impact on power rate if data set is deviated from normality. Even if the sample size among groups are equal, power rate would still be at unsatisfactory level.

Since we know that traditional statistical procedures are quite sensitive when dealing with non-normality, in this study, effect of the distribution also been investigated in terms of Type I error rate and power rate. For that purpose, three types of distribution with different level of skewness and kurtosis were chosen to evaluate on the impact of distributions on the procedures. These distributions include the standard normal distribution which represented distribution with zero skewness, chi-square distribution with three degrees of freedom represented moderate skewness and g-and-h distribution with $g = h = 0.5$ represented extremely skewed and heavy tailed distribution.

Chi-square distribution with three degrees of freedom(χ_3^2) represented moderately skewed distribution. Three degree of freedom is chosen for Chi-square distribution because it has a moderate skewness with skewness and kurtosis of $\gamma_1 = 1.63$ and $\gamma_2 = 4.00$ respectively (Othman et al., 2004). On the other hand, the g-and-h distribution with $g = 0.5$ and $h = 0.5$ which represented extremely skewed distribution (Hoaglin, 1985), has undefined theoretical value for $\gamma_1$and γ_2. In g-and-h distribution, g represents the level of skeweness and h represents the tails level in the distribution. As their respective value increases, the skewness and the heaviness of tail also will increase accordingly (Wilcox, 1997, 2012).

3.3.4 Variance Heterogeneity

Variance heterogeneity (heteroscedasticity) is another concern in testing the equality of location measure. In *ANOVA* test, when heteroscedasticity exists, the Type I error rate will inflate and power rate will subside even though sample sizes are equal among the groups (Scheffe, 1959; Schneider & Penfield, 1997; Mendes & Yigit, 2012; Fan & Hancock, 2012). Kulinskaya et al. (2003) also claimed of misleading results when heteroscedasticity exists in one-way *ANOVA*. The inflations of Type I error rate increases in tandem with the degree of heterogeneity exist. Type I error rate inflates in a lesser degree when the sample sizes are equal with small heterogeneity of variances (Box, 1954; Sawilosky, 1990). However, under moderate (e.g. 1:1:6) or large (e.g. 1:1:12) heterogeneity, the inflation becomes larger even with equal sample sizes among the groups (Rogan & Keselman, 1977; Tormarkin & Serlin, 1986; Sharma & Kibria, 2012). In Fan and Hancock (2012), other *ANOVA*-based tests (Welch's test, Brown–Forsythe test, James' second-order test, Alexander–Govern test) also show increase in Type I error rate and loss of power rate when there is heteroscedasticity.

In this study, variance with ratio of 1:36 being assigned across the groups in order to study the impact of heteroscedasticity on Type I error rate and power rate for the proposed procedure. This ratio has been applied in previous study of Syed Yahaya (2005). Keselman et al. (1998) also used ratios of 24:1 and 29:1 respectively in one-way and in completely randomized factorial designs, ratio as high as 17,977:1 was even cited in Wilcox (2003). Thus, for this study, the ratio of 1:36 seems to be reasonable in representing the extreme heteroscedasticity.

In this study the heterogeneous variance was set as 1:36 for two groups design and 1:1:1:36 for the four groups design. Another ratio that will be suggested for this study is 16: 36 for two groups and 1:4:16:36 for four groups design. The latter ratios represent moderate changes across groups (moderate heteroscedasticity) (Abdullah, Syed Yahaya, & Othman, 2011) while the earlier represent sudden changes across groups (Extreme Heteroscedasticity) (Keselman et al., 2007).

3.3.5 Nature of Pairings

When unequal sample sizes are paired with unequal variances, two types of pairings i.e. positive and negative pairings will emerge. These pairings have impact on the Type I error rate (Keselman et al., 1998; Keselman et al., 2004; Othman et al., 2004; Syed Yahaya, 2005; Fan & Handcock, 2012). A positive pairing exists when the largest number of group observations is paired with the largest group variance and the smallest number of group observations is paired with the smallest group variance. Meanwhile, negative pairing exists when largest number of group observations is paired with the smallest group variance, and the smallest number of group observations is paired with the largest group variance.

Type I error rate is easily inflated when there is a negative pairing (Box, 1953; Snedecor & Cochran, 1980; Spector, 1993; Syed Yahaya, 2005; Fan & Handcock, 2012). Based on Othman et al. (2004), positive and negative parings usually will produce conservative and liberal Type I error rate respectively. Due to the different effect on the performance of the procedure, in this study, the robustness of each proposed procedure is also evaluated under nature of paring.

3.4 Design Specification

The variables discussed in Section 3.3 were manipulated to create several conditions that could highlight the strength and weakness of the procedures. The conditions are displayed in Table 3.1 to Table 3.5 below which are condition of distribution, Design specification for the balanced *J* = 2, Design specification for the unbalanced *J* = 2, Design specification for the balanced *J* = 4 and Design specification for the unbalanced *J* = 4 respectively with the details has been discussed in section 3.3.

Table 3.1

Conditions of Distribution

Conditions	**Distributional Shape**
Perfect	Normal
Moderate Departure	Chi-square
Extreme Departure	$g = 0.5$, $h = 0.5$

Table 3.2

Design specification for the balanced J = 2

Group Sizes		**Group Variances**	
1	**2**	**1**	**2**
20	20	1	1

Table 3.3

Design specification for the unbalanced J = 2

Variance Heteroscedasticity	Pairing	Group Sizes		Group Variances	
		1	**2**	**1**	**2**
Extreme	Positive	15	25	1	36
	Negative	15	25	36	1
Moderate	Positive	15	25	16	36
	Negative	15	25	36	16

Table 3.4

Design specification for the balanced J = 4

Group Sizes				Group Variances			
1	**2**	**3**	**4**	**1**	**2**	**3**	**4**
20	20	20	20	1	1	1	1

Table 3.5

Design specification for the unbalanced J = 4

Variance Heteroscedasticity	Pairing	Group Sizes				Group Variances			
		1	**2**	**3**	**4**	**1**	**2**	**3**	**4**
Extremely	Positive	10	15	25	30	1	1	1	36
	Negative	10	15	25	30	36	1	1	1
Moderate	Positive	10	15	25	30	1	4	16	36
	Negative	10	15	25	30	36	16	4	1

3.5 Data Generation for Simulation Study

Simulated data was used in this study to test on the performance of the proposed procedure in each condition. The data were generated using SAS/IML version 9.2. Below is the data generation procedure for each distributional shape investigated, the full *SAS/IML* programming of *WMOM-H* procedures is attached in the Appendix A

i. Standard normal distribution

  a. Data were generated using *SAS* generator *RANNOR* (*SAS*, Institute, 2011) by setting mean as 0 and standard deviation as 1.

```
MTEMP = RANNOR(J(N,1,SSEED));
YTEMP = MTEMP[1:N];
```

ii. Chi-square distribution with three degrees of freedom (χ_3^2)

  a. Data were generated by initially generate three standard normal variates using (a), followed by squaring each of the three standard normal variates and sum them up.

```
TEMP=RANNOR(J(N,3,SSEED));
YTEMP = TEMP[,##];
MUCENT = 2.50;
YTEMP = YTEMP - MUCENT;
```

iii. *g*-and-*h* distribution with $g = h = 0.5$;

  a. Data were generated by initially generate standard normal variates, Z_{ij}, using (a), followed by converting standard normal variate to random variable equation through the equation

$$Y_{ij} = \begin{cases} \dfrac{\exp\left(gZ_{ij}\right) - 1}{g} \exp\left(\dfrac{hZ_{ij}^2}{2}\right), & g \neq 0 \\ Z_{ij} \exp\left(\dfrac{hZ_{ij}^2}{2}\right), & g = 0 \end{cases} \tag{3.4}$$

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

TEMP = MTEMP[1:N];

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

In g-and-h distribution, g controls the degree of skewness of the distribution while h controls the heaviness of the tails. The distribution will be symmetric when g = 0. As g and h increase, the distribution will be more skewed and the tails get heavier respectively. However, if g = h = 0, then $Y_{ij} = Z$, which is a standard normal distribution (Wilcox, 1997, 2012).

For each design in this study, 5000 datasets were simulated. Basically, the minimum datasets of 1000 are almost enough to yield the same result as a full distribution for a test at 5% level of significance (Manly, 2007). For this study, the significance level, α, was set at 0.05. However, better sampling limits were obtained when using 5000 datasets if compared to the used of 1000 datasets (Manly, 2007). As a result, this study proceeds with 5000 datasets. Each of these simulated datasets were bootstrapped 599 times for hypothesis testing.

3.6 The Settings of Central Tendency Measures for Power Analysis

Generally, there are three pattern of variability for power rate. Those are minimum, intermediate and maximum variability (Cohen, 1988), which indicate the deviation from the null hypothesis. In this study, we only focus on maximum variability pattern. Anyway, value of central tendency measures for the alternative hypothesis was determined at first before can proceed for the power test. Effect size index, *f*, which measures the degree of deviation from no effect, is the one determine each of the pattern variability. Table 3.6 below is the conventional level proposed for the effect size index, *f*.

Table 3.6

Values of effect size, f with respect to number of groups by Cohen (1988)

	Number of Groups	
Effect Size, *f*	**Two**	**Four**
Small	0.20	0.10
Medium	0.50	0.25
Large	0.80	0.40

This study covers three distributions as shown in Table 3.1. Each of the distribution was match number of groups, sample size across the groups, level of variance heterogeneity and nature of paring as presented in Table 3.2 to Table 3.5.

3.6.1 Two Groups Case

According to Cohen (1988), the effect size index, f for J = 2 is the absolute difference between two centre measures divided by their common within-population standard deviation, as in the equation below;

$$f = \frac{|m_1 - m_2|}{\sigma} \tag{3.5}$$

The test is considered as non-directional and standardized, where m_1 and m_2 are population means and σ is the standard deviation of either population. For this study the value of m_1 and m_2 is replace by the proposed robust estimators, *WMOM*.

This study only focus on maximum variability, so the effect size index, *f*, is 0.80. There is slightly different in the definition of f for the balanced (equal sample size and equal variances across the group) and unbalanced design (unequal sample size and unequal variance across group). For the balanced design, there is a common within-population σ, thus, f is defined as in Equation 3.5. Meanwhile, for the unbalanced design, since the variances are not equal, the equation requires a slight modification, by which pool variance is required in the denominator. The calculation of pool variance is as below;

$$\sigma' = \sqrt{\frac{(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2}{n_1 + n_2 - 2}} \tag{3.6}$$

3.6.1.1 Balanced design (*J* = 2)

As shown in Table 3.2, the sample sizes are $n_1 = n_2 = 20$ and the variances are $\sigma_1 = \sigma_2 = 1$ for balanced design, J = 2. For the maximum pattern variability, the appropriate effect size, f = 0.80. If m_1 = 1, and f = 0.80, the value for m_2 is determined by Equation 3.5 which is;

$$\frac{|1 - m_2|}{1} = 0.80$$

$$m_2 = 1.80$$

So, the setting of central tendency measures for *J* = 2 in our study is (1, 1.80).

3.6.1.2 Unbalanced design (*J* = 2)

For the unbalanced design, the sample sizes are n_1 = 15 and n_2 = 25 (refer to Table 3.3) with two levels of variance heterogeneity. The first level, $\sigma_1^2 = 1$ and $\sigma_2^2 = 36$ represents extreme heterogeneity while the other level with $\sigma_1^2 = 16$ and $\sigma_2^2 = 36$ represents moderate heterogeneity.

Using the suggested sample sizes and group variances, the pooled variance, $\sigma'$is calculated as in Equation 3.6 followed by calculating for the m_2. The calculation is almost the same as the balanced design, except that the standard deviation, σ, is replaced by the calculated pooled variance.

Extreme variance heterogeneous:

$$\sigma' = \sqrt{\frac{(15-1)1+(25-1)36}{15+25-2}}$$

$$\sigma' = 4.81$$

$$\frac{|1-m_2|}{4.81} = 0.80$$

$$m_2 = 4.85 \ \approx 5$$

Moderate variance heterogeneous:

$$\sigma' = \sqrt{\frac{(15-1)16+(25-1)36}{15+25-2}}$$

$$\sigma' = 5.35$$

$$\frac{|1 - m_2|}{5.35} = 0.80$$

$$m_2 = \ 5.28 \approx 5$$

Therefore, the setting of central tendency measure for *J* = 2 in our study is (1, 5). For the unbalanced design, there are conditions for positive and negative pairings however we only focused on positive as it generate more variation (Syed Yahaya, 2005). Table 3.7 shows the summary of the setting of central tendency measures for *J* = 2 under unbalanced design.

Table 3.7

The settings of central tendency measures for J = 2 unbalanced design

Variability	**Variance**	**Effect size, *f***	**Location measure**	**Location measure**
			1, m_1	**2, m_2**
Maximum	Extreme	0.8	1.0	5.0
	Moderate			

3.6.2 Four Groups Case

When the number of groups is greater than two, the relationship between the effect size and range of standardized means depends upon the range over their means dispersion as represented by

$$f = \frac{\sigma_m}{\sigma} \tag{3.7}$$

where σ_m is the standard deviation of the population means expressed in original scale units and σ is the standard deviation within the population.

Within those four means, the largest and smallest values of the mean are used to determined *d,* which is the range of the standardized means such that

$$d = \frac{m_{max} - m_{min}}{\sigma} \tag{3.8}$$

where m_{max} and m_{min} are the largest and the smallest of the four means, while σ is the standard deviation within the population. Anyway, Equation 3.7 only suitable for unequal sample size, thus, for equal sample size, which has a simpler approach, will be discussed in the next sub-section.

Under balanced design (equal sample size and equal variance across groups), the relationship between *f* and *d* for a given number of groups (*J*) is fixed. With regards to the value of *d* and *J*, Cohen (1988) has set the standard generalized pattern for each degree of variability (minimum, intermediate and maximum). Thus, based on the standard generalized pattern we are able to set the central tendency measure for the maximum variability. Table 3.8 presents the standard generalized pattern of *d* for maximum variability pattern under four groups case, *J* = 4.

Table 3.8

The standard pattern variability for J = 4 by Cohen (1988)

Degree of Variability	**Pattern variability**
Maximum	$-\frac{1}{2}d, -\frac{1}{2}d, \frac{1}{2}d, \frac{1}{2}d$

3.6.2.1 Balanced Design (*J* = 4)

Under balanced design, the relationship between *f* and *d* for maximum variability pattern depends whether the group size is even or odd as shown below:

When J in even number,

$$d_{max} = 2f \tag{3.9}$$

When *J* in odd number,

$$d_{max} = f\frac{2J}{\sqrt{J^2 - 1}} \tag{3.10}$$

In our study, Equation 3.9 being applied since the group size is four (*J* = 4). Based on table 3.6, the effect size index, *f*, for maximum pattern variability for four groups is 0.40, thus, the *d* value is

$$d_{max} = 2(0.40) = 0.8$$

As a result, the dispersion of the central tendency measures is (-0.4, -0.4, 0.4, 0.4) and Table 3.9 is the summary of the generalized and the central tendency measure dispersion respectively as suggested by Cohen (1988).

Table 3.9

Dispersion of central tendency measures corresponding to the pattern variability for J = 4 balanced design.

Variability pattern	**Generalized dispersion**	**Dispersion of central tendency measures**
Maximum	$-\frac{1}{2}d, -\frac{1}{2}d, \frac{1}{2}d, \frac{1}{2}d$	-0.4, -0.4, 0.4, 0.4

3.6.2.2 Unbalanced Design (*J* = 4)

As in section 3.6.2, Equation 3.7 and 3.8 are used to determine the value of *f* and *d* respectively for the setting of central tendency measure. According to the central tendency measures for maximum variability pattern proposed by Cohen (1988) which is $\left(-\frac{1}{2}d, -\frac{1}{2}d, +\frac{1}{2}d, +\frac{1}{2}d\right)$, the setting of central tendency measures for this condition is (-1, -1, 1, 1). This setting represents the setting used in Keselman et al. (2004), Othman et al. (2004) and Syed Yahaya (2005).

As per Equation 3.7, standard deviation of the population means (σ_m) expressed in original scale unit which is defined as below;

$$\sigma_m = \sqrt{\frac{\sum_{j=1}^{J} n_j (m_j - m)^2}{N}} \qquad (3.11)$$

$$m = \frac{\sum_{j=1}^{J} n_j m_j}{N}$$

for $j = 1, \dots \dots, J$

Standard deviation within the populations (σ) for *J* = 4, it is defined as

$$\sigma = \sqrt{\frac{(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2 + (n_3 - 1)\sigma_3^2 + (n_4 - 1)\sigma_4^2}{n_1 + n_2 + n_3 + n_4 - 4}} \qquad (3.12)$$

Given that $n_1 = 10, n_2 = 15, n_3 = 25, n_4 = 30$, $N = 80$ and $m_1 = -1, m_2 = -1, m_3 = 1, m_4 =$. Thus,

$$m = \frac{10(-1) + 15(-1) + 25(1) + 30(1)}{80} = \frac{30}{80} = 0.375$$

$$\sigma_m = \sqrt{\frac{10(-1 - 0.375)^2 + 15(-1 - 0.375)^2 + 25(1 - 0.375)^2 + 30(1 - 0.375)^2}{80}}$$

$$\sigma_m = \sqrt{\frac{68.75}{80}} = 0.927$$

For extreme variance heterogeneous with $\sigma_1^2 = 36, \sigma_2^2 = 1, \sigma_3^2 = 1, \sigma_4^2 = 1$

$$\sigma = \sqrt{\frac{(10-1)36 + (15-1)1 + (25-1)1 + (30-1)1}{10+15+25+30-4}}$$

$$\sigma = \sqrt{\frac{391}{76}} = 2.268$$

Therefore, the effect size index,

$$f = \frac{0.927}{2.268} = 0.41 \approx 0.40$$

Whilst, for moderate change variance design (moderate heteroscedasticity), the central tendency measure is set as (-1.3, -1.3, 1.3, 1.3) which d = 2.6. The proof that this dispersion was from the maximum variability pattern with the effect size index, f = 0.40 is shown as below;

Given that $n_1 = 10, n_2 = 15, n_3 = 25, n_4 = 30$, N = 80, $m_1 = -1.3, m_2 = -1.3, m_3 = 1.3$ and $m_4 = 1.3$. Thus,

$$m = \frac{10(-1.3) + 15(-1.3) + 25(1.3) + 30(1.3)}{80} = \frac{39}{80} = 0.488$$

$$\sigma_m$$

$$= \sqrt{\frac{10(-1.3-0.488)^2 + 15(-1.3-0.488)^2 + 25(1.3-0.488)^2 + 30(1.3-0.488)^2}{80}}$$

$$\sigma_m = \sqrt{\frac{116.19}{80}} = 1.205$$

For moderate variance heterogeneous with $\sigma_1^2 = 36, \sigma_2^2 = 16, \sigma_3^2 = 4, \sigma_4^2 = 1$

$$\sigma = \sqrt{\frac{(10-1)36 + (15-1)16 + (25-1)4 + (30-1)1}{10 + 15 + 25 + 30 - 4}}$$

$$\sigma = \sqrt{\frac{673}{76}} = 2.976$$

Therefore, the effect size index,

$$f = \frac{1.205}{2.976} = 0.405 \approx 0.4$$

Table 3.10

Dispersion of central tendency measures corresponding to the pattern variability for J = 4 unbalanced design

Pattern Variability	**Variance Heterogeneity**	**Generalized dispersion**	**Dispersion of central tendency measures**
Maximum	Extreme	$-\frac{1}{2}d, -\frac{1}{2}d, \frac{1}{2}d, \frac{1}{2}d$	-1, -1, 1, 1
	Moderate	$-\frac{1}{2}d, -\frac{1}{2}d, \frac{1}{2}d, \frac{1}{2}d$	-1.3, -1.3, 1.3, 1.3

In this study, we will apply the setting of central tendency measures of (-1, -1, 1, 1) and (-1.3, -1.3, 1.3, 1.3) for extreme heterogeneous and moderate level of heterogeneous variance design respectively.

3.7 Bootstrap Method

Bootstrap is a computer-based methodology to obtain a more accurate estimation of central tendency measure as compared to the traditional statistical method (Efron & Tibshirani, 1986). Bootstrap method was applied in this study because *MOM-H* statistic's sampling distributions are intractable.

Thus, the percentile bootstrap method which is widely used by most robust statisticians to assess statistical significance is suggested. This method has been used for example by Othman et al. (2004) and Syed Yahaya (2005) to get the significance level for *MOM-H* statistic. According to Babu et al. (1999), this method is expected to give better approximation especially under moderate sample size. Besides, Wilcox (1997, 2012) recommended using bootstrap methods in small sample size in order to get good control of Type I error rate.

Keselman et al. (2002) also discovered that Type I error control could be improved by combining bootstrap method with robust based central tendency measure (Keselman et al., 2002). A study from Westfall and Young (1993) demonstrated that by combining bootstrap method and trimmed means would result in better control of Type I error rate. This assumption was supported by the asymptotic results garnered by Hall and Padmanabhan (1992). Wilcox et al. (1998) in their work, combined trimmed means with bootstrap methods, and obtained good control of Type I error rate. Their study compared robust statistics by Welch (1951), Box (1954), and Alexander and Govern (1994) and their results produced non liberal Type I error rate. Thus, as mentioned by Wilcox (1997, 2012), the practicality of bootstrap is irrefutable as it is widely used. For power rate, Beran (1986) discovered that the power function is almost similar between bootstrap method and the classical *t*-test.

The time taken for the bootstrap computation mainly based on the bootstrap replication, *B*. According to Efron and Tibshirani (1993), *B* = 50 is enough to give a reliable estimation, but larger *B* is needed for percentiles estimation. Thus, they suggested that *B* should be at least 500 or 1000 in order to make the variability adequately low for estimated percentile. Davision and Hinkley (1997) also suggested that *B* should be at least 500 in order to obtain accurate results as there is large variability in the percentile estimation if number of simulation is less than 100. However, when choosing the value of *B*, Hall (1986) suggested that the value of *B* should be chosen so that $1-\alpha$ is a multiple of $(B+1)^{-1}$ and $1-\alpha=$ .95 is the primary focus of the study. The small adjustment was proven in Wilcox et al. (1998), whereby *B* = 599 has decreased the liberal Type I error rate of Welch statistic compared to *B* = 600. Thus, bootstrap replication, *B* =599 is chosen in this study.

3.7.1 *MOM-H* and *WMOM-H* with Bootstrap Method

The performance of the proposed procedure was assessed based on their Type I error rate and power rate. Percentile bootstrap method was used to compute the *p*-value which consequently used to calculate Type I error rate and power rate. To obtain the statistical significance for *WMOM-H* statistic through percentile bootstrap method, below are the steps to calculate the *p*-values;

i. *WMOM-H* statistic is calculated based on available data.

ii. Observation (n_j) from each group are resample (with replacement) randomly to obtain bootstrap sample.

iii. Each bootstrap was centred with estimated *MOM* or *WMOM* respectively such that $C_{ij}^{*} = Y_{ij}^{*} - WMOM$.

iv. *WMOM-H**is calculated using the respective value of C_{ij}^{*}.

v. Step (i) to (iv) were repeated for *B* times (*B =599*) to generate

vi. $WMOM - H_1^*$, $WMOM - H_2^*$, $WMOM - H_3^*, \ldots$, $WMOM - H_B^*$

vii. The *p*-value is calculated by *(*Number *of WMOM-H** > *WMOM-H)/B*.

viii. Lastly, step (i) to (vi) were repeated for 5000 times (simulation) and the average value is computed to obtain Type I error rate.

The calculation of Type I error rate and power rate follow the same steps, except for the setting of the central tendency measures in the power rate calculation. The central tendency measures to compute Type I error rate were always set to be zero in order to remain the true null hypothesis. On the other hand, the central tendency measures set to obtain statistical test power are not zero but vary according to the effect size and pattern variability. The full SAS/IML programming of *WMOM-H* procedures is attached in the Appendix A.

CHAPTER FOUR

ANALYSIS AND FINDINGS

4.1 Introduction

In this chapter, we are going to compare the proposed procedure, *W-MOMH* with *MOM-H* and their parametric and nonparametric counterparts in terms of Type I error rate and power rate of a test. As mentioned in chapter 3, to measure the robustness of the proposed procedure, the procedures have been exposed to various conditions which include balanced and unbalanced sample sizes, equal and unequal variances (moderate or extreme differences), nature of pairing of sample sizes to variances and type of distributions.

This chapter is organized based on the investigated conditions, namely number of groups, which is then breakdown to balanced and unbalanced sample sizes. The comparison is based on Type I error rate and power rate of a test, which are summarized in the form of tables.

The first column of each table displays the different types of distribution selected in this study with different levels of skewness and kurtosis. These distributions are the standard normal distribution, chi-square distribution with three degrees of freedom and *g*-and-*h* distribution with $g = h = 0.5$ which represents distribution with zero skewness, moderate skewness and extreme skewness with heavy-tailed respectively. The second column lists the two nature of pairings, positive (*+ve*) and negative (*-ve*) pairings based on the pairing assignment between sample sizes and group variances. The pairing is further divided into moderate ("*m +ve*" and "*m -ve*") or extreme ("*+ve*" and "*-ve*") changes according to the level of heterogeneity of the variances (moderate and extreme heteroscedasticity). The pairing column only exists in unbalanced design as there is no pairing exist in the balanced design. The rest 4 columns record the Type I error rate obtained from the respective procedures (as stated on the column header).

Bradley's (1978) liberal criterion of robustness is applied in this study to evaluate the robustness of the procedure under different conditions. Based on the criterion a test with 5% of statistical significant level should produce Type I error rate between 0.025 and 0.075 in order to be considered as robust under certain condition.

Three existing procedures are being compared to the proposed procedure *WMOM-H* in this study, which includes *MOM-H*, classical parametric procedures, and nonparametric procedures. For two group's case, *t*-test and Mann-Whitney represent parametric and nonparametric procedure respectively. Whilst, Analysis of Variance (*ANOVA*) and Kruskal-Wallis is respectively represented the four groups' case. The goal of the *WMOM-H* is to test the equality of the groups such that

$$H_0: \theta_1 = \theta_2 = \ldots = \theta_j$$

$$H_1: At\ least\ \ one\ \theta_j\ different\ from\ the\ others$$

4.2 Type I error rate for *J* = 2

For *J* = 2, the null hypothesis is $H_0: \theta_1 = \theta_2$ while alternative hypothesis is $H_1: \theta_1 \neq \theta_2$ where θ represents the location measure. Table 4.1 and 4.2 presents the empirical Type I error rate for balanced and unbalanced design.

4.2.1 Balanced Design (*J* = 2)

The tests for balanced design were conducted with groups having equal number of observations and equal variances across the groups. Table 4.1 shows the results for *J* = 2. The results indicate that regardless of distributions, almost all the procedures investigated fulfilled Bradley's robust criterion except for *MOM-H* under extremely skewed and heavy-tailed distribution.

Table 4.1

Empirical Type I error rate for balanced design, J = 2

	Robust Procedure		**Parametric Procedure**	**Nonparametric Procedure**
Distribution	***MOM-H***	***WMOM-H***	***t*-test**	***MW***
Normal	0.0410	0.0526	0.0528	0.0526
Chi-square	0.0422	0.0526	0.0500	0.0566
g = *h*= 0.5	0.0244	0.0396	0.0288	0.0526
Grand Average	***0.0359***	***0.0483***	***0.0439***	***0.0539***

Based on the grand average, *WMOM-H* generates the best average values among the procedures, which is not only nearest to the nominal level, but its value does not exceed the nominal level. Mann-Whitney (*MW*) generates the highest average Type I error rate (0.0539), but its deviation from the nominal level is less than *t*-test (0.0439). On the aspect of winsorization, the average result between *MOM-H* and *WMOM-H* shows great improvement on *WMOM-H*. The Type I error rate for *WMOM-H* is closer to the nominal level, which indicates that winsorization can increase the robustness of the procedure.

With regards to distributional shapes, *W-MOMH* and *MW* perform equally good (0.0526) in controlling Type I error rate surpassing *t*-test (0.0528) under the normal distribution. For moderately skewed distributions, *t*-test shows perfect control of

Type I error rate (0.500), while for extremely skewed and heavy-tailed distribution; *MW* performs the best (0.0526).

In general, the result for balanced design with *J* = 2 shows that proposed procedure has better control of Type I error rate.

4.2.2 Unbalanced Design (*J* = 2)

Under unbalanced design with group *J* = 2, several tests were carried out which paired unequal number of observations, with unequal variance across groups. Table 4.2 shows that nearly 73% of the results across the assigned condition fulfilled Bradley's robust criterion, and majority produced by robust procedure (i.e. *MOM-H and WMOM-H*).

The proposed procedure *WMOM-H* has the same performance as the balanced design as shown by the "Grand Average" value. The value (0.0530) falls within the range of 0.025 to 0.075 and closest to the nominal level compared to the other procedures. The next better procedure is *MOM-H*, followed by *t*-test and Mann-Whitney (*MW*), Type I error rate of 0.0418, 0.0621 and 0.0683, respectively. Additionally, *WMOM-H* shows further improvement across various distributions and pairing, compared to the original procedure (i.e. *MOM-H)*. It also has good control of Type I error rate regardless of the extreme or moderate changes in variances (extreme and moderate heteroscedasticity), with the overall results being within Bradley's criterion.

Overall *MOM-H* also produces results within the Bradley's criterion of robustness, with the exception of positive pairing, combined with moderate heteroscedasticity design under extreme skewed and heavy-tailed distribution. In contrast, the traditional statistical *t*-test and *MW* (parametric and nonparametric procedures) show failures in controlling Type I error rate, especially in the case of extreme heteroscedasticity conditions.

Table 4.2

Empirical Type I error rate for unbalanced design, J = 2

		Robust Procedure		**Parametric Procedure**	**Nonparametric Procedure**
Distribution	**Pairing**	***MOM-H***	***WMOM-H***	***t-test***	***MW***
Normal	+*ve*	0.0496	0.0628	0.0198	0.0448
	-*ve*	0.0470	0.0570	0.1268	0.1086
	m +*ve*	0.0388	0.0486	0.0360	0.0420
	m -*ve*	0.0384	0.0504	0.0704	0.0600
	Average	***0.0435***	***0.0547***	***0.0633***	***0.0639***
Chi-square	+*ve*	0.0626	0.0684	0.0238	0.0666
	-*ve*	0.0642	0.0674	0.1678	0.1312
	m +*ve*	0.0382	0.0538	0.0334	0.0502
	m -*ve*	0.0478	0.0556	0.0800	0.0770
	Average	***0.0532***	***0.0613***	***0.0763***	***0.0813***
g = *h* = 0.5	+*ve*	0.0328	0.0532	0.0118	0.0426
	-*ve*	0.0324	0.0436	0.1048	0.0976
	m +*ve*	0.0222	0.0380	0.0238	0.0422
	m -*ve*	0.0276	0.0368	0.0464	0.0568
	Average	***0.0288***	***0.0429***	***0.0467***	***0.0598***
Grand Average		***0.0418***	***0.0530***	***0.0621***	***0.0683***

The result for the unbalanced design with the *J* = 2 generally suggests that the proposed procedure (i.e. *WMOM-H*) generates better control of Type I error rate, even under extreme heterogeneity variance across groups with unequal observations, and this is noticeably better compared to traditional statistical procedures.

4.3 Type I error rate for *J* = 4

The null hypothesis for *J* = 4 is $H_0: \theta_1 = \theta_2 = \theta_3 = \theta_4$ while alternative hypothesis is $H_1: At\ least\ \ one\ \theta_j\ different\ from\ the\ others$ where θ represents the location measure. Tables 4.3 and 4.4 present the empirical Type I error rate for both balanced and unbalanced designs for this case. The results in the following subsections show that the proposed procedure has good control of Type I error rate for both balanced and unbalanced designs.

4.3.1 Balanced Design (*J* = 4)

In terms of balanced design for *J* = 4, several tests were conducted with all of the groups having equal number of observations and equal variances, across groups. The results in Table 4.3 show that *WMOM-H* has a considerable improvement compared to *MOM-H* in terms of Type I error rate control.

Table 4.3

Empirical Type I error rate for balanced design, J = 4

	Robust Procedure		**Parametric Procedure**	**Nonparametric Procedure**
Distribution	***MOM-H***	***WMOM-H***	***ANOVA***	***KW***
Normal	0.0256	0.0420	0.0518	0.0498
Chi-square	0.0170	0.0304	0.0450	0.0440
g = *h* = 0.5	0.0098	*0.0238*	0.0290	0.0498
Grand Average	***0.0175***	***0.0321***	***0.0419***	***0.0479***

However, based on the grand average, Kruskal-Wallis (*KW*) generates the optimal average values of Type I error rate (0.0479), followed by *ANOVA* (0.0419), *WMOM-H* (0.0321) and *MOM-H* (0.0175). The Type I error rate for *WMOM-H* were within Bradley’s criterion for both Normal and moderate skewed (Chi-square) distributions, but was slightly below the lowest limit of the criterion for extreme skewed and heavy tail distributions. The overall robustness represented by the "Grand Average" showed improvement from 0.0175 to 0.0321, compared to the original procedure (*MOM-H*).

Across distributions, *KW* ranked the highest in terms of controlling Type I error rate with values consistently close to the nominal level.

The results for balanced design with J = 4 generally show that the proposed procedure (*WMOM-H*) has better control of Type I error rate compared to the original procedure (*MOM-H*).

4.3.2 Unbalanced Design (J = 4)

In unbalanced design for J = 4, several tests were conducted on the pairing of unbalanced sample size and unequal variances. Table 4.4 shows that only 69% of the results across the different conditions tested fulfilled Bradley's robust criterion. Those which did not fulfill the robust criterion, majority were from the traditional statistical procedures.

Moreover, the proposed procedure, *WMOM-H*, generated the optimal grand average value for Type I error rate (0.0524), followed by *MOM-H* (0.0405), Kruskal-Wallis (*KW*; 0.0843) and *ANOVA* (0.1422) with the *KW* and *ANOVA* value exceeding the criterion interval. Additionally, *WMOM-H* showed an improvement in robustness compared to the original procedure, with the grand average of Type I error closer to the nominal level of 0.050. Furthermore, *WMOM-H* was the only procedure rate that was fully in control of Type I error rate across different distributions and pairing designs, regardless of extreme or moderate variances in heterogeneity. *MOM-H* also generated similar results, with the exception of the moderate heteroscedasticity design with extremely skewed and heavy-tailed distributions.

Table 4.4

Empirical Type I error rate for unbalanced design, J = 4

		Robust Procedure		**Parametric Procedure**	**Nonparametric Procedure**
Distribution	**Pairing**	***MOM-H***	***WMOM-H***	***ANOVA***	***KW***
Normal	+*ve*	0.0486	0.0644	0.0336	0.0448
	-*ve*	0.0528	0.0622	0.2850	0.1158
	m +ve	0.0426	0.0560	0.0288	0.0326
	m -ve	0.0404	0.0518	0.2224	0.1312
	Average	***0.0461***	***0.0586***	***0.1425***	***0.0811***
Chi-square	+*ve*	0.0646	0.0722	0.0596	0.0668
	-*ve*	0.0660	0.0714	0.3254	0.1258
	m +ve	*0.0408*	*0.0574*	*0.0328*	*0.0466*
	m -ve	*0.0368*	*0.0474*	*0.2646*	*0.1578*
	Average	***0.0521***	***0.0621***	***0.1706***	***0.0993***
g = *h* = 0.5	+*ve*	0.0300	0.0432	0.0256	0.0442
	-*ve*	0.0290	0.0422	0.2400	0.1022
	m +ve	0.0190	0.0356	0.0128	0.0348
	m -ve	0.0150	0.0250	0.1760	0.1082
	Average	***0.0233***	***0.0365***	***0.1136***	***0.0724***
Grand Average		***0.0405***	***0.0524***	***0.1422***	***0.0843***

The failure of *ANOVA* and *KW* to be in the robust criterion was very much influenced by the negative pairing of both moderate and extreme heteroscedasticity, which lead to high grand average of Type I error rate (> 0.0750).

With respective to the average by distribution, *MOM-H* obtained the highest rank of Type I error rate control for zero skewed (Normal) and moderate skewed (Chi-square) distributions with average Type I error rate of 0.0461, 0.0521 respectively. Meanwhile, *WMOM-H* produced the best Type I error rate control for extremely skewed with heavy tail distribution, producing average Type I error rate of 0.0365.

The results for the unbalanced design with *J* = 4 show that *WMOM-H* has better control of Type I error rate across different conditions compared to the traditional statistical procedures.

4.4 Power rate of Test for *J* = 2

As mentioned in Section 3.6, this study focused only on maximum variability. Therefore, the results presented in the subsequent sections only presents result of maximum variability. Similar to Type I error rate, the null hypothesis and alternative hypothesis are $H_0: \theta_1 = \theta_2$ and $H_1: \theta_1 \neq \theta_2$ respectively where θ represents the location measure. The power rate of the test for both balanced and unbalanced designs is presented in Tables 4.5 and 4.6. The results show that, in general, the proposed procedure has better power rate compared to *MOM-H*, *t*-test and Mann-Whitney for both balanced and unbalanced designs.

4.4.1 Balanced Design (*J* = 2)

In terms of balanced design, both groups were assigned with equal number of observations (sample size) and equal variance across groups. The results in Table 4.5 show that the proposed procedure improved in power rate for each distribution compared to the original procedure (*MOM-H*). It was also found to possess a higher power rate compared to the *t*-test when performing under extremely skewed and heavy tailed distributions (*g* = *h* = 0.5).

Across the Grand Average values, Mann-Whitney (*MW*) generated the optimal power rate (0.4514) among the four different procedures, followed by *WMOM-H*, *t*-test, and *MOM-H*, with grand average values of 0.3961, 0.3412 and 0.3383 respectively. With reference to the type of distributions, *MW* generated the best power rate across the various distributions, compared to other procedures.

Table 4.5

Empirical power rate for balanced design, J = 2

	Robust Procedure		**Parametric Procedure**	**Nonparametric Procedure**
Distribution	***MOM-H***	***WMOM-H***	***t-test***	***MW***
Normal	0.5344	0.6604	0.6866	0.6650
Chi-square	0.1412	0.1740	0.1806	0.2522
$g = h = 0.5$	0.3392	0.3538	0.1564	0.4370
Grand Average	***0.3383***	***0.3961***	***0.3412***	***0.4514***

WMOM-H generally generated better power rate compared to *t*-test. The proposed procedure also improved considerably compared to the original procedure under *J* = 2 balanced design.

4.4.2 Unbalanced Design (*J* = 2)

For the unbalanced design where tests were conducted on unequal number of observations paired with unequal variances. Similar to the balanced design, the results in Table 4.6 show that the proposed procedure has a considerable hike in power rate for various distributions, compared to the original procedure. In addition, it generated better power rate than *t*-test for the overall comparison (Grand Average). However, among the four investigated procedures, based on the Grand Average, Mann-Whitney (*MW*) generated the best power rate of 0.4724 followed by *WMOM-H* (0.4501), *t*-test (0.3782) and *MOM-H* (0.3726).

Table 4.6

Empirical power rate for unbalanced design, J = 2

		Robust Procedure		**Parametric Procedure**	**Nonparametric Procedure**
Distribution	**Pairing**	***MOM-H***	***WMOM-H***	***t*-test**	***MW***
Normal	*+ve*	0.7672	0.8710	0.7458	0.7562
	-ve	0.5208	0.6358	0.8298	0.7142
	m +ve	0.5262	0.6490	0.6172	0.6134
	m -ve	0.4408	0.5482	0.6806	0.6042
	Average	***0.5638***	***0.6760***	***0.7184***	***0.6720***
Chi-square	*+ve*	0.1744	0.3700	0.3056	0.1656
	-ve	0.2006	0.1864	0.2172	0.3342
	m +ve	0.1316	0.2106	0.1720	0.1610
	m -ve	0.1526	0.1756	0.1802	0.2810
	Average	***0.1648***	***0.2357***	***0.2188***	***0.2355***
$g = h = 0.5$	*+ve*	0.5436	0.6742	0.2308	0.6596
	-ve	0.3880	0.3656	0.2616	0.5672
	m +ve	0.3198	0.3850	0.1340	0.3958
	m -ve	0.3054	0.3294	0.1638	0.4160
	Average	***0.3892***	***0.4386***	***0.1976***	***0.5097***
Grand Average		***0.3726***	***0.4501***	***0.3782***	***0.4724***

According to the results depicted in Table 4.6, the *WMOM-H* procedure generated better power rate in almost all conditions compared to original procedure, except in the case of negative pairing with extreme heterogeneity variances under moderately skewed (Chi-square) and extremely skewed and heavy tails distributions ($g = h = 0.5$). Comparing the Average results (each distribution) for *WMOM-H* procedure to the parametric *t*-test, we observed that *WMOM-H* performed better in moderately skewed (Chi-square) and extremely skewed and heavy tails distributions ($g = h = 0.5$).

The results for unbalanced designs with $J = 2$ reveal that the proposed procedure has better power rate compared to *MOM-H* and *t*-test, especially in moderately and extremely skewed and heavy tails distributions.

4.5 Power rate of Test for *J* = 4

In terms of *J* = 4, the null hypothesis and alternative hypothesis are $H_0: \theta_1 = \theta_2 = \theta_3 = \theta_4$ and $H_1: At\ least\ \ one\ \theta_j\ different\ from\ the\ others$ respectively where θ represents the location measure. The power rate of test for balanced and unbalanced designs is presented in Tables 4.7 and 4.8 respectively. The results show that the proposed procedure has better power rate compared to *MOM-H*, in both balanced and unbalanced designs.

4.5.1 Balanced Design (*J* = 4)

For *J* = 4 with balanced design, tests were conducted with all groups having equal number of observations, with equal variance across all groups. The results in Table 4.7 show the empirical power rate for the balanced design with maximum pattern variability. According to the results, *WMOM-H* showed considerable improvement compared to the original procedure in terms of empirical power rate.

Table 4.7

Empirical power rate for balanced design, J = 4

	Robust Procedure		**Parametric Procedure**	**Nonparametric Procedure**
Distribution	***MOM-H***	***WMOM-H***	***ANOVA***	***KW***
Normal	0.6260	0.7828	0.8364	0.8148
Chi-square	0.0836	0.1408	0.2016	0.3042
***g* = *h* = 0.5**	0.2702	*0.3120*	0.1298	0.5506
Grand Average	***0.3266***	***0.4119***	***0.3893***	***0.5565***

Kruskal-Wallis (*KW*) generated the optimal power rate in the overall performance (refer to Grand Average), followed by *WMOM-H*, *ANOVA*, and *MOM-H*, with grand average power rate of 0.5565, 0.4119, 0.3893 and 0.3266 respectively. Across three different distributions, *ANOVA* and *KW* have generated the highest power rate under zero skewness distribution for *ANOVA*, moderately skewed and extremely skewed with heavy tailed distributions for *KW*.

In general, the results for balanced design for *J* = 4 show that *WMOM-H* generates better power rate compared to the original procedure, and is the more powerful compared to investigated procedures under extremely skewed and heavy tailed distributions except *KW*.

4.5.2 Unbalanced Design (*J* = 4)

In the unbalance design for group *J* = 4, tests were conducted with an unequal number of observations for each group, and with unequal variance across groups. With reference to Table 4.8, the results suggest that *WMOM-H* generates better power rate compared to the original procedure for all the conditions tested. The highest rank of power rate for the *J* = 4 unbalanced design belonged to Kruskal-Wallis (*KW*), followed by *ANOVA*, *WMOM-H* and *MOM-H* respectively, with an overall average power rate of 0.6351, 0.3722, 0.2597 and 0.1813 respectively.

With regards to the average summary by distribution, *KW* has the highest power rate for all of the tested distributions. The proposed procedure was found to have better performance across all designs compared to *MOM-H*. At the same time, it also performed better than *ANOVA* under positive pairing in extreme and moderate variances heterogeneity across all types of distributions. Compared to *KW*, *WMOM-H* has better power rate in the case of moderate variance heterogeneity across groups with positive pairing for Chi-Square distributions.

Table 4.8

Empirical power rate for unbalanced design, J = 4

		Robust		Parametric	Nonparametric
Distribution	**Pairing**	***MOM-H***	***WMOM-H***	***ANOVA***	***KW***
Normal	*+ve*	0.2878	0.4324	0.2714	0.9980
	-ve	0.2046	0.2668	0.8696	0.9994
	m +ve	0.4374	0.6178	0.4084	0.6204
	m -ve	0.2920	0.3842	0.8160	0.6984
	Average	***0.3055***	***0.4253***	***0.5914***	***0.8291***
Chi-square	*+ve*	0.0726	0.1486	0.1464	0.4342
	-ve	0.1174	0.1204	0.3692	0.7296
	m +ve	0.0698	0.1682	0.1144	0.0754
	m -ve	0.0970	0.1114	0.2734	0.3824
	Average	***0.0892***	***0.1372***	***0.2259***	***0.4054***
***g* = *h* = 0.5**	*+ve*	0.1496	0.2278	0.0706	0.8020
	-ve	0.1340	0.1724	0.3284	0.9054
	m +ve	0.1702	0.2924	0.5600	0.4156
	m -ve	0.1426	0.1736	0.2390	0.5606
	Average	***0.1491***	***0.2166***	***0.2995***	***0.6709***
Grand Average		***0.1813***	***0.2597***	***0.3722***	***0.6351***

The results for the unbalanced design with *J* = 4 generally shows that the proposed procedure has better power rate compared to *MOM-H*. Additionally, it demonstrated better performance for positive pairing compared to negative pairing, for all types of distributions. *WMOM-H* is the most powerful among the four tested procedures under positive pairing with moderate heterogeneity variance design for Chi-Square distributions.

4.6 Real Data Analysis

In this section, two different sets of real data on medical manufacturing were applied on the proposed procedure (*WMOM-H*), the original procedure (*MOM-H*), parametric and nonparametric procedures for comparing groups. The first set of the data consist of two groups measurement from Supplier Quality Engineering (SQE) while the second set of the data consist of four group of measurement from Research and Development (R&D). Thus, the discussion is split in two different subsections, 4.6.1 and 4.6.2 representing two (J = 2) and more than two groups (J = 4) respectively. Some basic exploratory analyses were also performed before testing the stated procedures.

4.6.1 Real Data Analysis (*J* = 2)

The source of data for J = 2 is from the Supplier Quality Engineering (SQE) department in a medical product manufacturing industry. The data were collected through an electrical test with intensity, dB as the measurement output. The SQE department need to determine whether there is any difference in performance for the new batch testing head compared to the currently use batch. In the measurement process, 20 and 22 units of testing head from the currently used batch and the new batch has been measured respectively. According to Table 4.9, data of current engineering used batch testing head was non-normally distributed while data of new batch testing head was approximate normally distributed with unequal variance across groups.

Table 4.9

Descriptive statistic of real data with J = 2

		Batch of Testing Head	
		Current	**New**
	Test		
Normality Test	**Shapiro-Wilk Test**	0.0381	0.6316
Equal Variance Test	**Levene's Test**	0.0022	
	Statistic		
	N	20	22
	Mean	0.0935	-0.0436
	Median	0.1750	-0.0600
	Standard Deviation	0.2383	0.1386

In order to test whether there is a difference across 2 group using the stated procedures, the null hypothesis and alternative hypothesis was set as follows:

$$H_0 : \theta_{current} = \theta_{new}$$

$$H_1 : \theta_{current} \neq \theta_{new}$$

Table 4.10

p-value of procedure test on real data with J = 2

Procedure	***WMOM-H***	***MOM-H***	***t-test***	***MW***
p-value	0.0417	0.0634	0.0310	0.0606

The results after the analysis indicate that there was a statistical difference intensity, dB, for Group 1 and Group 2 detected by *WMOM-H* and *t*-test. However, *MOM-H* procedure and *MW* unable to detect the differences at 5 % level of significance. The *p*-value generated from the *WMOM-H* procedure was 0.0417 while 0.0310 for *t*-test. With reference to Table 4.2, the real data matched the design of extreme skewed and heavy tailed distributions with moderate change of variances (negative paring), "*m - ve*". According to the simulation results in Table 4.2, the *t*-test showed better control of Type I error rate compared to *WMOM-H* and this match with our real data analysis results.

4.6.2 Real Data Analysis (*J* = 4)

The source of data for *J* = 4 was from the Research and Development (R&D) department in a health product manufacturing industry. In their new product development, there exists a product with different designs, and their engineering team desires to determine whether there is a difference in intensity, dB for the design with different parameter setting. Due to the time limitation in achieving the timeline, they only manage to produce and measure 25, 16, 26 and 6 units of products for design 1, 2, 3 and 4 respectively. According to Table 4.11, design 3 and 4 are normally distributed, but this is not the case for design 1 and 2. Besides, equal variance tests show that there are equal variances across the groups.

Table 4.11

Descriptive statistic of real data with J = 4

		Design			
		1	**2**	**3**	**4**
Test					
Normality Test	**Shapiro-Wilk Test**	0.0303	0.0274	0.3808	0.1705
Equal Variance Test	**Levene's Test**	0.5375			
Statistic					
N		25	16	26	6
Mean		0.1612	0.1431	0.1826	0.3283
Median		0.1800	0.1650	0.1900	0.3550
Standard Deviation		0.0947	0.1409	0.1070	0.0673

In order to test whether there is a difference across groups for J = 4 using the stated procedures, the null and alternative hypotheses were set as follows:

$$H_0: \theta_1 = \theta_2 = \theta_3 = \theta_4$$

$$H_1: At\ least\ \ one\ \theta_j\ different\ from\ the\ others$$

Table 4.12

p-value of procedure test on real data with J = 4

Procedure	***WMOM-H***	***MOM-H***	***ANOVA***	***KW***
p-value	0.0000	0.0050	0.0060	0.0120

The results of the analysis show that there is a significant difference in average intensity, dB, across groups. The *p*-value generated are less than 0.05 for all the tested procedures. From Table 4.12, the *WMOM-H* procedure generate the lowest *p*-value which is 0.000, among other tested procedures. This indicates that the proposed procedure shows better detection across the 4 groups compared to other procedures.

CHAPTER FIVE

CONCLUSION

5.1 Introduction

Traditional parametric procedures are known to falter at correctly providing an accurate test result when the required assumptions are not fulfilled. Assumptions such as data are normally distributed and variances are equal must be fulfilled so that the parametric procedures can produce convincing results with no inflation of Type I error rate, simultaneously increasing the power to detect differences. Although nonparametric procedures can be the alternative when the aforementioned problems occur, the main weakness of these procedures include loss of information because of the ranking used in the measurement and also the demand for a larger sample size to reject any false hypothesis. Thus, this study aims to overcome these problems by suggesting a new robust location estimators known as winsorized Modified One-step *M*-estimator (*WMOM*) on *H*-statistic to compare groups. As highlighted in the earlier chapters, the proposed estimator originated from Modified One-step *M*-estimator (*MOM*), an asymmetric trimmed mean, which is winsorized to produce a better robustness effect.

As presented in Chapter 4, the proposed procedure, *WMOM-H*, has been compared with *MOM-H* and their parametric and nonparametric counterparts in terms of Type I error rate and power rate of test across different investigated designs. The proposed procedure, *WMOM-H*, and the original *MOM-H* procedure were simulated 5000 times with a significant level of 0.05; and bootstrap method was employed to test the hypothesis. To compare the robustness of the procedure under different investigation conditions, Bradley’s (1978) liberal criterion of robustness was applied, based on the criterion, a test with a 5% significant level should produce a Type I error rate between 0.025 and 0.075 for the procedure to be considered robust under a particular condition. In addition, we also compared the power rate of the test to assess for any improvement in the power rate of *MOM-H* to W*MOM-H* under various conditions.

5.2 Performance comparison between *MOM-H* and *WMOM-H*

In terms of Type I error rate, we can observe that the proposed procedure, the winsorized approach, *WMOM-H*, showed an outstanding performance compared to the original procedure, *MOM-H*. Based on Table 5.1, 70% of the conditions test under *WMOM-H* produced the smallest disparity with the nominal level, 0.05, while only 30% produced by *MOM-H*. In addition, the robustness improved by 17% (from 80% to 97%) using the proposed *WMOM-H* procedure, whereby, these 17% were from those investigated design that unable performed well in original procedure, *MOM-H* (Type I error rate below 0.025).

Table 5.1

Overall summary of Type I error rate for MOM-H and WMOM-H

Comparison Criteria	*MOM-H*	*WMOM-H*
Minimal Delta Relative to 0.05	30%	70%
0.025 ≤ Type I error rate ≤ 0.075	80%	97%
Type I error rate < 0.025	20%	3%
Type I error rate > 0.075	0%	0%

In reference to Table 5.1, the result shows that the 30% *WMOM-H* with delta Type I error rate higher than *MOM-H* was from an unbalanced design for both the group sizes of two and four. In addition, these involve both positive and negative pairings with extreme variance heterogeneity from Normal and moderately skewed (Chi-square) distribution; with the exception of the group of two that has a moderately skewed distribution, the negative pairing in moderate variance heterogeneity also generated a higher delta value compared to the original procedure.

Nevertheless, these 30% designated condition still fall within Bradley’s criteria of robustness when the test was run with the proposed procedure, *WMOM-H*. Moreover, *WMOM-H* outperformed the extremely skewed and heavy tailed (g = h = 0.5) distribution in any paring and variance heterogeneity design for all group sizes. In regard to the balanced design, *WMOM-H* showed significant improvement (robustness) as compared to *MOM-H*, however, the Type I error rate for the four groups with extremely skewed and heavy tailed (g = h = 0.5) distribution did not fall within Bradley’s criteria of robustness. Nonetheless, the Type I error rate was improved from 0.0098 (*MOM-H*) to 0.0238 (*WMOM-H*).

In terms of the power rate of the test, *WMOM-H* consisted of the largest portion (nearly 97.3%) with the highest power rate of the test compared to the original procedure, *MOM-H*, with reference to Table 5.2. In other words, *WMOM-H* had shown improved robustness in regard to the power rate of the test.

Table 5.2

Overall summary of power rate of test for MOM-H and WMOM-H

Comparison Criteria	***MOM-H***	***WMOM-H***
Higher Power Rate of Test	7%	93%

The 7% higher power rate of the *MOM-H* procedure involved 2 designed conditions, two groups unbalanced design with moderately skewed distribution, and extremely skewed and heavy tailed distributions. In addition, these 2 designed conditions include negative paring and extreme variance heterogeneity. Although *WMOM-H* did not perform well performed for these 2 designs, the difference was very minimal, which was less than 7% difference. On the other hand, the improvement of the power rate of the test for *WMOM-H* throughout the investigated conditions ranged from 2.56% to 140.97%.

In conclusion, the proposed procedure, *WMOM-H*, showed outstanding performance compared to the original procedure, *MOM-H. WMOM-H* improved the robustness of the procedure for both Type I error rate and the power rate of the test.

5.3 Performance comparison between *WMOM-H* and Traditional Procedures

In this section, the proposed procedure is compared with the respective traditional procedure. The parametric procedure is represented by *t*-test and *ANOVA* while the nonparametric procedure is represented by Mann-Whitney and Kruskal-Wallis for two and four groups respectively.

Table 5.3

Overall summary of Type I error rate for WMOM-H and traditional procedures

Comparison Criteria	***WMOM-H***	**Parametric Procedure**	**Nonparametric Procedure**
Minimal Delta Relative to 0.05	53%	13%	37%
0.025 ≤ Type I error rate ≤ 0.075	97%	50%	67%
Type I error rate < 0.025	3%	17%	0%
Type I error rate > 0.075	0%	33%	33%

It can be observed from Table 5.3 that the proposed procedure has the highest percentage (nearly 53%) of Type I error rate with minimal difference to the nominal level, 0.05, across the designed conditions compared to the parametric (13%) and nonparametric (37%) procedures. This indicates that the proposed procedure *WMOM-H* was more robust than the traditional procedures. Moreover, each of the procedure, when tested across different designation conditions showed that the *WMOM-H* procedure was able to fulfill Bradley's criterion by 97% across different designs but not for parametric and nonparametric procedures, which can only achieve 50% and 67% respectively. Both the parametric and nonparametric procedures generated 33% liberal Type I error rate (> 0.075 of the Bradley's interval) each while *WMOM-H* was clean from this should be avoided situation. Meanwhile, the comparison for the lower bound interval, the parametric procedure generated the highest with 17% followed by only 3% from *WMOM-H* and none for the non-parametric procedure.

The outstanding performance of the *WMOM-H* procedure was largely attributed to its strength in generating a Type I error rate that was close to the nominal level for unbalanced design across different parings and variance homogeneity.

Table 5.4

Overall summary of power rate of test for WMOM-H and traditional procedures

Comparison Criteria	***WMOM-H***	**Parametric Procedure**	**Nonparametric Procedure**
Highest Power Rate of Test	20%	20%	60%

In terms of the power rate of the test, *WMOM-H* only generated 20% of the highest power rate compared to traditional procedures. There was no specific pattern in the results for the power test, but the percentage of generating highest power rate seemed to be inclined to the nonparametric procedure.

In conclusion, although the nonparametric procedure, generated better power rate, the *WMOM-H* procedure still shows outstanding performance compared to traditional procedures, as it was able to control Type I error rate within the Bradley's robustness criterion across differently designed tests. When testing for robustness, the ability to control Type I error rate is the utmost importance (Hayes, 2005).

5.4 Implication

The aim of this research is to develop a new robust procedure in testing the equality of the central tendency measure, which is able to control the Type I error rate and improve the power rate of the test as well. The developed procedure, *WMOM-H,* was proven to be robust compared to its original procedure, *MOM-H,* and traditional procedures (parametric and nonparametric procedures). Investigation on the robustness showed that 97% of the Type I error rate for *WMOM-H* fulfilled the robustness criterion from a total of 30 investigated conditions. *WMOM-H* performed well for both J = 2 and J = 4 in balanced and unbalanced sample sizes. However, the procedure did not perform as expected for J = 4 balanced design under extremely skewed and heavy tailed distributions generating Type I error rate of 0.0238. Nevertheless, the original procedure, *MOM-H*, also failed to fulfill Bradley's robust criterion for this condition.

For the power rate of the test, *WMOM-H* improved the power rate by 93% throughout the investigated design compared to its original procedure *MOM-H*. Although *WMOM-H* yielded a lower value compared to *MOM-H* for the 2 investigated designs, the difference was very minimal, that was less than 7%. In regards to comparison with the traditional procedure, although the highest power rate of the test percentage for *WMOM-H* was only 20% compared to the parametric and nonparametric procedures, the 20% correspond to those investigated design that achieved low power rate in traditional procedures.

Overall, *WMOM-H* performed well compared to the other investigated procedures. Although the power rate of the test for this procedure was inconsistent across the investigated designs, the *WMOM-H* was still able to consistently controlled its Type I error rate across different designs. Moreover, this procedure was also able to generate better power rate of the test value by 20% out of the 30 investigated designs compared to traditional procedure.

5.5 Suggestion for Future Research

As mentioned in Section 5.1, parametric procedures would only have performed well when assumptions are fulfilled and the nonparametric procedure requires a large sample size collection to reject false hypothesis. Thus, the focus of this study is to generate a procedure which can perform well across all types of conditions, without worrying about the assumptions.

Our study has proved that the proposed procedure, *WMOM-H*, was able to control its Type I error rate consistently within Bradley's robust criterion. However, a lower and inconsistent power rate across investigation designs was the weakness of this procedure. At any rate, since *WMOM* has the strength to consistently control the Type I error rate across different designs, future researchers should investigate *WMOM* as one of the central tendency measure for post-hoc test and even extend this study other discipline such as in in Statistical Process Control (*SPC*), especially for machines or testers that consistently generate outliers across time before preventive maintenance is performed.

As for *H*-statistic, even though *WMOM-H* has significantly improved robustness compared to *MOM-H*, other winsorized central tendency measures such as adaptive winsorized mean should also be considered as a replacement to the current *MOM* or *WMOM* procedures, so that a more consistent increase in the power rate of the test can be achieved.

REFERENCES

Abdullah, S., Syed Yahaya, S. S., & Othman, A. R. (2011). Modified Alexander-Govern test as alternative to t-test and ANOVA F test. *Sains Malaysiana, 40*(10), 1187-1192.

Ahmad Mahir, R.., & Al-Khazaleh, A. M. H. (2009). New method to estimate missing data by using the asymmetrical winsorized mean in a time series. *Applied Mathematical Sciences, 3*(35), 1715 – 1726.

Alan, O., Phyllis, S., & John, Q. (2008). *The importance of teaching power in statistical hypothesis testing*. Paper session presented at the Northeast Decision Sciences Annual Meeting, Brooklyn, New York.

Alexander, R. A. & Govern, D. M. (1994). A new and simpler approximation for ANOVA under variance heterogeneity. *Journal of Educational Statistic, 19*(2), 91 – 101. doi: 10.3102/10769986019002091

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

Beran, R. (1986). Simulated power functions. *The Annals of Statistics, 14*(1), 151-173.

Box, G. E. P. (1953). Non-normality and tests of variances. *Biometrika, 40*(3/4), 318 – 355. doi: 10.2307/2333350

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. *Annals of Mathematical Statistics*, *25*(2), 290 – 302.

Bradley, J. V. (1968). *Distribution-free statistical tests*. Englewood Cliffs, NJ: Prentice Hall.

Bradley, J. V. (1978). Robustness?. *British Journal of Mathematical and Statistical Psychology*, *31*(2), 144-152.

Chernick, M. R. (2008). *Bootstrap methods: a guide for practitioners and researchers* (2nd ed.). Newtown, PA: Wiley-Interscience.

Clark-Carter, D. (1997). The account taken of statistical power in research. *British Journal of Psychology, 88*(1), 71-83. doi: 10.1111/j.2044-8295.1997.tb02621.x

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1992a). A power primer. *Psychological Bulletin*, *112*(1), 155-159. doi: 10.1037/0033-2909.112.1.155

Cohen, J. (1992b). Statistical power analysis. *Current Directions in Psychological Science*, *1*(3), 98-101.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their applications. (Cambridge series in statistical and probabilistic mathematics)*. United States of America: Cambridge University Press.

Dixon W. J. (1960). Simplified estimation from censored normal samples. *The Annals of Mathematical Statistics, 31*(2), 385-391.

Dixon W. J., & Tukey J. W. (1968). Approximate behavior of the distribution of winsorized t (trimming/winsorization 2). *Technometrics*, *10*(1), 83-98. doi: 10.2307/1266226

Efron, B., & Tibshirani, R. J. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, *1*(1), 54 – 77.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. United States of America: Chapman & Hall Inc.

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods. *American Psychologist, 63*(7), 591-601.

Fan, W., & Hancock, G. R. (2012). Robust means modeling: an alternative for hypothesis testing of independent means under variance heterogeneity and nonnormality. *Journal of Educational and Behavioral Statistics*, *37*(1), 137–156. doi: 10.3102/1076998610396897

Guo, Jiin-Huarng & Luh, Wei-Ming (2000). An invertible transformation two-sample trimmed t-statistic under heterogeneity and nonnormality. *Statistic & Probability Letters. 49*(1), 1-7. doi:10.1016/S0167-7152(00)00022-5

Haddad, F. S., Syed Yahaya S. S., & Alfaro J. L. (2013). Alternative Hotelling's T^2 charts using winsorized modified one-step M-estimator. *Quality and Reliability Engineering International*, *29*(4), 583-593. doi: 10.1002/qre.1407

Hall, P. (1986). On the number of bootstrap simulations required to construct a confidence interval. *The Annals of Statistics*, *14* (4), 1453-1462.

Hall, P., & Padmanabhan, A. R. (1992). On the bootstrap and the trimmed mean. *Journal of Multivariate Analysis*, *41*(1), 132-153. doi:10.1016/0047-259X(92)90062-K

Hampel, F. (1968). *Contribution to the Theory of Robust Estimation*. (Ph.D. thesis). University of California, Berkeley.

Hayes, A. F. (2005). *Statistical methods for communication science*. New Jersey: Lawrence Erlbaum Associates, Inc.

Hoaglin, D. C. (1985). Summarizing shape numerically: The g-and h-distributions. In D. Hoaglin, F. Mosteller & J. Tukey (Eds), *Exploring data tables, trends, and shapes* (pp. 461–513). New York, NY: Wiley.

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*(348), 909-923.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics, 35*(1), 73 – 101. doi:10.1214/aoms/1177703732

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

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., …Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, *68*(3), 350-386. doi: 10.3102/00346543068003350

Keselman, H. J., Wilcox, R. R., Lix, L. M., Algina, J., & Fradette, K. (2007). Adaptive robust estimation and testing. *British Journal of Mathematical and Statistical Psychology*, *60*(2), 267–293. doi: 10.1348/000711005X63755

Keselman, H. J., Wilcox, R. R., Othman, A. R., & Fradette, K. (2002). Trimming, transforming statistics, and bootstrapping: circumventing the biasing effects of heteroscedasticity and non-normality. *Journal of Modern Applied Statistical Methods, 1*(2), 288 – 399.

Keselman, H. J., Othman, A. R., Wilcox, R. R., & Fradette, K. (2004). The new and improved two-sample t-test. *American Psychological Society, 15*(1), 57-51.

Keselman, H. J., Algina, J., Lix, L., Wilcox, R. R., & Deering, K. (2008). A generally robust approach for testing hypotheses and setting confidence intervals for effect sizes. *Psychological Methods, 13*(2), 110-129. doi: 10.1037/1082-989X.13.2.110.

Kulinskaya, E., Staudte, R. G., & Gao, H. (2003). Power approximations in testing for unequal means in a one-way ANOVA weighted for unequal variances. *Communications in Statistics - Theory and Method, 32*(12), 2353-2371. doi: 10.1081/STA-120025383

Lix, L. M., & Keselman, H. J. (1998). To trim or not to trim: tests of location equality under heteroscedasticity and non-normality. *Educational and Psychological Measurement, 58*(3), 409-442. doi: 10.1177/0013164498058003004

Manly, B. F. J. (2007). *Randomization, bootstrap and Monte Carlo methods in biology (3rd ed.).* Boca Raton, FL: Chapman & Hall/CRC.

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

Md Yusof, Z., Abdullah, S., & Syed Yahaya, S. S. (2012a). Type I error rate of parametric, robust and nonparametric methods for two group cases. *World Applied Sciences Journal*, *16*(12), 1817-1819.

Md Yusof, Z, Abdullah, S., Syed Yahaya, S. S., & Othman, A. R. (2012b), A robust alternative to the t-test. *Modern Applied Science, 6*(5), 27-33. doi:10.5539/mas.v6n5p27

Mendes, M., & Akkartal, E., (2010). Comparison of ANOVA F and WELCH tests with their respective permutation versions in terms of Type I error rate and test power. *Kafkas Univ Vet Faj Derg, 16*(5). 711-716.

Mendes, M., & Yigit, S. (2012), Comparison of ANOVA-F and ANOM tests with regard to Type I error rate and test power. *Journal of Statistical Computation and Simulation*, *83*(11), 1-12. doi: 10.1080/00949655.2012.679942

Murphy, K. R., Myors, B. & Wolach, A. (2008). *Statistical power analysis: a simple and general model for traditional and modern hypothesis tests (3rd ed.)*. New York: Routledge.

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

Reed, J. F., & Stark, D. B. (1996). Hinge estimators of location: robust to asymmetry. *Computer Methods and Programs Biomedicine*, *49*(1), 11-17.

Rivest, L. P. (1994). Statistical properties of Winsorized means for skewed distributions. *Biometrika, 81*(2), 373-383. doi: 10.2307/2336967

Rogan, J. C., & Keselman, H.J. (1977). Is the ANOVA F-test robust to variance heterogeneity when sample sizes are equal? An investigation via a coefficient of variation. *American Educational Research Journal*, *14*(4), 493 –498. doi: 10.3102/00028312014004493

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

Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, *88*(424), 1273-1283. doi: 10.1080/01621459.1993.10476408

Sawilowsky, S.S. (1990). Nonparametric tests of interaction in experimental design. *Review of Educational Research, 60*(1), 91-126. doi: 10.3102/00346543060001091

Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and Type II error properties of the *t*-test to departures from population normality. *Psychological Bulletin, 111*(2), 352-360. doi: 10.1037/0033-2909.111.2.352

Schneider, P. J., & Penfield, D. A. (1997). Alexander and Govern's approximation: providing an alternative to ANOVA under variance heterogeneity. *The Journal of Experimental Education, 65*(3), 271-286.

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

Schrader, R. M., & Hettmansperger, T. P. (1980). Robust Analysis of Variance Based Upon a Likelihood Ratio Criterion. *Biometrika*, *67*(1), 93-101. doi: 10.1093/biomet/67.1.93

Sharma, D., & Kibria, B. M. G., (2012). On some test statistics for testing homogeneity of variances: a comparative study. *Journal of Statistical Computation and Simulation,* *83*(10), 1-20. doi: 10.1080/00949655.2012.675336

Siegel, S. (1957). Nonparametric statistics. *The American Statistician, 11*(3), 13-19.

Snedecor, G. W., & Cochran, W. G. (1980). *Statistical methods (7th ed.).* Ames, IA: Iowa University Press.

Spector, P. E. (1993). *SAS programming for researchers and social scientists*. Newbury Park : Sage Publication Inc.

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

Syed Yahaya, S. S., Othman, A. R. & Keselman, H. J. (2004a). Testing the equality of location parameters for skewed distributions using S1 with high breakdown robust scale estimators. In M. Hubert, G. Pison, A. Struyf, & S. Van Aelst (Eds.), *Theory and Applications of Recent Robust Methods, Series: Statistics for Industry and Technology* (pp. 319-328). Basel: Birkhauser.

Syed Yahaya, S. S., Othman, A. R., & Keselman, H. J. (October, 2004b). *An alternative approach for testing location measures in the one-way independent group design*. Paper presenting session of the International Conference on Statistics and Mathematics and Its Applications in the Development of Science and Technology. Bandung, Indonesia.

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

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

Tomarkin, A. J., & Serlin, R. C. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin. 99*(1), 90 – 99.

Tukey J. W., & McLaughlin D. H. (1963). Less vulnerable confidence and significance procedures for location based on a single sample: trimming/winsorization 1. Sankhyā: *The Indian Journal of Statistics, Series A, 25*(3), 331-352.

Welch, B. L. (1951). On the comparison of several mean values: an alternative approach. *Biometrika, 38*(3/4), 330-336.

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

Wilcox, R. R. (1994). A one-way random effects model for trimmed means. *Psychometrika, 59* (3), 289-307.

Wilcox, R. R. (1995). ANOVA: the practical importance of heteroscedastic methods, using trimmed means versus means, and designing simulation studies. *British Journal of Mathematical and Statistical Psychology, 48*(1), 99-114. doi: 10.1111/j.2044-8317.1995.tb01052.x

Wilcox, R. R. (1997). *Introduction to robust estimation and hypothesis testing (3rd ed.).* New York: Academic Press.

Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods?. *American Psychologist, 53*(3), 300–314. doi: 10.1037//0003-066X.53.3.300

Wilcox, R. R. (2003). *Applying contemporary statistical techniques*. San Diego: Academic Press.

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing (3rd ed.).* New York: Academic Press.

Wilcox, R. R., Charlin, V. L., & Thompson, K. L. (1986). New Monte Carlo results on the robustness of the ANOVA F, W and F* statistics. *Communications in Statistics-Simulations, 15*(4), 933-943. doi: 10.1080/03610918608812553

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

Wilcox, R.R., & Keselman, H. J. (2003a). Modern robust data analysis methods: measures of central tendency. *Psychological Methods, 8*(3), 254–274. doi: 10.1037/1082-989X.8.3.254

Wilcox, R. R., & Keselman H. J. (2003b). Repeated measures ANOVA based on a modified one-step M-estimator. *Journal of British Mathematical and Statistical Psychology*, *56*(1), 15 – 26. doi: 10.1348/000711003321645313

Wilcox, R. R., Keselman, H. J., & Kowalchuk, R. K. (1998). Can test for treatment group equality be improved? The bootstrap and trimmed means conjecture. *British Journal of Mathematical and Statistical Psychology, 51*(1), 123-134. doi: 10.1111/j.2044-8317.1998.tb00670.x

Wilcox, R. R., Keselman H. J., Muska, J., & Cribbie, R. (2000). Repeated measures ANOVA: Some new results on comparing trimmed means and means. *British Journal of Mathematical and Statistical Psychology, 53*(1), 69-82.

Yang, K., Li, J., & Gao, H. (2006). The impact of sample imbalance on identifying differentially expressed genes. *BMC Bioinformatics, 7*(4), S8. doi: 10.1186/1471-2105-7-S4-S8

APPENDIX A

SAS/IML Programming for *WMOM-H*

```
***USING THE MOM ESTMATOR ON THE H STATISTIC***;
OPTIONS PS=40;
OPTIONS NOCENTER;
PROC IML;
RESET NONAME;

**PREPARING DATA FOR CALCULATING WMOM-ESTIMATOR**;
(Please Refer To Author If Need Full Programming)
START DATAMOD(Y, CRIT, YMAT)  GLOBAL (NX, NTOT, WOBS, BOBS);
NTOT = NROW(Y);
WOBS = NCOL(Y);
BOBS = NCOL(NX);
YT = J(NTOT, WOBS, 0);
GMAD = J(WOBS, BOBS, 0);
GMED = J(WOBS, BOBS, 0);
F = 1;
M = 0;
DO I = 1 TO BOBS;
.
.
.
.
.
.
FINISH;

**VARIABLE WINSORIZING BASED ON CRITERIA VECTOR**;
(PLEASE REFER TO AUTHOR IF NEED FULL PROGRAMMING)
START WINSMOD(YMAT, CRIT, WINSOR, MUBARM, H) GLOBAL(NX, NTOT, WOBS,
BOBS);
WINSOR = J(WOBS, BOBS, 0);
F = 1;
M = 0;
.
.
.
.
.
.
FINISH;

**FINDING THE P-VALUE OF THE H STATISTIC REQUIRES BOOTSTRAP**;
**GENERATING BOOTSTRAP SAMPLE**;
(PLEASE REFER TO AUTHOR IF NEED FULL PROGRAMMING)
START BOOTDAT(Y, WINSOR, YB) GLOBAL(NX, NTOT, WOBS, BOBS, SEED);
F = 1;
M = 0;
.
.
.
.
```

```
.
.
FINISH;

**CALCULATING BOOTSTRAP H STATISTIC**;
(PLEASE REFER TO AUTHOR IF NEED FULL PROGRAMMING)
START BOOTSTAT(YB, HB) GLOBAL(NX, NTOT, WOBS, BOBS, SEED);
.
.
.
.
.
.
FINISH;

*********TRIAL RUN ON BOOTSTRAPPING WITH GENERATED DATA***********;
SSEED=439839383;
CPOPVAR = {1 1 1 1};
CNX = {20 20 20 20};
CPOPMN = {0 0 0 0};
CN = CNX[,+];
COND = NROW(CPOPVAR);
NSIM = 5000;
F = 1;

**NUMBER OF BOOTSTRAP SAMPLES**;
NUMSIM = 599;
**SEED FOR BOOTSTRAPPING**;
SEED = 40389;

COUNTER = 0;
ALPHA = 0.05;

****GENERATE DATA FOR CONDITIONS****;
(PLEASE REFER TO AUTHOR IF NEED FULL PROGRAMMING)
DO K = 1 TO NSIM;
  DO I = 1 TO COND;
.
.
.
.
.
.
    RUN WMOM1;
    IF (RESULTS[2] <= ALPHA) THEN COUNTER = COUNTER + 1;
  END; *DO I;
END; *DO K;

DO I = 1 TO COND;
   V = CPOPVAR[I,];
   S = CNX[I,];
   M = CPOPMN[I,];
   COUNT = COUNTER/NSIM;
   PRINT 'STUDY CONDITIONS ARE:';
   PRINT 'ALPHA IS:' ALPHA[FORMAT = 5.2];
   PRINT 'GROUP POPULATION VARIANCES:' V[FORMAT = 4.0];
   PRINT 'GROUP SAMPLE SIZES:' S[FORMAT = 4.0];
   PRINT 'GROUP MEANS:' M[FORMAT = 4.0];
   PRINT 'TEST FOR:4pemmn' COUNT[FORMAT = 6.5];
END; *DO I;
```