The copyright © of this thesis belongs to its rightful author and/or other copyright owner. Copies can be accessed and downloaded for non-commercial or learning purposes without any charge and permission. The thesis cannot be reproduced or quoted as a whole without the permission from its rightful owner. No alteration or changes in format is allowed without permission from its rightful owner.

UUM
Universiti Utara Malaysia

DEVELOPMENT AND MODIFICATION OF *H*-STATISTIC WITH WINSORIZED APPROACH MEANS

Universiti Utara Malaysia

TEH KIAN WOOI

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)

TEH KIAN WOOI

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

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

"MODIFIED H-STATISTIC WITH WINSORIZED APPROACH MEANS"

UUM
Universiti Utara Malaysia

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 : ***19 Januari, 2017.***

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: ***January 19, 2017.***

Pengerusi Viva: **Assoc. Prof. Dr. Rahela Abdul Rahim** Tandatangan
(Chairman for VIVA) *(Signature)*

Pemeriksa Luar: **Assoc. Prof. Dr. Zulkifley Mohamed** Tandatangan
(External Examiner) *(Signature)*

Pemeriksa Dalam: **Dr. Nor Aishah Ahad** Tandatangan
(Internal Examiner) *(Signature)*

Nama Penyelia/Penyelia-penyelia: **Dr. Suhaida Abdullah** Tandatangan
(Name of Supervisor/Supervisors) *(Signature)*

Nama Penyelia/Penyelia-penyelia: **Assoc. Prof. Dr. Zahayu Md Yusof** Tandatangan
(Name of Supervisor/Supervisors) *(Signature)*

Tarikh:
(Date) ***January 19, 2017***

Permission to Use

In presenting this thesis in fulfilment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the Universiti Library may make it freely available for inspection. I further agree that permission for 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

UUM Universiti Utara Malaysia

Abstrak

Ujian-*t* pelajar dan ujian-*F* ANOVA adalah ujian statistik klasik untuk membandingkan dua atau lebih kumpulan bebas. Kedua-duanya adalah ujian yang berkuasa apabila data tertabur normal dan mempunyai varians homogen. Walau bagaimanapun, data dengan pencirian tersebut adakalanya sukar untuk dipenuhi dalam kehidupan sebenar dan akan memberi kesan kepada kawalan kadar ralat Jenis I dan mengurangkan kuasa ujian statistik tersebut. Statistik-*H* adalah statistik teguh namun hanya mampu menunjukkan prestasi yang baik hanya pada set data tidak normal. Statistik ini telah diinovasikan dengan penganggar *MOM* dan ditandai sebagai *MOM-H*. Oleh yang demikian, dalam kajian ini, dua statistik-*H* terubah suai dengan min menggunakan pendekatan ter*Winsor* adalah dicadangkan untuk menangani ketidakpatuhan kedua-dua pencirian tersebut. Statistik yang dicadangkan adalah statistik-*H* dengan min ter*Winsor* (*WM*) dan statistik-*H* dengan min ter*Winsor* suai (*AWM*) yang masing-masing ditandai sebagai *WM-H* dan *AWM-H*. Menggunakan pengubahsuaian ini, prestasi ujian lebih baik bukan sahaja pada ketidaknormalan, tetapi juga pada keheterogenan varians. Pendekatan ini menggunakan nilai awal iaitu 15% dan 25% nilai pe*Winsor*an Pendekatan *WM* me*Winsor* secara simetri manakala *AWM* me*Winsor* secara tersuai mengikut bentuk taburan berdasarkan penganggar engsel, *HQ* dan HQ_1. Statistik *WM-H* terdiri daripada 15*WM-H* dan 25*WM-H*, manakala *AWM-H* terdiri daripada 15*WHQ-H*, 25*WHQ-H*, 15WHQ_1-*H* dan 25WHQ_1-*H*. Prestasi ujian yang dicadangkan adalah dinilai dengan menggunakan Kadar Ralat Jenis I dan kuasa ujian berdasarkan kajian simulasi. Semua keputusan daripada ujian yang dicadangkan dibandingkan dengan ujian statistik-*H* yang asal, *MOM-H* dan statistik klasik. Pada taburan terpencong, *WM-H* menunjukkan prestasi lebih baik berbanding dengan yang lain tetapi setanding dengan *MOM-H*. Secara keseluruhan ujian yang dicadangkan dapat memberikan hasil yang lebih baik daripada *MOM-H* dan ujian statistik klasik pada keadaan tertentu. Ujian yang dicadangkan juga ditentusahkan menggunakan set data sebenar.

Kata kunci: Pendekatan ter*Winsor*, Penganggar engsel, Kadar ralat Jenis I, Kuasa ujian, Statistik-*H*

Universiti Utara Malaysia

Abstract

Student's *t*-test and ANOVA *F*-test are the classical statistical tests for comparing two or more independent groups. Both are powerful tests when data is normally distributed and variances are homogenous. However, the data with these properties sometime is difficult to be met in real-life will affect the Type I error rates control and reduce statistical power of the tests. *H*-statistic is a robust statistic but performs well only under non-normality dataset. This statistic had been invented with *MOM* estimator denoted as *MOM-H*. Therefore, in this study, two modified *H*-statistic with mean using Winsorizing approach are proposed to handle both violated properties. The proposed statistics are the *H*-statistic with Winsorized mean (*WM*) and the *H*-statistic with adaptive Winsorized mean (*AWM*) which denoted as *WM-H* and *AWM-H*, respectively. Using this modification, the tests perform better not only under non-normality, but also under heterogeneity of variances. The approach use predetermined values of 15% and 25% Winsorization. The *WM* is Winsorizing symmetrically while the *AWM* is Winsorizing adaptively according to the shape of distribution based on hinge estimators, *HQ* and HQ_1. The *WM-H* statistic consists of 15*WM-H* and 25*WM-H*, whereas the *AWM-H* comprises of 15*WHQ-H*, 25*WHQ-H*, 15WHQ_1-*H* and 25WHQ_1-*H*. The performances of the proposed tests are evaluated using Type I error rates and power of test based on simulation study. All the results from the proposed tests are compared with the original *H*-statistic, *MOM-H* and classical statistical tests. The findings indicate that 15*WHQ-H* performs the best for two groups case especially under heavy tailed distribution. Under skewed distribution, *WM-H* has better performance to others but comparable to *MOM-H*. In overall the proposed tests are able to give better results than the *MOM-H* and the classical statistical tests under certain conditions. The proposed tests are also validated using real dataset.

Universiti Utara Malaysia

Keywords: Winsorizing approach, Hinge estimator, Type I error rates, Power of test, *H*-statistic

Acknowledgement

This study can be successfully accomplishing, firstly, I would like to express my thousands of gratitude to my supervisor Dr. Suhaida Abdullah for her guidance, advices and continuous support in completing the thesis. I also would like to express my gratitude to my co-supervisor Dr. Zahayu Md. Yusof for her advice which help to improve my work. Another person who would like to express my gratitude to is Prof. Dr. Sharipah Soaad Syed Yahaya, who the person helping and encouraging me to pursue my postgraduate degree in Master of Science (Statistics). Without the guidance, advices and help from them, my thesis would not have been possible.

I would like also to thank my family, course mates and friends, for their encouragement, moral and spiritual support. Finally, I would like to express my appreciation to Univerisiti Utara Malaysia for providing a chance for me to pursue my postgraduate degree.

UNIVERSITI UTARA MALAYSIA ILMU BUDI BAKTI UUM Universiti Utara Malaysia

Table of Contents

Universiti Utara Malaysia

UUM
Universiti Utara Malaysia

Universiti Utara Malaysia

UUM
Universiti Utara Malaysia

List of Tables

Universiti Utara Malaysia

UUM Universiti Utara Malaysia

List of Figures

Universiti Utara Malaysia

List of Appendices

UNIVERSITI UTARA MALAYSIA
ILMU BUDI BAKTI
UUM
Universiti Utara Malaysia

CHAPTER ONE
INTRODUCTION

1.1 Background

In the case of employing classical procedures in comparing independent groups, the normality of distribution and the homogeneity of variances among the groups are the primary concerns that will affect the analysis results. The devastating effect on controlling Type I errors rate and reducing the statistical power will happen when dispersion in these criteria occurs. (Syed Yahaya, 2005; Syed Yahaya, Othman, & Keselman, 2006; Keselman, Algina, Lix, Wilcox, & Deering, 2008). In order to deal with these violation of assumptions, the alternative procedures such as non-parametric procedure may be employed. However, the use of this procedure may cause loss of information as this procedure is testing on the ranking value rather than on the original parametric value (Siegel, 1957).

Universiti Utara Malaysia

Besides the non-parametric procedure, another common method used to deal with the violation of normality is simple data transformation. In other words, each observation of the data is transformed by taking inverse, logarithms, square roots, or other transformations, before performing test analysis (Rasmussen, 1989; Wilcox & Keselman, 2003a). Based on Rasmussen's study in 1989, an accurate transformation may provide better control of Type I error rate and increase the statistical power under more non-normal distribution. However, for the mildly skewed data or the data that have groups with skewed data in opposite directions, it may not be advantageous. Furthermore, the transformations are complicated to perform and wrong or inaccurate transformation being chosen will affect the accuracy of the analysis results.

Another alternative method is the robust statistics approaches proposed to deal with the violation of assumptions problem when classical statistical procedures are used. Robust statistics is very powerful and able to perform well in terms of Type I error rate control and maintain adequate power rate, even under non-normal and heterogeneity of variances data (Erceg-Hurn & Mirosevich, 2008). As a result, the better solution to deal with the violating assumptions in statistical test analysis is by moving forward to robust statistics that able to derive a better test and improved performance compared to those from traditional Student's *t*-test and analysis of variance (ANOVA) *F*-test.

In recent years, several robust statistical tests have been proposed to deal with the non-normal distribution and heteroscedasticity data. For example, the Welch test was developed to handle the problem of heteroscedasticity whereas the *H*-statistic was proposed to deal with the non-normality data (Welch, 1947; Welch, 1951; Othman, Keselman, Padmanabhan, Wilcox, & Fradette, 2004). However, the challenge in robust statistical tests development is to obtain a good test with better control of Type I error rate as well as able to achieve high in power under violated assumptions condition.

Universiti Utara Malaysia

The *H*-statistics was originally proposed by Schrader and Hettmansperger (1980) in which it is readily adaptable to any central tendency measure. It gives reasonably good results in comparison when using *M*-estimator, but it is not recommended for central tendency comparison that uses mean or even trimmed mean (Wilcox, 2012). Keselman, Wilcox, Othman, and Fradette (2002) replaced modified one-step *M*-estimator (*MOM*) as central tendency measure in *H*-statistic (denote as *MOM-H*) and this robust statistic was further studied by Othman et al. (2004). In their study, they found that this robust statistic is able to produce better control in Type I error rate and increase the statistical power under skewed distribution. Syed Yahaya (2005) furthers the study on *MOM-H*. She proposed the approach of *MOM-H* with robust scale estimators, which are *MADn*, *Sn*, and *Tn*. She found that *MOM-H* is well in control over probability of Type I error under normal distribution but fair under non-normal distribution.

1.2 Problem Statement

The statistical test is a tool to compare and test the equality of central tendency of two or more independent groups. However, the commonly used central tendency measure such as mean is sensitive to the presence of outliers. The presence of outliers will lead to wrong interpretation of the results of the statistical analysis. The robust central tendency measure is one of the recommended statistics used to deal with the problem of the presence of outliers by using trimming approach to trim the observations at both tails of distribution (Wilcox & Keselman, 2003a). The usual trimmed mean is one of the trimming approaches that simply removes or trims a fixed proportion on both tails (left and right) of the distribution symmetrically and obtains the mean by average the remaining data. This trimmed amount is a predetermined trimming percentage without checking for outliers beforehand (Wilcox & Keselman, 2003a). If 0% proportion of trimming is set, the result of the trimmed mean is treated as a sample mean, and the trimmed mean is equal to a median when the maximum possible amount of trimming which is 50% for each tail (Wilcox, 2003). However, the number of data to be trimmed is a question to be asked before performing the trimming process.

Universiti Utara Malaysia

In previous study, Huber (1972) suggested that the trimmed mean with the trimming percentage between 1% and 10% should be an almost efficient estimate and commented that 5% trimming percentage may provide even better on the average. He also proposed 15% trimmed mean to deal with the not-so-long-tailed distribution. The

15% of trimming percentage was also recommended by Mudholkar, Mudholkar and Srivastava (1991). This is because they found that the 15% of trimming percentage is adequate for most practical situations. Stigler (1977) suggested 10% trimmed mean for the data sets that exhibit a slight tendency toward more extreme value. Rosenberger and Gasko (1983) proposed 20% of trimming in general to achieve a relatively small standard error, and suggested 25% of trimming for small sample sizes. Wilcox and Keselman (2003b) showed that 25% to 50% trimmed mean or a proportion of trimmed mean of at least 25% will have good control over the probability of Type I error but have the potential of substantially reducing the power rate. Under normality, Wilcox (2003) proposed 20% of trimming as a reasonable default value based on small standard error achieving criteria, as well as to avoid the low power and obtain good control of Type I error rate.

Universiti Utara Malaysia

Beside the trimmed amount, the trimming process is also another factor of concern. As mentioned before, the trimming process of usual trimmed mean is symmetrical trimming and may cause the loss of important information. This trimming is performed by trimming the data symmetrically on left and right tails without checking the pattern of distribution. It simply removes the data according to predetermined trimmed proportion on both tails even under normal distribution. However, the data should not be trimmed at all when the data is normally distributed. For data with skewed distribution, the usual trimmed mean also trims the data symmetrically regardless of the shape of the data distribution. In practical, this type of data should be trimmed more on the right if the distribution is skewed to the right and vice versa.

To deal with the skewed distributed data, the asymmetric trimmed mean proposed by Hogg (1974) is applied. The asymmetric trimmed mean is similar to the usual trimmed mean, but it allows asymmetric trimming according to the shape of distribution as its strength. The asymmetric trimmed mean was modified by Keselman, Wilcox, Lix, Algina, and Fradette (2007) with the use of the hinge estimator to determine the proportion of trimming. This approach is called adaptive trimmed mean using hinge estimator. The hinge estimator determines the trimming proportion of the adaptive trimmed mean based on the shape of the distribution where the proportion of trimming for right tail is more than left tail when the distribution is skewed to the right and vice versa. The hinge estimator was proposed by Reed and Stark (1996) where seven hinge estimators was proposed. The HQ and HQ_1 are the best among the top estimators recommended by Keselman et al. (2007). They are used in adaptive trimmed mean as these two estimators provide better control of Type I error rate in Welch test.

There are two descriptions for trimming proportion or trimming percentage. First, the trimming proportion is defined as the trimmed amount for each tail and this is used in usual trimmed mean. The other description is used in Keselman et al. (2007) where the proportion of trimming is the total proportion required to trim from the sample data. As mentioned before, the adaptive trimmed mean using hinge estimator is similar to the usual trimmed mean in which it needs a predetermined trimming percentage. However, its proportion of trimming is the total trimmed proportion. In Keselman et al. (2007), the results showed that 15% and 25% of trimming provided better control of Type I error rate and higher statistical power. Under extremely non-normal distribution, 10% of trimming provided good performance in their study. On the other hand, 15% trimming using HQ_1 obtained a good control of Type I error rate and is recommended by Abdullah (2011).

Wilcox and Keselman (2003b) proposed modified one-step *M*-estimator (*MOM*) which was transformed from the Staudte and Sheather (1990) one-step *M*-estimator. This is another central tendency measure proposed to deal with skewed distributed data. The trimming process is done empirically based on the outlier detection criteria. *MOM* trims the data flagged as outliers. There is a possibility of no trimming at all or different amounts of trimming in each tail (Wilcox & Keselman, 2003b). From the study of Wilcox and Keselman (2003b), *MOM* has obtained relatively good power under normality, and has performed well and provided good control in Type I error rate when the sample size is small. *MOM* has a substantially smaller standard error, but may not be the case when comparing with mean sometimes (Wilcox, 2012). However, the outlier is very difficult to detect and failure of outlier detection will affect the power rate of a statistical test (Wilcox, 2003).

Universiti Utara Malaysia

In general, the means determined based on the trimming approaches (usual trimmed mean, adaptive trimmed mean and empirically trimmed mean) are obtained by calculating the average of the remaining data after the trimming process. These trimming procedures may cause the sample size to be reduced or always changing. Furthermore, this will cause the loss of important information when the data is discarded. The Winsorized approach mean is another statistical measurement of central tendency suggested by Charles P. Winsor as mentioned in Dixon (1960). Instead of discarding the data, the discarded data is replaced with the highest and lowest values of the remaining data, then the Winsorized mean (symmetrically) calculates by averaging all the data (Tukey & McLaughlin, 1963). Ahmad Mahir and Al-Khazaleh (2009) found that the asymmetrical Winsorized mean consistently performed better than other methods used in their study. The Winsorized approach mean preserves the sample size after the trimming process. By mention at the motivation, modified *H*-statistic with Winsorized approach mean (symmetrically and adaptively using hinge estimator) were proposed to improve the performance in controlling the Type I error rate and obtain high power under non-normal distribution, heterogeneity of variances and the unbalance of sample sizes conditions.

1.3 Objective of the Study

The main goal of this study is to modify the *H*-statistic with Winsorized approach mean in order to be applicable under various data conditions without worrying for any violation of the assumption. In order to meet this objective, several sub-objectives to be accomplished as listed below:

i. Develop the modified test statistic known as modified *H*-statistic with Winsorized mean (*WM-H*) and modified *H*-statistic with adaptive Winsorized mean (*AWM-H*) that use the Winsorized approach mean (symmetrically or adaptively) as the central tendency measure in *H*-statistic.
ii. Investigate the performance of the proposed test statistic in terms of Type I error rate and power.
iii. Compare the robustness of the test statistic against the existing procedures (*t*-test, *F*-test and *MOM-H*) under various conditions.
iv. Investigate the ability of the proposed procedures on real data (education/health/finance/manufacturing).

Universiti Utara Malaysia

1.4 Significance of the Study

The success of this study will be able to move forward the experimental design methodology in various fields. Normality of data and homogeneity of variances among treatment groups have been the main concerns of analysts when they perform their analysis as the results may be biased if the conditional assumption is not fulfilled.

Therefore, the new methodology proposed in this study will be advantageous to the researchers in various areas (especially experimental sciences). In particular, the researchers can proceed with their original data without much concern regarding the characteristic of their data.

1.5 Organization of the Thesis

There are five chapters in this study. The first chapter describes the problem faced when the statistical test uses classical procedures and reviews the alternative procedures proposed by other researchers to deal with the problem. Besides, the robust statistics and the proposed procedures are also briefly introduced in Chapter One. The procedure and the robust central tendency measures used in this study are further reviewed and described in detail with some of the past and contemporary research findings in Chapter Two. Chapter Two also explains some important terminologies such as Type I error rate, statistical power and bootstrap method.

Universiti Utara Malaysia

In Chapter Three of this study, the way to conduct the empirical investigation is described. The design and selection of the test conditions to evaluate the performance of the procedures are discussed in this chapter. This chapter also discusses the data generation of four types distribution from the standard normal. Another section in Chapter Three describes the setting of central tendency measures for the power analysis that focuses on the pattern and the effect size.

The results of the evaluation on the procedures are discussed in Chapter Four. The performance of the procedures was evaluated via the ability of Type I error rate control and statistical power performance. Chapter Five ends this study with discussion and conclusion as well as suggestions for further studies.

CHAPTER TWO
LITERATURE REVIEW

2.1 Introduction

Statistical test is a powerful tool to test the equality of the central tendency measure of the independent treatment groups. However, non-normality and heteroscedasticity are the challenges when classical statistic procedures are used. For example, analysis of variance (ANOVA) is one of the widely used classical parametric statistical tests and its ability to detect true differences would be seriously hampered when the assumptions of normality and homoscedasticity are not met (Syed Yahaya, 2005). Moreover, in real data analysis, these assumptions are hard to attain. This violation of assumptions will affect the results and lead to wrong interpretation of the data analysis when classical procedures are used (Erceg-Hurn & Mirosevich, 2008). Therefore, researchers had made effort to seek for alternative statistical procedures that are able to provide accurate results under any condition even when the assumptions are violated. This is the answer of why robust statistical test procedures are developed and keep moving forward, to alleviate the problems inherent with the violation of assumptions when parametric statistical tests are used. The robustness of the statistical tests is assessed from its ability of controlling Type I error rate and power.

Universiti Utara Malaysia

Type I error rate is also known a significance level. It is often denoted by α and defined as the probability of rejecting the null hypothesis when it is true. The null hypothesis of the central tendency measure equality test as

$$H_0 \colon \theta_1 = \theta_2 = \cdots = \theta_j,$$

where θ_i is the measure of central tendency for group i distribution, F_i: $i = 1, 2, \ldots, j$. The probability of the Type I error or the rate of Type I error (denoted as ρ) is the proportion or percentage of significant statistical test to check the equality of the central tendency measures. It should be a small value and close to the nominal (significance) value. The robustness of the statistical tests is referring to the ability of a procedure to maintain the ρ value at its nominal level, α. Based on the robustness criterion proposed by Bradley (1978), a procedure is considered robust if its ρ is within 0.5α and 1.5α. For nominal level, $\alpha = 0.05$, the interval for ρ should be from 0.025 to 0.075. The ρ of a test tends to differ from the nominal level, α when the violation assumptions occur and the null hypothesis is true (Bradley, 1978).

The power of the statistical test is the probability that the false null hypothesis is correctly rejected, that is, the probability that a statistically significant result is obtained (Cohen, 1988, 1992b). The power of statistical test is denoted by $1 - \beta$, where β is the Type II error. The Type II error rate is the probability of failing to reject the false null hypothesis. In convention, Cohen (1992a, 1992b) proposed 80% as the desired value of power for general use. However, 50% of statistical power is the minimum accepted value and values smaller than this are likely to fail (reject null hypothesis) than to succeed (Murphy, Myors & Wolach, 2008). The power of the statistical test depends on the significance level, α, the sample size, n, and the effect size, f (Cohen, 1992b).

In terms of α, the lower α value without predicted direction results in lower statistical power. While for sample size, the larger the sample size, the more precise the results and the higher the statistical power (Cohen, 1988). Cohen (1992b) defined the effect size as the difference between the null hypothesis and alternative hypothesis. The smaller effect size means smaller discrepancy between the null hypothesis and alternative hypothesis, thus results in lower power of a statistical test. Therefore, the higher statistical power can be obtained by the increment of α or sample size, or by larger effect size.

2.2 Measure of Central Tendency

The measure of central tendency is the key factor of statistical test. One of the purpose of comparing independent groups test is to testing the equality of the central tendency measure for two or more distributions. For example, the Student's *t*-test is calculated by taking the difference between two means divided by the standard error,

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \tag{2.1}$$

where $\bar{X}_i$ is the measure of central tendency for group i distribution, F_i: $i = 1, 2, \ldots, j$. S_i^2 and n_i are the group i, $i = 1, 2, \ldots, j$ variance and sample size respectively. Mean is a commonly used central tendency measure. However, it is sensitive to the presence of outliers and this will affect the results of the statistical test. Therefore, a robust measure of central tendency is required to obtain a better result of statistical test in controlling Type I error rate and higher statistical power even under the conditions of violated assumptions. The usual trimmed mean is one of the robust measures of central tendency that is commonly used. It discards the data symmetrically on both tails. Besides, there are also asymmetric trimmed mean that allows asymmetrical trimming, and empirically trimmed mean, which trims the outlier only.

Universiti Utara Malaysia

2.2.1 Usual Trimmed Mean

The usual trimmed mean (γ-trimmed mean) is one of the earlier robust central tendency measures that is used as an alternative central tendency measure in a statistical test. It is done by symmetrically trimming the smallest, the largest or both values of the data according to a percentage of trimming which is fixed without assessing the characteristic of the data distribution (Keselman et al., 2007). The mean is calculated by the average of the remaining trimmed data. For instance, $X_{ij} = (x_{1j}, x_{2j}, \dots, x_{nj})$ are a random sample in ordered and associated with the j^{th} group and γ-trimmed mean is defined as

$$m_t(\gamma) = \frac{1}{q} \sum_{i=1+k}^{n-k} X_{ij} \qquad (2.2)$$

where γ is the trimming percentage where the proportion of observations will be trimmed from each tail of the distribution. The $k = [n\gamma]$ is rounded down to the nearest integer and $q = n - 2k$, so that $q = n - 2n\gamma$.

When conducting the γ-trimmed mean, the trimming process must be done with caution due to two drawbacks of γ-trimmed mean. In particular, the quantity of the data to be trimmed and the issue of symmetrically trimming must be considered. This is to avoid the loss of important information during the process. Furthermore, the shape of the data distribution, whether normal or non-normal is unable to be identified without further assessment (Wilcox & Keselman, 2003a). When the data is skewed or is a heavy-tailed distribution, the data should be trimmed asymmetrically according to the shape of the distribution rather than trimmed symmetrically.

Universiti Utara Malaysia

2.2.2 Adaptive Trimmed Mean

The adaptive trimmed mean is an alternative robust central tendency measure that is able to deal with the drawbacks of γ-trimmed mean. The adaptive trimmed mean is done asymmetrically by referring to the characteristic of the data distribution before executing the trimming process (Keselman et al., 2007). The adaptive trimmed mean was originally proposed by Hogg (1974) and is defined as

$$m_t(\gamma_\ell, \gamma_u) = \frac{1}{q} \sum_{i=1+k_1}^{n-k_2} X_{ij} \qquad (2.3)$$

where γ_ℓ and γ_u are the lower and upper trimming percentage respectively. So, $k_1 = [n\gamma_\ell]$, $k_2 = [n\gamma_u]$ and $q = n - 2k_1 - k_2$.

Keselman et al. (2007) adopted the adaptive trimmed mean by Hogg (1974) with collaboration with hinge estimators from Reed and Stark (1996) to obtain a better measure of central tendency. The hinge estimators used in the adaptive trimmed mean suit the trimming process to the shape of data distribution. Reed and Stark (1996) defined seven hinge estimators namely *HQ*, *HQ*$_1$ and *HH*$_3$ from tail length measure, and *HQ*$_2$, *HH*$_1$, *HSK*$_2$ and *HSK*$_5$ from measure of skewness. Among these seven hinge estimators, *HQ* and *HQ*$_1$ are recommended as the best and most efficient estimators with a good control of Type I error rate in Welch test (Keselman et al., 2007; Reed & Stark, 1996).

UUM Universiti Utara Malaysia

2.2.2.1 Hinge estimator, *HQ*

The hinge estimator *HQ* was defined by Reed and Stark (1996) from the tail-length measure *Q*, which was proposed by Hogg (1974)

$$Q = \frac{U_{.05} - L_{.05}}{U_{.5} - L_{.5}} \tag{2.4}$$

where L_α is the mean of the smallest αn observations and U_α is the mean of the largest αn observations. The Q can be used to classify the symmetric distributions as light-tailed ($Q < 2$), medium-tailed ($2 \le Q \le 2.6$), heavy-tailed ($2.6 \le Q \le 3.2$) and very heavy-tailed ($Q > 3.2$) (Reed & Stark, 1996).

Reed and Stark (1996) defined *HQ* as

$$HQ = \frac{UW_Q}{UW_Q + LW_Q} \tag{2.5}$$

where $UW_Q = \frac{\sum_{j=1}^{J} n_j(U_{.05} - L_{.05})}{\sum_{j=1}^{J} n_j}$ and $LW_Q = \frac{\sum_{j=1}^{J} n_j(U_{.5} - L_{.5})}{\sum_{j=1}^{J} n_j}$. The γ is the total percentage to be trimmed from the data. So, the γ_ℓ is the lower proportion of trimming,

$$\gamma_\ell = \gamma HQ = \gamma \frac{UW_Q}{UW_Q + LW_Q} \tag{2.6}$$

and the upper proportion to be trimmed is calculated by

$$\gamma_u = \gamma - \gamma_\ell \tag{2.7}$$

From the study of Keselman et al. (2007), the *HQ* is the second best hinge estimator in controlling Type I error rate.

2.2.2.2 Hinge estimator, *HQ*1

Another hinge estimator defined by Reed and Stark (1996) based on Hogg's (1974) tail-length measure, Q_1 is HQ_1,

$$Q_1 = \frac{U_{.2} - L_{.2}}{U_{.5} - L_{.5}} \tag{2.8}$$

Q_1 is the tail-length measure proposed by Hogg (1974) and it is used for the classification of symmetric distributions as light tail distribution when Q_1 is smaller than 1.81, medium tail distribution when Q_1 is from 1.81 to 1.87 and heavy tail distribution when Q_1 is larger than 1.87. Its algorithm is similar to *HQ*, and differs from *HQ* in the criteria of UW_{Q_1} where α is equal to 0.2. The HQ_1 was defined by Reed and Stark (1996) as

$$HQ_1 = \frac{UW_{Q_1}}{UW_{Q_1} + LW_{Q_1}} \tag{2.9}$$

Keselman et al. (2007) recommended adaptive trimmed mean using HQ_1 in heteroscedastic statistic as it is able to obtain better control of Type I error rate and good power to detect the effects. Abdullah (2011) also recommended HQ_1 as it obtained good control of Type I error rate in her study.

Universiti Utara Malaysia

2.2.3 Modified one-step *M*-Estimator (*MOM*)

The modified one-step *M*-estimator (*MOM*) proposed by Wilcox and Keselman (2003b) was transformed from one-step *M*-estimator (Staudte and Sheather, 1990). It is quite similar to the adaptive trimmed mean in terms of the data trimming approach, which is based on data distribution. However, the difference in *MOM* is the trimming process that is done empirically based on the outlier detection criteria, in which X_{ij} is flagged as an outlier if

$$|X_{ij} - \widehat{M}_j| > 2.24(MADn) \tag{2.10}$$

where

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

$MADn = \frac{MAD}{0.6745}$

MAD = median of the value $|x_{1j} - \widehat{M}_j|, |x_{2j} - \widehat{M}_j|, \cdots, |x_{nj} - \widehat{M}_j|$

Universiti Utara Malaysia

MOM was defined by Wilcox and Keselman (2003b) as

$$\hat{\theta} = \sum_{i=k_{1j}+1}^{n_j-k_{2j}} \frac{X_{ij}}{n_j-k_{1j}-k_{2j}} \tag{2.11}$$

where

$X_{ij} =$ the i^{th} ordered observations in group j.

$n_j =$ number of observations for group j.

$k_{1j} =$ number of observations X_{ij} that $(X_{ij} - \widehat{M}_j) < -2.24\, MADn$

$k_{2j} =$ number of observations X_{ij} that $(X_{ij} - \widehat{M}_j) > 2.24\, MADn$

MOM is able to adapt in small sample sizes statistical test with 2.24 as suggested by Wilcox and Keselman (2003b). The proposed constant of 2.24 is due to reasonably good efficiency obtained under normality for *n*, which is less than 100. *MOM* also provides relatively good power under normality and good control of the Type I error rate under non-normal distribution and small sample size (Wilcox & Keselman, 2003b). However, the trimming process based on outlier detection criteria will lead to unsatisfactory test result when it fails to detect the outlier.

2.2.4 Winsorized Mean (*WM*)

Winsorized mean (*WM*) is another robust central tendency measure, which determines the mean by using the Winsorizing approach suggested by Charles P. Winsor. It appears similar to the usual trimmed mean, but it substitutes the trimmed values with the nearest retained values, rather than trimming them away (Tukey & McLaughlin, 1963; Dixon & Tukey, 1968). This allows the sample size to remain the same. The *WM* modified from Equation 2.2 is given as

$$m_w(\gamma) = \frac{1}{n}\sum_{i=1+k}^{n-k} X_{ij} + k(X_{1+kj} + X_{n-kj}) \qquad (2.12)$$

where

γ = the proportion of the observations from each tail of the distribution that will be replaced

$k = [n\gamma]$ = the number of substituted values taken to the nearest rounded down integer

Dixon (1960) and Rivest (1994) showed that using the *WM* as the central tendency measure under normally and skewed distributed data provided better results in their studies.

As mentioned above, the *WM* appear like the usual trimmed mean. Therefore, its Winsorizing process is similar to the trimming process. The data is winsorized symmetrically regardless of skewed or heavy-tailed distributed data. Therefore, the loss of information also occurs on *WM* when performing the Winsorizing process. The data should be winsorized adaptively according to the shape of the distribution for skewed or heavy-tailed distributed data.

Universiti Utara Malaysia

2.2.5 Adaptive Winsorized Mean (*AWM*)

To deal with the problem of symmetric Winsorizing as mentioned in previous sub-section, adaptive Winsorized mean (*AWM*) is the alternative robust central tendency measure to be used. The *AWM* is able to perform the Winsorizing process according to the shape of the distribution, either symmetrical or asymmetrical. The *AWM* is defined as

$$m_w(\gamma_\ell, \gamma_u) = \frac{1}{n}\sum_{i=1+k_1}^{n-k_2} X_{ij} + k_1 X_{1+k_1 j} + k_2 X_{n-k_2 j} \tag{2.13}$$

which is modified from Equation 2.3, where γ_ℓ and γ_u are the lower and upper Winsorizing proportions respectively. So, $k_1 = [n\gamma_\ell]$ and $k_2 = [n\gamma_u]$.

The Winsorizing percentage of *AWM* is set priory as Winsorized mean (*WM*). However, the amount to winsorize for each tail depends on the shape of the data. Then, the average of after adaptive Winsorizing process data us taken to get the *AWM*.

Universiti Utara Malaysia

The aforementioned robust central tendency measure such as adaptive trimmed mean using hinge estimator and *WM* proved to produce better results in statistical testing. Therefore, this study proposes the use of Winsorizing approach (symmetrical or asymmetrical using hinge estimator). In order to deal with the non-normal data, this study suggests the *AWM* rather than the *WM*. However, the *WM* is also proposed to assess its performance in proposed statistical test, *H*-statistic.

2.3 *H*-statistic

H-statistic is one of the robust statistical procedures that readily adapts to any central tendency measure. It gives reasonably good results when using *M*-estimator, but it is not recommended for mean or even trimmed mean (Wilcox, 2012). Schrader and Hettmansperger (1980) originally proposed this robust statistical test and defined as

$$H = \frac{1}{N}\sum_{j=1}^{J} n_j\left(\hat{\theta}_j - \hat{\theta}_.\right)^2 \tag{2.14}$$

$$N = \sum_j n_j$$

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

Othman et al. (2004) modified this statistic with modified one-step *M*-estimator (*MOM*) as its central tendency measure, $\hat{\theta}$ rather than *M*-estimator. This modified procedure was named as *MOM-H*. *MOM-H* proved that it is able to provide better control of Type I error rate even under skewed distributions. Syed Yahaya (2005) adopted scale estimators of *Sn* and *Tn* in *MOM-H* in her study. As a result, *MOM-H* produced better control of Type I error rate under normal distribution and showed reasonable performance under non-normal data. The *MOM* estimator used the outlier detection as its trimming criteria, but the outlier is not easy to detect. This may cause result in no outliers been being detected and the power rate will be affected (Wilcox, 2003).

Universiti Utara Malaysia

In this study, the Winsorizing approach mean, *WAM* (common naming for both *WM* and *AWM*) are suggested as the central tendency measures to substitute the *MOM* estimator in *H*-statistic. The Winsorizing percentages selected for *WAM* used in this study were 15% and 25%. According to the studies of Keselman et al. (2007) and Abdullah (2011), 15% and 25% of trimming provided better controlling of Type I error rate and high statistical power. Thus, six new procedures based on *H*-statistic were generated namely 15*WM-H*, 25*WM-H*, 15*WHQ-H*, 25*WHQ-H*, 15WHQ_1-*H* and 25WHQ_1-*H*.

2.3.1 15*WM-H*

15*WM-H* uses 15%-Winsorized mean (15*WM*) as the central tendency measure of *H*-statistic. First of all, calculate the 15%-Winsorized mean with Equation 2.12 where γ = 15% that might be discarded and replace 15% data from left tail and 15% data from right tail with the smallest and largest value of the remaining data respectively after the data are discarded.

After $15WM$ is calculated, compute the *H*-statistic by using Equation 2.14 with $15WM$ as the central tendency measure, $\hat{\theta}_j$. These steps apply to subsection 2.3.2 with 25% as its Winsorizing percentage.

2.3.2 25*WM-H*

25*WM-H* is the *H*-statistic using 25%-Winsorized mean (25*WM*) as its central tendency measure. This procedure is almost similar with 15*WM-H*. The difference is it uses 25% as its Winsorizing percentage. When calculating the Winsorized mean, 25*WM-H* uses γ = 25% in Equation 2.12. Next, the same step that uses Equation 2.14 in 15*WM-H* is applied in 25*WM-H* to compute its *H*-statistic.

2.3.3 15*WHQ-H*

15*WHQ-H* is the modified *H*-statistic with 15%-*AWM* using hinge estimator *HQ* as its central tendency measure. To obtain 15*WHQ-H*, calculate the *HQ* hinge estimator by using Equation 2.5. Then, use the value calculated from Equation 2.5 to compute the lower and upper Winsorizing proportions by using Equation 2.6 and 2.7 with 0.15 (15%) as its total Winsorizng percentage, γ.

$$\gamma_\ell = .15HQ = .15\,\frac{UW_Q}{UW_Q - LW_Q} \tag{2.15}$$

$$\gamma_u \;=\; .15 \;-\; \gamma_\ell \tag{2.16}$$

After the upper and lower proportions of Winsorization are calculated, use Equation 2.13 to calculate 15*WHQ*. Finally, the *H*-statistic is calculated by using Equation 2.14.

Universiti Utara Malaysia

2.3.4 25*WHQ-H*

25*WHQ-H* is another procedure as discussed in 2.3.3 that uses *HQ* to calculate its lower and upper Winsorizing proportions. However, this procedure uses 25% (0.25) as its total Winsorizing percentage as shown below

$$\gamma_\ell = .25HQ = .25\frac{UW_Q}{UW_Q - LW_Q} \qquad (2.17)$$

and

$$\gamma_u = .25 - \gamma_\ell \qquad (2.18)$$

Follow all steps in 2.3.3 and replace Equation 2.15 and 2.16 with Equation 2.17 and 2.18 respectively to calculate $\hat{\theta}_j$ (Equation 2.13) and *H*-statistic (Equation 2.14) as the modified *H*-statistic with 25%-*AWM* using hinge estimator *HQ*.

2.3.5 15*WHQ*$_1$-*H*

Universiti Utara Malaysia

The way to calculate modified *H*-statistic with 15%-*AWM* using hinge estimator *HQ*$_1$ (denoted as15*WHQ*$_1$-*H*) is similar to the steps in 2.3.3. The only difference is the use of hinge estimator, *HQ*$_1$ to calculate the Winsorizing proportion for 15%-*AWM*. This procedure uses Equation 2.9 to replace Equation 2.5 in procedure 2.3.3 to calculate the lower proportion of Winsorizing

$$\gamma_\ell = .15HQ_1 = .15\frac{UW_{Q_1}}{UW_{Q_1} - LW_{Q_1}} \qquad (2.19)$$

The next step is to calculate the upper proportion of Winsorizing by

$$\gamma_u = .15 - \gamma_\ell \qquad (2.20)$$

Following that, use Equation 2.13 to calculate the $\hat{\theta}_j$ and finally, the *H*-statistic is computed by using Equation 2.14.

2.3.6 25*WHQ*1-*H*

In this procedure, 25% of Winsorization is used to calculate 25%-*AWM* using hinge estimator HQ_1 in the modified *H*-statistic (denoted as 25WHQ_1-*H*). The Winsorizing proportion of 25%-*AWM* using hinge estimator HQ_1 is calculated by

$$\gamma_\ell = .25HQ_1 = .25\frac{UW_{Q_1}}{UW_{Q_1} - LW_{Q_1}} \tag{2.21}$$

$$\gamma_u = .25 - \gamma_\ell \tag{2.22}$$

Equation 2.21 and 2.22 are then used in Equation 2.13 to calculate 25%-*AWM* using hinge estimator HQ_1, $\hat{\theta}_j$. Finally, the *H*-statistic is computed by using the Equation 2.14. The results of the *H*-statistic depend on the modified procedures with *AWM* and the *AWM-H* is able to obtain better performance under various conditions.

Since the sampling distribution of *H*-statistic is unknown, the percentile bootstrap method was used to obtain the *p*-value (Othman et al., 2004).

Universiti Utara Malaysia

2.4 Bootstrap method

The bootstrap method is a resampling method that resamples and replaces the data from a dataset introduced by Efron (1979). It is able to measure accuracy to a statistical estimate routinely that is far too complicated for traditional computation of statistical analysis (Efron & Tibshirani, 1986, 1993). The bootstrapping is simple and can help in increasing the accuracy of the test statistic (Md Yusof, Abdullah, Syed Yahaya, & Othman, 2012). The bootstrap method is also able to deal with the issue of the unknown distribution of a statistical test.

From the study of Syed Yahaya (2005), the advantages of the bootstrap were listed, which includes:

i. wide applicability – requires no theoretical calculation, and is available no matter how mathematically complicated the estimator may be,

ii. increased accuracy,

iii. and ability to take advantage of modern computing and completely automatic.

In terms of obtaining good control of Type I error rate, the bootstrap technique can be advantageous. Some version of the percentile bootstrap method generally has practical advantage when measures of central tendency that are relatively insensitive to outliers are used (Wilcox & Keselman, 2003a). The bootstrap procedure for this study is discussed in the next chapter.

UUM
Universiti Utara Malaysia

CHAPTER THREE
METHODOLOGY

3.1 Introduction

Recently, a few robust statistical tests for the comparison of independent groups have been developed to obtain better results in controlling Type I error rates and higher statistical power under non-normality and heteroscedasticity. The *H*-statistic is one of these robust procedures that has been proven to be successful for those purpose as mentioned in the previous chapter. Furthermore, this test statistic is readily adaptable in any central tendency measures but not recommended for mean or even trimmed mean (Wilcox, 2012). Therefore, this study proposes a few statistical tests that use the Winsorized approach mean (Winsorized mean, *WM*, and adaptive Winsorized mean using hinge estimator, *AWM*) in *H*-statistic namely *WM-H* and *AWM-H*.

UUM Universiti Utara Malaysia

To evaluate the performance of these proposed procedures in terms of Type I error rate control and statistical power, five variables were manipulated to simulate various conditions that can be used to highlight the strengths and weaknesses of these statistical tests. These five variables were type of distribution, number of groups, sample sizes, degree of variance heterogeneity and nature of pairing. The performances of the proposed procedures were assessed using 5000 simulated datasets which were generated using SAS generator RANNOR of SAS/IML Version 9.3 (2011) and with a 5% statistical significance level ($\alpha = 0.05$).

3.2 Proposed Statistical Procedures

In this study, the *H*-statistic was modified to obtain an improved procedure capable of handling the problem of non-normality and heteroscedasticity. The *H*-statistic was modified by replacing the original central tendency measure with a 15% and 25%

winsorizing percentage of *WM* which is denoted as *WM-H*. Another modification of these *H*-statistic test denoted as *AWM-H* which used a 15% and 25% winsorizing percentage of *AWM* (using hinge estimators, *HQ* and HQ_1) to replace the original central tendency measure. The 15% and 25% of trimming percentage provided the best results in controlling Type I error rates and high statistical power in the studies of Keselman et al. (2007) and Abdullah (2011). Therefore, 15% and 25% of winsorization were chosen to represent the minimum and maximum winsorizing percentage in this study. Figure 3.1 illustrates the proposed procedures for this study.

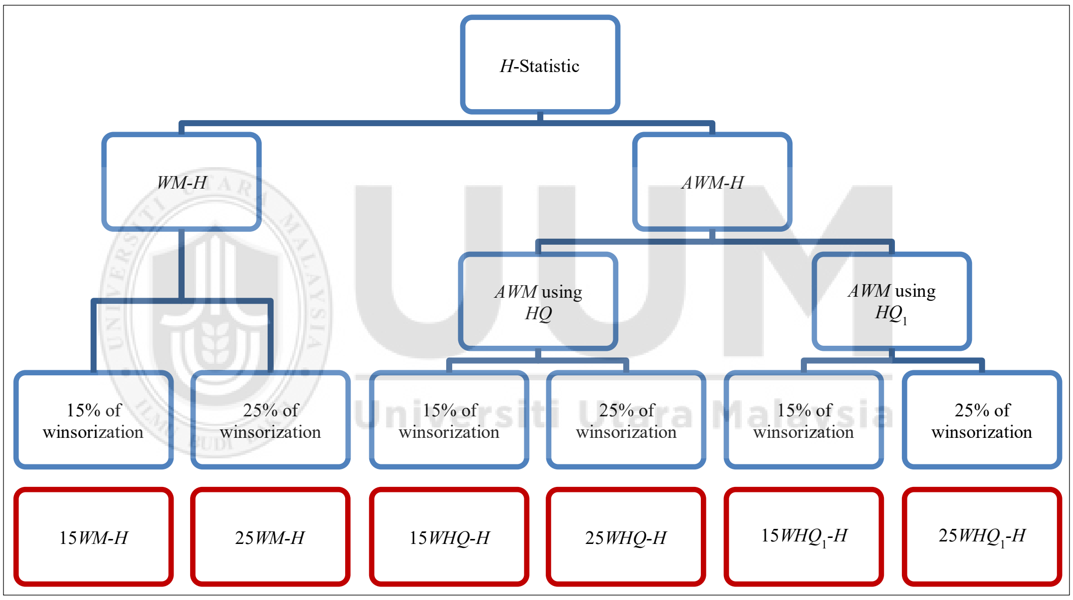


Figure 3.1. Statistical tests with the corresponding robust central tendency measure and percentage of winsorization

In general, this study generated six new *H*-statistic procedures namely 15*WM-H*, 25*WM-H*, 15*WHQ-H*, 25*WHQ-H*, 15WHQ_1-*H*, and 25WHQ_1-*H*. These six new procedures were compared to the *MOM-H* and classical procedures (Student's *t*-test and ANOVA, *F*-test).

3.3 Manipulation of Variables

The proposed procedures were tested under various conditions generated from manipulating five variables which are discussed as follow.

3.3.1 Type of Distribution

Violating the assumption of normality is one of the problems faced when using the classic statistical test in comparing the central tendency measures of two or more treatment groups. Under non-normal distribution conditions, the classical procedures such as Student's *t*-test and ANOVA *F*-test will yield inadequate and unsatisfactory results in controlling the Type I error rates and statistical power (Bradley, 1968; Syed Yahaya, 2005). In this study, four conditions of distribution with different levels of skewness and kurtosis (tail's length) were used to test the effects of distribution on the proposed procedures. These four types of distribution were generated by using the *g*- and -*h* distribution introduced by Tukey (1977) and extensively studied by Hoaglin, Mosteller and Tukey (1983), and Wilcox (2012).

Universiti Utara Malaysia

The *g*- and -*h* distribution was generated by transforming the standard normal distribution, *Z* using the following equation:

$$Y_{ij}=\begin{cases} \frac{\exp(gZ_{ij})\text{-}1}{g}\exp\left(\frac{hZ_{ij}^{2}}{2}\right), & g\neq 0 \quad (3.1) \\ Z_{ij}\exp\left(\frac{hZ_{ij}^{2}}{2}\right), & g=0 \quad (3.2) \end{cases}$$

where *g* and *h* are the non-negative constants with both effecting the skewness and kurtosis of distribution respectively. The increase in the *g* and *h* values will raise the values of the skewness and kurtosis leading to the distributions being skew and heavier distribution tails (Wilcox, 2012).

The first distribution selected was a standard normal distribution (g = h = 0) which represented zero skewness (κ_1 = 0) and was light-tailed (κ_2 = 0.3). The values of g = 0 and h = 0.5 were selected for symmetric heavy-tailed distribution with no skewed (κ_1 = 0) but heavy-tailed (κ_2 = undefined). The g = 1 and h = 0 values were chosen to represent the skewed normal tailed. In the study by Wilcox (2012), the g = 1 and h = 0 values corresponded to a lognormal distribution and it is skewed with a relatively light-tailed (κ_1 = 6.2, κ_2 =114). The forth distribution, which is a skewed heavy-tailed distribution generated from the *g*- and -*h* distribution with g = 1 and h = 0.5, was selected to represent a distribution with a level of skewness similar to a lognormal distribution but heavy-tailed (κ_1 = κ_2 = undefined). These four conditions of distribution were chosen in line with the assumptions that the proposed procedures are able to perform well under any condition which lies in between the normal and extreme values of the skewness and kurtosis of distribution. These four types of distribution are shown in Table 3.1 as follow.

Universiti Utara Malaysia

Table 3.1

The skewness and kurtosis of g- and -h distributions

g	h	**Skewness, κ_1**	**Kurtosis, κ_2**	**Distribution**
0	0	0	0.3	Normal
0	0.5	0	Undefined	Symmetry heavy tailed
1	0	6.2	114	Skewed normal tailed
1	0.5	Undefined	Undefined	Skewed heavy tailed

(Source: Wilcox, 2012)

3.3.2 Number of Groups

Besides the type of distribution, the number of groups, or group sizes is another condition being considered to evaluate the robustness of the proposed procedures. Therefore, this study used two and four groups as one of the conditions to assess the performances of the proposed procedures. The groups of two ($J = 2$) is the minimum number of groups for comparing independent groups. Furthermore, these group sizes often use the classical Student's *t*-test to determine the equality of central tendency measures due to its ability to perform well under normality. For the case of more than two groups, four groups ($J = 4$) were chosen as it is the moderate number of groups as studied by Lix and Keselman (1998).

3.3.3 Sample Sizes

In most data, the imbalance of sample sizes usually occurs in the biomedical field (Yang, Li, & Gao, 2006). The existence of these unbalanced sample sizes could affect the ability of statistical tests to control for Type I error rates and statistical power (Wilcox, 2003). Thus, both unbalanced and balanced sample sizes were taken into consideration when developing the proposed procedures of assessment for this variable in the study.

Universiti Utara Malaysia

For cases involving two groups, a total number of 40 samples were chosen (as sample size, N) because it has been commonly used in previous studies such as Othman et al. (2004), Syed Yahaya (2005), and Keselman et al. (2007). The number of observations (n_j) for each group under unequal sample size conditions were designated as n_1=15 and n_2=25 whereas n_1=n_2=20 was set for equal sample sizes.

For case with more than two groups, Othman et al. (2004), and Keselman et al. (2007) used the total number of sample size, $N = 70$ and $N = 90$ to study the effects of unequal sample sizes and both were able to provide reasonable results in controlling for Type I error rates. However, these two values are not suitable for a balanced sample size with 4 groups. Therefore, N = 80 is used in this study because it is between the two sample size values mentioned above and it is believed that this total number of sample will perform well in controlling the Type I error rates. Syed Yahaya (2005) also used the total number of sample, N = 80, and assigned the n_j for cases with unequal sample sizes as n_1 = 10, n_2 = 15, n_3 = 25 and n_4 = 30. Therefore, the same n_j values were used in this study to evaluate the effects of unequal sample sizes in the case of J = 4 for the proposed procedures. In contrast, $n_1 = n_2 = n_3 = n_4 = 20$ was assigned for cases with equal sample sizes.

3.3.4 Degree of Variance Heterogeneity

The unequal variances or variance heterogeneity is another problem faced when classical procedures were used in testing the equality of central tendency measures of two or more treatment groups. Even with normal and balanced sample sizes, the violation of the equal variance assumption will give unsatisfactory results for Type I error rate controls and statistical power (Wilcox, 1994).

The ratio of variances was selected as 1:1 (1:1:1:1) and 1:36 (1:1:1:36) to investigate the performance of the proposed procedures under equal and unequal variances respectively. The ratios of 1:36 (J =2) and 1:1:1:36 (J = 4) are the extreme heterogeneity of variances condition that are used to evaluate the performances of statistical procedures (Othman et al., 2004; Syed Yahaya, 2005; Keselman et al., 2007). Keselman et al. (2007) reported that the ratio of 1:36 (1:1:1:36) was large and reasonable to evaluate the effectiveness of the proposed procedures to perform under a ‘potentially’ extreme condition. Consequently, with the selected ratios, the performances of the proposed procedures were evaluated with equal and unequal variance conditions.

3.3.5 Nature of Pairing

From previous researches, the pairings of sample sizes with variances might provide different results in terms of the Type I error rate control (Othman et al., 2004; Syed Yahaya, 2005; Keselman et al., 2007). There are two types of nature pairings known as positive pairing and negative pairing. Both pairings were formed when unequal sample sizes are paired with unequal variances (Syed Yahaya, 2005). In cases where the smallest sample size, n_j is paired with the smallest variance, and the largest sample size, n_j is paired with the largest variance, this condition represents positive pairing (Othman et al., 2004; Syed Yahaya, 2005; Keselman et al., 2007). Negative pairing, on the other hand, refers to cases where the smallest n_j is paired with the largest variance and *vice versa*. To evaluate the robustness of the proposed procedures with the effects of nature pairing, both positive and negative pairings were included as one of the variables for this study.

Universiti Utara Malaysia

3.4 Design of Specification

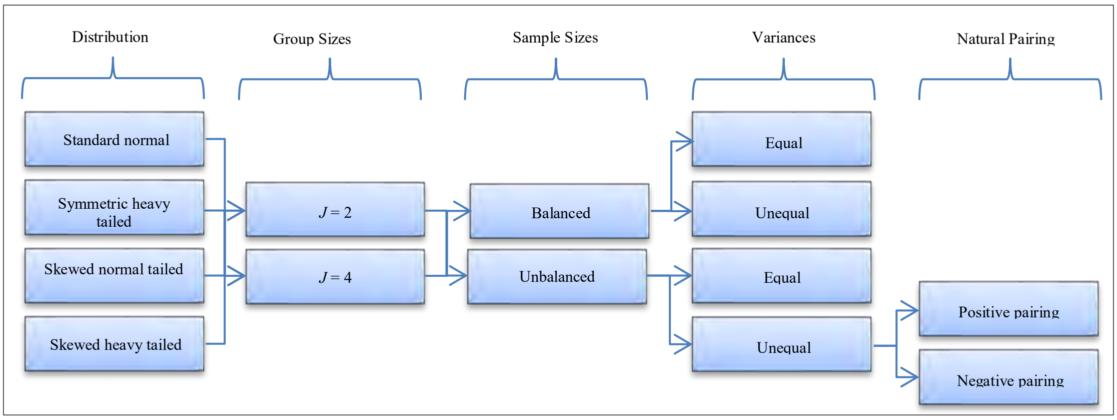


Figure 3.2. The conditions for investigating the robustness of the proposed procedures

To assess the robustness of the proposed procedures, the five variables thoroughly discussed in the previous sub-sections were manipulated. As shown in Figure 3.2, the manipulation of the five variables created a few test conditions that can highlight the strengths and weaknesses of these proposed procedures.

Table 3.2

Design specification for the balanced or unbalanced and J = 2 or J = 4 conditions

Groups	**Sample Sizes**				**Group Variances**				**Pairing**
$J = 2$, Balanced	20	20	-	-	1	1	-	-	-
	20	20	-	-	1	36	-	-	-
$J = 2$, Unbalanced	15	25	-	-	1	1	-	-	-
	15	25	-	-	1	36	-	-	Positive
	15	25	-	-	36	1	-	-	Negative
$J = 4$, Balanced	20	20	20	20	1	1	1	1	-
	20	20	20	20	1	1	1	36	-
$J = 4$, Unbalanced	10	15	25	30	1	1	1	1	-
	10	15	25	30	1	1	1	36	Positive
	10	15	25	30	36	1	1	1	Negative

Universiti Utara Malaysia

In general, the test conditions can be categorised into three conditions. The first is the perfect condition where the sample sizes and variances for each groups are equal and the shape of distribution is normal. The second condition is the mild departure condition with the combination of either balanced sample sizes and equal variances with non-normal distribution, or unbalanced sample sizes and unequal variance with normal distribution. The extreme departure condition is the third condition with a combination of unbalanced sample sizes and unequal variances with non-normal distribution.

Table 3.2 illustrates the design specification for J = 2 and J = 4. With the design specifications, the robustness of the proposed procedures was evaluated in terms of Type I error rates and statistical power.

3.5 Data Generation

In this study, the SAS/IML Version 9.3 (SAS, 2011) generator RANNOR was used to run the simulations to generate the pseudo-random variates. The g- and -h distribution was used to generate four types of distribution by transforming the standard normal variables to random variables via Equation 3.1 or 3.2 according to the selected g and h values.

For the symmetry distributions where $g = 0$, the central tendency measure is equal to zero so the null hypothesis is not affected when multiplying each Y_{ij} by σ_j to obtain unequal variances (Wilcox, 1994). For skewed distributions where $g > 0$, the central tendency measure is not equal to zero and the population mean of the g- and -h distribution is

$$\mu_{g\,h} = \frac{1}{g(1-h)^{1/2}}(e^{g^2/2(1-h)} - 1) \qquad (3.3)$$

(Hoaglin, 1985). Thus, the observations, Y_{ij}, from each simulated skewed distribution should subtract the population central tendency parameter (θ) as

$$X_{ij} = Y_{ij} - \theta \qquad (3.4)$$

before multiplying by σ_j, to ensure that the null hypothesis remains true. To compute the θ, one million observations from each investigated distributions were generated with the θ being eventually determined by using robust location estimators (Othman et al., 2004; Wilcox and Keselman, 2003b; Keselman et al., 2007). These θ values were then used to standardize the variates from each replication to assure that the null hypothesis remains true in every case.

For the cases with unequal variances, each X_{ij} from Equation 3.4 were multiplied by the square root of ${\sigma_j}^2$ to obtain a distribution with a standard deviation σ_j as

$$X_{ij} = Y_{ij} - \theta \times \sqrt{{\sigma_j}^2} \tag{3.5}$$

The σ_j, which is used to multiply with X_{ij} are not the actual values of the standard deviations (variances) and these values more aptly reflect the ratio of the variances (standard deviations) between the groups as in Table 3.2 (Wilcox, 1994).

For the statistical power analyses, the groups' central tendency measures were set to not be zero allowing the values of the groups' central tendency measures to vary according to the suggested effect size and pattern variability (refer to Section 3.6). Therefore, θ_j added to Equation 3.6 as

$$X_{ij} = Y_{ij} - \theta \times \sqrt{{\sigma_j}^2} + \theta_j \tag{3.6}$$

to conform to either Type I error rate or power analysis. The θ_j is always set at zero for Type I error rates assessment while the values of θ_j are depends on the settings of the central tendency measures for power analysis.

For each design, irrespective of the cases or the number of groups investigated, 5000 datasets for each conditions were performed using a 0.05 statistical significance level ($\alpha = 0.05$) to estimate the Type I error rates and the statistical power. In the study by Manly (2007), a minimum of 1000 datasets were sufficient in analysing the results at a 5% level of significance. However, better sampling limits will be obtained 99.9% of

Universiti Utara Malaysia

the time when using 5000 datasets instead of 1000 datasets (Manly, 2007). Therefore, for this study, 5000 datasets were chosen as the number of randomizations and each of these simulated datasets were then bootstrapped 599 times (refer to Section 3.7).

3.6 The Settings of Central Tendency Measures for Power Analysis

For the Type I error rate assessments, the first step is to define each group's central tendency measure to equal zero (H_0: $\mu_1 = \mu_2 = \cdots = \mu_n = 0$). The same setting is also required prior to statistical power analyses; however, the difference is the setting of central tendency measures should be defined according to the alternative hypothesis rule (H_1: $\mu_1 \neq \mu_2 \neq \cdots \neq \mu_n$). From previous studies such as Keselman, Wilcox, Algina, Fradette, and Othman (2004), Othman et al. (2004), and Syed Yahaya (2005), the values were defined according to the conventional values of small, medium and large effect sizes as proposed by Cohen (1988) and studied by Cohen (1992a), Cohen (1992b), and Murphy, Myors, and Wolach (2008). Table 3.3 showing the conventional values of effect sizes for *J* = 2 and *J* = 4 in describing large, medium, and small effects. Therefore, the different effect size values will provide dissimilar settings of central tendency measures for power analyses.

Universiti Utara Malaysia

Table 3.3

The conventional values for small, medium and large effects

Effect size	**Group size**	
	***J* = 2**	***J* = 4**
Small	0.20	0.10
Medium	0.50	0.25
Large	0.80	0.40

The distinction between the null hypothesis, H_0 and the alternative hypothesis, H_1 is defined as the effect size (Cohen, 1992b). The larger effect size will lead to easier to detect the effect of statistical tests and provide a larger value of statistical power (Murphy, Myors, & Wolach, 2008). However, the small and medium effect sizes are also included in this study to compare the statistical power performance of the proposed procedures.

Although the setting of central tendency measure is based on the effect size, *f*, the *f* needs to be translated to a range of standardized central tendency measures, *d*, before proceeding with any calculations. According to Cohen (1988), the *d* is the distance between the smallest and the largest of the *J* central tendency measures which is originally a mean as defined as:

$$d = \frac{\theta_{max} - \theta_{min}}{\sigma} \tag{3.7}$$

where θ_{max} and θ_{min} are the largest and the smallest of *J* central tendency measures respectively and σ is the (common) standard deviation within the populations:

$$\sigma = \sqrt{\frac{(n_1-1)\sigma_1^2+(n_2-1)\sigma_2^2+\cdots+(n_j-1)\sigma_j^2}{n_1+n_2+\cdots+n_j-j}} \tag{3.8}$$

As mentioned in the previous sub-section, the cases involving $J = 2$ and $J = 4$ were included in the power analyses of this study. For each case, the balanced and unbalanced designs were studied. Therefore, the settings of the central tendency measure for power analyses were set by according to the conditions shown in Table 3.2.

Universiti Utara Malaysia

3.6.1 Cases with Two Groups (*J* = 2)

For situations where *J* = 2, the *d* is the distance between standardized means divided by the (common) standard deviation within the population and defined as *f* by Cohen (1988):

$$d = f = \frac{|\theta_1 - \theta_2|}{\sigma} \quad (3.9)$$

with θ_1 and θ_2 representing the population means and σ being the standard deviation within the population. In this study, the population means were replaced with the central tendency measures that was discussed in the previous chapter.

In terms of balanced and unbalanced design, the *d*s were defined by using Equation 3.8 and Equation 3.9. The balanced design uses balanced sample sizes and variances for the calculation whereas the unbalanced design uses the pairing of unbalanced sample sizes with variances in both equal and unequal conditions.

Universiti Utara Malaysia

3.6.1.1 Balanced Design (*J* = 2)

Since the small, medium and large effect sizes as shown in Table 3.3 were used to study the power performance of the proposed procedures, the settings of the central tendency measure can be defined by using Equation 3.9 for balanced sample sizes with equal variances ($n_1 = n_2; \sigma_1^2 = \sigma_2^2$) which is one of the conditions in Table 3.2. The settings of the central tendency measure are defined as,

Small effect size, $0.2 = \frac{|1 - \theta_2|}{1}$

Medium effect size, $0.5 = \frac{|1 - \theta_2|}{1}$

Large effect size, $0.8 = \frac{|1 - \theta_2|}{1}$

if $\theta_1 = 1$ and the σ calculated from Equation 3.8 with sample sizes and variances are $n_1 = n_2 = 20$ and $\sigma_1^2 = \sigma_2^2 = 1$ respectively.

Then, the θ_2 for each effect size are

Small effect size, $\theta_2 = 1.2$

Medium effect size, $\theta_2 = 1.5$

Large effect size, $\theta_2 = 1.8$

Therefore, the settings of the central tendency measure for $J = 2$ with balanced sample sizes and equal variances were set as shown in Table 3.4.

Table 3.4

The setting of central tendency measures for power analysis under J = 2 equal variances for balanced design and unbalanced design

Effect size	**(θ_1, θ_2)**
Small, $f = 0.2$	(1, 1.2)
Medium, $f = 0.5$	(1, 1.5)
Large, $f = 0.8$	(1, 1.8)

Universiti Utara Malaysia

A similar procedure was used to calculate the setting of central tendency measure for the condition of balanced sample sizes which are paired with unequal variances ($n_1 = n_2; \sigma_1^2 \neq \sigma_2^2$). However, the (common) standard deviation within the population, σ, needs to be calculated by using Equation 3.8 as

$$\sigma = \sqrt{\frac{(20-1)1+(20-1)36}{20+20-2}} = 4.30$$

where $n_1 = n_2 = 20$ and $\sigma_1^2 = 1, \sigma_2^2 = 36$. Then, used Equation 3.9 to obtain the central tendency measure settings as

Small effect size, $0.2 = \frac{|1-\theta_2|}{4.30}$

Medium effect size, $0.5 = \frac{|1-\theta_2|}{4.30}$

Large effect size, 0.8 = $\frac{|1-\theta_2|}{4.30}$

If $\theta_1 = 1$ and the σ = 4.30, thus the θ_2 are

Small effect size, θ_2 = 1.86

Medium effect size, θ_2 = 3.15

Large effect size, θ_2 = 4.44

The *J* = 2 settings of the central tendency measure for balanced sample sizes with unequal variances are shown in Table 3.5.

Table 3.5

The setting of central tendency measures for power analysis under J = 2 unequal variances for balanced design

Effect size	**(θ_1, θ_2)**
Small, *f* = 0.2	(1, 1.86)
Medium, *f* = 0.5	(1, 3.15)
Large, *f* = 0.8	(1, 4.44)

Universiti Utara Malaysia

3.6.1.2 Unbalanced Design (*J* = 2)

The same steps were used to determine the settings of the central tendency measure for the unbalanced design (*J* = 2) power analyses. The values of sample sizes and variances for the unbalanced design with either equal or unequal variances followed the values which were shown in Table 3.2. For the unbalanced design with equal variances, the setting of central tendency measures is the same as the balanced design with equal variances. This is because the (common) standard deviation within the population, σ, is 1 for the cases involving equal variances that $\sigma_1^2 = \sigma_2^2 = 1$ regardless the sample sizes are unbalanced ($n_1 = 15$, $n_2 = 20$) or balanced. Therefore, the setting of central tendency measures for this case was the same as shown in Table 3.4.

For the unbalanced design with unequal variances ($n_1 \neq n_2$; $\sigma_1^2 \neq \sigma_2^2$), it was divided according to positive pairing and negative pairing. When $n_1 = 15$, $n_2 = 20$, $\sigma_1^2 = 1$ and $\sigma_2^2 = 36$, this condition is the positive pairing of unbalanced design and the σ is calculated by using Equation 3.8 as

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

and $\sigma = 4.81$ being the result. Then, the settings of the central tendency measure were defined by using Equation 3.9 with $\sigma = 4.81$ and if $\theta_1 = 1$

Small effect size, $0.2 = \frac{|1-\theta_2|}{4.81}$

Medium effect size, $0.5 = \frac{|1-\theta_2|}{4.81}$

Large effect size, $0.8 = \frac{|1-\theta_2|}{4.81}$

Universiti Utara Malaysia

Therefore, $\theta_2 = 1.96, 3.41$ and 4.85 for small, medium and large effect sizes respectively. The results for the settings of the central tendency measure for power analysis under $J = 2$ unbalanced design with positive pairing are shown in Table 3.6.

Table 3.6

The setting of central tendency measures for power analysis under J = 2 positive pairing for unbalanced design

Effect size	**(θ_1, θ_2)**
Small, $f = 0.2$	(1, 1.96)
Medium, $f = 0.5$	(1, 3.41)
Large, $f = 0.8$	(1, 4.85)

Besides positive pairing, negative pairing is also considered in the study of statistical power. The negative pairing for unbalanced design is formed when $n_1 = 15$ and $n_2 = 20$ are associated with $\sigma_1^2 = 36$ and $\sigma_2^2 = 1$. By using Equation 3.8, the σ was computed as

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

$$= 3.73$$

With $\sigma = 3.73$, the central tendency measure settings for power analysis were calculated by using Equation 3.9, where if $\theta_1 = 1$

Small effect size, $0.2 = \frac{|1-\theta_2|}{3.73}$

Medium effect size, $0.5 = \frac{|1-\theta_2|}{3.73}$

Large effect size, $0.8 = \frac{|1-\theta_2|}{3.73}$

Thus, θ_2 for unbalanced design with negative pairing under *J* = 2 for small, medium and large effect size were shown in Table 3.7.

Universiti Utara Malaysia

Table 3.7

The setting of central tendency measures for power analysis under J = 2 negative pairing for unbalanced design

Effect size	**(θ_1, θ_2)**
Small, *f* = 0.2	(1, 1.75)
Medium, *f* = 0.5	(1, 2.87)
Large, *f* = 0.8	(1, 3.98)

3.6.2 Cases with Four Groups

For the cases where *J* is more than two, the Equation 3.9 is not suitable to determine the setting of central tendency measures. For this case, Cohen (1988) proposed the spread of the means by a value similar to the standard deviation instead of the distance. Then, this value is divided with the common standard deviation of the population as shown

$$f = \frac{\sigma_m}{\sigma} \tag{3.10}$$

where, the σ_m for the balanced design is

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

$$m = \frac{\sum_{j=1}^{J} m_j}{J}, \qquad J = 1, 2, \cdots, j \tag{3.12}$$

and for the unbalanced design is

$$\sigma_m = \sqrt{\frac{\sum_{j=1}^{J} n_j(m_j - m)^2}{N}} \tag{3.13}$$

$$m = \frac{\sum_{j=1}^{J} n_j m_j}{N}, \qquad J = 1, 2, \cdots, j \tag{3.14}$$

Universiti Utara Malaysia

The standard deviation of the population means signifies the standard deviation within the population (Cohen, 1988).

As discussed earlier, the value of *f* needs to be converted to *d* by using Equation 3.7 and Equation 3.8 before determining the setting of central tendency measures. When the case has means more than two, the relationship between the *f* and the *d* depends upon the dispersion of the means over their range (Cohen, 1988). Equation 3.7 defined the range between the largest and the smallest of the *J* means; then, the remaining *J* - 2 means are fall variously over the *d* and unable to be determined. Thus, Cohen (1988) proposed three patterns of variability to describe the relationship between *f* and *d* with a function of *J* means. The three patterns of variability described by Cohen (1988) are minimum, intermediate and maximum variability respectively.

The focus of this study is to assess the large effect of variability, thus the maximum variability was selected which yields the maximum standard deviation with *J* means spread at both extreme ends of the range (Cohen, 1988). Cohen (1988) proposed this pattern of variability with all even numbers of the means being

$$d = 2f \tag{3.15}$$

and when the number of means is odd

$$d = f\frac{2J}{\sqrt{J^2-1}} \tag{3.16}$$

The even number of groups ($J = 4$) was selected for this study, thus the Equation 3.15 was used for both balanced and unbalanced designs. By the definition of Cohen (1988), when the *J* is even, the number of *J* will evenly fall at $-\frac{1}{2}d$ and $+\frac{1}{2}d$. Therefore, the pattern variability for the four central tendency measures are $-\frac{1}{2}d, -\frac{1}{2}d, +\frac{1}{2}d, +\frac{1}{2}d.$

Universiti Utara Malaysia

3.6.2.1 Balanced Design ($J = 4$)

Similar to $J = 2$, all three effect sizes – small, medium and large – where $f = 0.1$, 0.25 and 0.4 respectively (refer to Table 3.3) were used for the proposed procedures' power assessment under $J = 4$ conditions. As mentioned in the previous sub-section, this study will focus on maximum variability as suggested by Cohen (1988); hence, the pattern variability of the central tendency measures is $-\frac{1}{2}d, -\frac{1}{2}d, +\frac{1}{2}d, +\frac{1}{2}d$ Then, the d for cases with balanced sample sizes paired with equal variances ($n_1 = n_2 = n_3 = n_4$; $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$) were calculated directly by using Equation 3.15 ($J = 4$ = even) and the small, medium and large effect sizes were

Small effect size, $d = 2(0.1) = 0.2$

Medium effect size, $d = 2(0.25) = 0.5$

Large effect size, $d = 2(0.4) = 0.8$

Thus, the dispersion of the central tendency measures for this pattern variability are as shown as in Table 3.8 for small, medium and large effect size respectively.

Table 3.8

The setting of central tendency measures for power analysis under J = 4 equal variances for balanced design

Effect size	**($\theta_1, \theta_2, \theta_3, \theta_4$)**
Small, *f* = 0.1	(-0.1, -0.1, 0.1, 0.1)
Medium, *f* = 0.25	(-0.25, -0.25, 0.25, 0.25)
Large, *f* = 0.4	(-0.4, -0.4, 0.4, 0.4)

For cases involving balanced sample sizes paired with unequal variances ($n_1 = n_2 = n_3 = n_4$; $\sigma_1^2 \neq \sigma_2^2 \neq \sigma_3^2 \neq \sigma_4^2$), the value of *d* is determined by using Equation 3.80, 3.10 to 3.12. First, the values of *m* are determined by using Equation 3.12 with the assumption that $m_1 = m_2 = -\frac{1}{2}d$, $m_3 = m_4 = \frac{1}{2}d$ as proposed by Cohen (1988) regarding the maximum variability pattern of the central tendency measures $\left(-\frac{1}{2}d, -\frac{1}{2}d, +\frac{1}{2}d, +\frac{1}{2}d\right)$, thus,

UUM Universiti Utara Malaysia

$$m = \frac{0}{4} = 0$$

Then, Equation 3.11 is used to calculate the σ_m as

$$\sigma_m = \sqrt{\frac{(-\frac{1}{2}d)^2 + (-\frac{1}{2}d)^2 + (\frac{1}{2}d)^2 + (\frac{1}{2}d)^2}{4}}$$

$$= \frac{1}{2}d$$

Next, the σ is computed by using Equation 3.80 with sample sizes, $n_1 = n_2 = n_3 = n_4 = 20$ and variances, $\sigma_1^2 = 1$, $\sigma_2^2 = 1$, $\sigma_3^2 = 1$, $\sigma_4^2 = 36$ as

$$\sigma = \sqrt{\frac{(20-1)1+(20-1)1+(20-1)1+(20-1)36}{20+20+20+20-4}}$$

$$= 3.12$$

Lastly, the d for small, medium and large effect sizes were determined by using Equation 3.10 as

Small effect size, $0.1 = \frac{1/2\, d}{3.12}$

$$d = 0.62$$

Medium effect size, $0.25 = \frac{1/2\, d}{3.12}$

$$d = 1.56$$

Large effect size, $0.4 = \frac{1/2\, d}{3.12}$

$$d = 2.50$$

As per result above, the dispersion of the central tendency measures for small, medium and large effect sizes with balanced sample sizes and unequal variances are illustrated in Table 3.9 below.

Universiti Utara Malaysia

Table 3.9

The setting of central tendency measures for power analysis under J = 4 unequal variances for balanced design

Effect size	**($\theta_1, \theta_2, \theta_3, \theta_4$)**
Small, f = 0.1	(-0.31, -0.31, 0.31, 0.31)
Medium, f = 0.25	(-0.78, -0.78, 0.78, 0.78)
Large, f = 0.4	(-1.25, -1.25, 1.25, 1.25)

3.6.2.2 Unbalanced Design ($J = 4$)

For the unbalanced design of $J = 4$, the settings of the central tendency measures were determined by using Equation 3.8, 3.10, 3.13 and 3.14 with f equal 0.1, 0.25 and 0.4 for small, medium and large effect sizes respectively. Let's $m_1 = m_2 = -\frac{1}{2}d$, $m_3 = m_4 = \frac{1}{2}d$ as proposed by Cohen (1988) for maximum variability pattern of the central tendency measures, Equation 3.13 and Equation 3.14 were used to determine the σ_m as

$$\sigma_m = \sqrt{\frac{\sum_{j=1}^{J} n_j(m_j - m)^2}{N}} = \sqrt{\frac{(n_1 + n_2)(n_3 + n_4)^2 + (n_3 + n_4)(n_1 + n_2)^2}{N^3} d^2}$$

where

$$m = \frac{\sum_{j=1}^{J} n_j m_j}{N} = \frac{n_3 + n_4 - n_1 - n_2}{2N} d$$

Then, σ is calculated by using Equation 3.8 as

$$\sigma = \sqrt{\frac{(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2 + \cdots + (n_j - 1)\sigma_j^2}{n_1 + n_2 + \cdots + n_j - j}}$$

UUM Universiti Utara Malaysia

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

Lastly, the d is calculated by using Equation 3.10 as

$$f = \frac{\sigma_m}{\sigma} = \frac{\sqrt{\frac{(n_1 + n_2)(n_3 + n_4)^2 + (n_3 + n_4)(n_1 + n_2)^2}{N^3} d^2}}{\sigma}$$

$$d = \sqrt{\frac{\sigma^2 \times f^2 \times N^3}{(n_1 + n_2)(n_3 + n_4)^2 + (n_3 + n_4)(n_1 + n_2)^2}} \qquad (3.17)$$

According to the previous studies by Keselman et al. (2004), and Othman et al. (2004), the setting of the central tendency measures, -1, -1, 1, 1 was developed for maximum variability with a large effect size with Syed Yahaya (2005) further studying for the statistical power analysis. This central tendency measure setting is based on the conditions described in Table 3.2 where $n_1 = 10$, n_2=15, n_3=25 and n_4=30 as the sample sizes and $\sigma_1^2 = 36$, $\sigma_2^2 = 1$, $\sigma_3^2 = 1$, and $\sigma_4^2 = 1$ as the variances. This condition is known as unbalanced design with negative pairing of unequal variances ($n_1 \neq n_2 \neq n_3 \neq n_4$; $\sigma_1^2 \neq \sigma_2^2 \neq \sigma_3^2 \neq \sigma_4^2$). Thus, the setting of the central tendency measures for maximum variability with large effect size is computed by Equation 3.8 and Equation 3.17 as

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

$$= 2.27$$

and

$$d = \sqrt{\frac{2.27^2 \times 0.4^2 \times 80^3}{(10+15)(25+30)^2 + (25+30)(10+15)^2}}$$

$$= 1.96 \approx 2$$

Based on maximum variability pattern of the central tendency measures, the setting of these measures for this condition was -1, -1, 1, 1 and this setting has been used in Keselman et al. (2004), Othman et al. (2004), and Syed Yahaya (2005) studies. However, *d* = 1.96 instead of *d* = 2 was used in this study to obtain results with higher accuracy. For small and medium effect sizes, the same steps were used as

$$\text{Small effect size, } d = \sqrt{\frac{2.27^2 \times 0.1^2 \times 80^3}{(10+15)(25+30)^2 + (25+30)(10+15)^2}}$$

$$= 0.49$$

$$\text{Medium effect size, } d = \sqrt{\frac{2.27^2 \times 0.25^2 \times 80^3}{(10+15)(25+30)^2 + (25+30)(10+15)^2}}$$

$$= 1.22$$

Therefore, the setting of central tendency measures for unbalanced design with negative pairing for small, medium and large effect sizes are shown in Table 3.10.

Universiti Utara Malaysia

Table 3.10

The setting of central tendency measures for power analysis under J = 4 negative paring for unbalanced design

Effect size	**($\theta_1, \theta_2, \theta_3, \theta_4$)**
Small, $f = 0.1$	(-0.25, -0.25, 0.25, 0.25)
Medium, $f = 0.25$	(-0.61, -0.61, 0.61, 0.61)
Large, $f = 0.4$	(-0.98, -0.98, 0.98, 0.98)

Besides the negative pairing of the unbalanced design, positive pairing is also used as mentioned previously. The condition of the positive pairing for unbalanced design involves the sample sizes $n_1 = 10$, n_2=15, n_3=25 and n_4=30 paired with variances $\sigma_1^2 = 1$, $\sigma_2^2 = 1$, $\sigma_3^2 = 1$, and $\sigma_4^2 = 36$. By using Equation 3.8 and Equation 3.17, the setting of the central tendency measures for maximum variability with small, medium and large effect sizes were computed as

Universiti Utara Malaysia

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

$$= 3.79$$

and

$$\text{Small effect size, } d = \sqrt{\frac{3.79^2 \times 0.1^2 \times 80^3}{(10+15)(25+30)^2 + (25+30)(10+15)^2}}$$

$$= 0.82$$

$$\text{Medium effect size, } d = \sqrt{\frac{3.79^2 \times 0.25^2 \times 80^3}{(10+15)(25+30)^2 + (25+30)(10+15)^2}}$$

$$= 2.04$$

$$\text{Large effect size, } d = \sqrt{\frac{3.79^2 \times 0.4^2 \times 80^3}{(10+15)(25+30)^2 + (25+30)(10+15)^2}}$$

$$= 3.27$$

Thus, the setting of central tendency measures for unbalanced design with positive pairing under J = 4 for small, medium and large effect size are illustrated as in Table 3.11.

Table 3.11

The setting of central tendency measures for power analysis under J = 4 positive paring for unbalanced design

Effect size	**($\theta_1, \theta_2, \theta_3, \theta_4$)**
Small, f = 0.1	(-0.41, -0.41, 0.41, 0.41)
Medium, f = 0.25	(-1.02, -1.02, 1.02, 1.02)
Large, f = 0.4	(-1.64, -1.64, 1.64, 1.64)

Universiti Utara Malaysia

Another condition under unbalanced design is unbalanced sample sizes paired with equal variances ($n_1 \neq n_2 \neq n_3 \neq n_4$; $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$). The setting of the central tendency measures for this condition was also computed by using Equation 3.8 and Equation 3.17 with sample sizes $n_1 = 10$, n_2=15, n_3=25 and n_4=30 and variances $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = 1$ as described in Table 3.2. The d for small, medium and large effect sizes were calculated as

$$\text{Small effect size, } d = \sqrt{\frac{1^2 \times 0.1^2 \times 80^3}{(10+15)(25+30)^2 + (25+30)(10+15)^2}}$$

$$= 0.22$$

$$\text{Medium effect size, } d = \sqrt{\frac{1^2 \times 0.25^2 \times 80^3}{(10+15)(25+30)^2 + (25+30)(10+15)^2}}$$

$$= 0.54$$

$$\text{Large effect size, } d = \sqrt{\frac{1^2 \times 0.4^2 \times 80^3}{(10+15)(25+30)^2 + (25+30)(10+15)^2}}$$

$$= 0.86$$

where the $\sigma = 1$ due to $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = 1$. For $J = 4$ unbalanced design that is paired with equal variances, the setting of central tendency measures for small, medium and large effect sizes are as shown in Table 3.12.

Table 3.12

The setting of central tendency measures for power analysis under J = 4 equal variances for unbalanced design

Effect size	**($\theta_1, \theta_2, \theta_3, \theta_4$)**
Small, $f = 0.1$	(-0.11, -0.11, 0.11, 0.11)
Medium, $f = 0.25$	(-0.27, -0.27, 0.27, 0.27)
Large, $f = 0.4$	(-0.43, -0.43, 0.43, 0.43)

UUM Universiti Utara Malaysia

With the settings of central tendency measures for statistical power analysis as above, the robustness of proposed methods under small, medium and large effect sizes with maximum variability for 2 and 4 group sizes can be evaluated.

3.7 Bootstrap

Bootstrap is a computer-based method that is able to routinely assess the accuracy of an estimator with high intricacy for traditional computation of statistical analysis (Efron & Tibshirani, 1986, 1993). It can provide reasonably accurate results for some problems of hypotheses testing (Wilcox, 2012).

When working with intractable sampling distributions, bootstrap methods are able to cope with this problem by resampling the random sample that has insufficient information about a population (Syed Yahaya, 2005). Additionally, better results for Type I error rates under small sample sizes can be obtained, but it should be cautioned that some of these bootstrap methods will actually provide worse results (Wilcox, 2012). The combination of bootstrap methods with certain robust central tendency measures can improve the control of Type I error rates (Keselman, Wilcox, Othman, & Fradette, 2002). Othman et al. (2004), and Syed Yahaya (2005) used the percentile bootstrap in their studies to obtain the significance level, *p*-value of *MOM-H* statistic. The percentile bootstrap is relatively effective in terms of controlling the Type I error rates with at least 20% of trimming but it is still able to perform reasonably well with 15% and even 10% trimming (Wilcox, 2012). When computing a *p*-value, the percentile bootstrap has faster execution time and do not lead to the computational problems (Wilcox, 2012).

Besides Type I error rates, the bootstrap methods also work on statistical power analysis with the use of percentile bootstrap in the studies of Othman et al. (2004), and Wilcox, Keselman, Muska, and Cribbie (2000). Ultimately, Wilcox (2012) found more than 1000 journal articles that have obtained great practical applications in both theoretical and simulation studies on the bootstrap methods.

Universiti Utara Malaysia

3.7.1 Modified *H*-statistic with Percentile Bootstrap

As discussed above, the percentile bootstrap could provide significant results in *MOM-H* statistic. Therefore, the similar approach has been employed in the modified *H*-statistic with proposed central tendency measures of Winsorized mean (*WM*) and adaptive Winsorized mean with hinge estimators, *HQ* and HQ_1 (*AWM*), and named as *WM-H* and *AWM-H* respectively. The percentile bootstrap was used to compute the *p*-values that represented the estimated Type I error rates or statistical power. The steps of the percentile bootstrap to obtain the *p*-values are as follow:

(a) Calculate the modified *H*-statistic with proposed central tendency measures based on the available data and denote it as *H*.

(b) Randomly sample and replace the n_j observations from the j^{th} group to obtain bootstrap samples of $X^*_{1j}, X^*_{2j}, \cdots, X^*_{n_j j}$.

(c) Centre each bootstrap sample with respective estimated central tendency measures, $\hat{\theta}$ as C^*_{ij}=$X^*_{ij} - \hat{\theta}_j$, i=1, 2, $\cdots$, n_j.

(d) Use the C^*_{ij} values to calculate the modified H-statistic and denote it as H^*.

(e) Repeat Step 2 to Step 4 B times to yield $H^*_1, H^*_2, \cdots, H^*_B$, where B = 599 appears to suffice in most situations when $n \geq 12$ (Wilcox, 2012).

(f) Obtain the p-value by using (# of $H^* > H$)/B.

To set the number of bootstrap replications, B, there is not a fixed value and it is according to approximations. Efron and Tibshirani (1993) found B = 50 was often enough to provide a good estimate of standard error. In Othman et al. (2004), and Syed Yahaya (2005), the B = 599 was used because it provided reasonable results in their respective studies. Besides that, the Wilcox, Keselman, and Kowalchuk (1998) study proved that the setting of B = 599 instead of 600 resulted to three liberal values of the Welch statistic decreasing in size to 0.074 (from 0.076), 0.077 (from 0.078), and 0.07 (from 0.077). Furthermore, the results from Hall (1986) also showed the advantages of choosing B = 599 rather than B = 600 because $1 - \alpha$ is a multiple of $(B + 1)^{-1}$ and $1 - \alpha$ = .95 is the primary focus of the study.

CHAPTER FOUR
RESULT OF ANALYSIS

4.1 Introduction

Normality and homogeneity are the two main concerns when performing a statistical analysis. The accuracy of the analysis result will be affected if any of these two assumptions are violated, which may result in wrong decision-making. The search for an alternative procedure that is able to obtain a better Type I error rate and with more statistical power under a violated assumption condition will be the focus of this study.

This study proposes a two-test statistic formed from the modification of an *H*-statistic with a Winsorized mean (*WM-H*) and an *H*-statistic with an adaptive Winsorized mean (*AWM-H*) as its central tendency measure. The Winsorized mean (*WM*) and adaptive Winsorized mean (*AWM*) require a predetermined percentage similar to the usual trimmed mean, but the trimmed value will be winsorized with the smallest or largest or both smallest and largest remaining data before the mean is computed. The difference between *WM* and *AWM* is that *AWM* requires a hinge estimator, so that it would be able to perform asymmetry winsorizing according to the distribution shape, whereas *WM* only able to perform symmetry winsorizing. In this study, 15% and 20% were selected as the predetermined percentage for winsorization and hinge estimators, *HQ* and HQ_1 were used in the *AWM* calculation. After selecting the predetermined winsorized percentage and hinge estimators, six procedures are formed as mentioned in Section 3.2.

A comparison between the robustness of the proposed procedures in terms of Type I error rates and statistical power was done with the original *MOM-H* and the classical procedures that are Student’s *t*-test (2-group tests) and ANOVA *F*-test (4-group tests).

Universiti Utara Malaysia

To highlight the strengths and weaknesses of the compared procedures, various variables were used, as stated in Section 3.3, such as the number of groups, the type of population distribution, sample size, the degree of variance heterogeneity, and the nature of pairing. A few test conditions were proposed by manipulating the variables. The compared procedures were then tested with these test conditions. All the results in terms of Type I error rates are outlined in Table 4.1 and Table 4.2.

As discussed in Section 3.3.1, four types of distribution generated by *g*- and -*h* distribution, shown in the first column of Table 4.1 and Table 4.2, are standard normal distribution ($g = h = 0$), symmetric heavy tailed distribution ($g = 0$; $h = 0.5$), skewed normal tailed distribution ($g = 1$; $h = 0$), and skewed heavy tailed distributions ($g = 1$; $h = 0.5$). The second and third columns of Table 4.1 and Table 4.2 represent the pairing of the sample sizes and variances. Two types of natural paring i.e. positive pairing and pairing will be formed as a result of the unbalanced sample sizes paired with unequal variances, as displayed in the fourth column of Table 4.1 and Table 4.2. The natural pairing does not apply to the balanced design and unbalanced design with equal variances. The fifth to tenth columns of Table 4.1 and Table 4.2 outline the Type I error rates of the proposed procedures formed via the modified *H*-statistic and compared with the Type I error rates of *MOM-H* and classical procedure (Student's *t*-test or ANOVA *F*-test) that are displayed in the eleventh and twelfth columns of Tables 4.1 and 4.2. The last row ("AVERAGE") for each distribution is the average of Type I error rates for each procedure, which corresponds to the type of distribution. The "GRAND AVERAGE" is the average value from all Type I error rates, obtained from each procedure and displayed in the last row of Table 4.1 and Table 4.2. For the power analysis, an additional column, which is the effect sizes column, is added after the column for natural pairing.

Universiti Utara Malaysia

4.2 Type I Error Rates

The robustness of the proposed procedures were evaluated using the robustness criterion proposed by Bradley (1978). According to the criterion, the proposed procedure is considered robust if its empirical Type I error rate (ρ) is within 0.5α and 1.5α. Therefore, at the 5% statistical significance level ($\alpha = 0.05$) used in this study, the procedure will only be considered robust in any manipulated condition design (refer to Section 3.3) if its empirical Type I error rate falls within 0.025 and 0.075 (marked with an underline).

When the empirical Type I error rate is smaller than 0.025, the procedure is considered conservative, whereas a liberal procedure entails an empirical Type I error rate that is larger than 0.075. The robustness criterion based on Bradley (1978) was chosen for this study since it had been widely used in study such as Othman et al. (2004), Syed Yahaya (2005) and Abdullah (2011). Besides the Bradley's (1978) robustness criterion, another criterion was also used, i.e. the procedure that can provide a Type I error rate closest to the nominal (significance) level of $\alpha = 0.05$ would be considered the best procedure (marked with bold font).

UUM
Universiti Utara Malaysia

4.2.1 Type I Error Rates for the Two-Group Test ($J = 2$)

The Type I error rate results for the Two-group test ($J = 2$) are illustrated in Table 4.1. All of the procedures were tested under four types of population distribution with five conditions generated from the manipulation of sample sizes and variances.

4.2.1.1 Standard normal distribution ($g = h = 0$)

When the procedures are tested under standard normal distribution ($g = h = 0$), all of the proposed procedures were determined to be robust, as their Type I error rates fell within Bradley's robustness criterion interval (0.025 and 0.075) for all five manipulated conditions, except for the Student's *t*-Test under unbalanced sample sizes paired with unequal variances either positive or negative pairing, which obtained Type I error rates of 0.0198 and 0.1268 respectively. However, the overall performance of the procedures fulfilled the robustness criterion based on their "AVERAGE" values and thus are considered robust.

4.2.1.2 Symmetric heavy tailed distribution (g = 0; h = 0.5)

The Type I error rates for all procedures grew smaller with an increase in kurtosis with zero skewness. When the procedures are tested under symmetric heavy tailed distribution (g = 0; h = 0.5) with balanced sample sizes and equal variances, none of the procedures met Bradley's robustness criterion with giving conservative Type I error rates, except for 15*WHQ-H* and the Student's *t*-test, which were considered robust with 0.0292 and 0.0356 as their respective Type I error rates All the procedures were determined to be robust when tested under unequal variances with 15*WHQ-H* found to be the best procedure under this condition, as it yielded a Type I error rate equal to 0.0430, which is the closest value to the nominal level of α = 0.05.

Universiti Utara Malaysia

Under unbalanced sample sizes, only 15*WHQ-H*, 15*WHQ*$_{1}$*-H*, and the Student's *t*-test fulfilled the robustness criterion with Type I error rates of 0.0300, 0.0258, and 0.0374, respectively, when variances are equal. When the condition consisted of unbalanced sample sizes with positive pairing, except for 25*WHQ*$_{1}$*-H* (0.0242) and *t*-test (0.0118), the other procedures satisfied Bradley's criterion of robustness. With negative pairing, all proposed procedures met the robustness criterion except for 15*WM-H* and the Student's *t*-test. The overall performance under symmetric heavy tailed distribution resulted in the "AVERAGE" values for all procedures falling within the robustness criterion interval except for the 15*WM-H* and 25*WHQ*$_1$-*H* where both of these procedures yielded conservative Type I error rates.

4.2.1.3 Skewed normal tailed distribution (*g* = 1; *h* = 0)

All proposed procedures were able to perform within Bradley's robustness criterion interval under a skewed normal tailed distribution (*g* = 1; *h* = 0) when the variances are equal, either with balanced or unbalanced sample sizes. Among the proposed procedures, 15*WHQ-H* was determined to be the best method, yielding the closest Type I error rates (0.0476 and 0.0524) to the nominal level, 0.05 under both conditions.

When the variances became unequal, only three procedures, 15*WM-H*, 25*WM-H,* and *MOM-H* performed well with Type I error rates that fell within the robustness criterion interval regardless of the sample size design or natural pairing. Besides the three procedures, the Student's *t*-test also performed well under positive pairing with a Type I error rate of 0.0370. 25*WQH*$_1$-*H* was also found to be robust under negative pairing besides the three procedures mentioned earlier. 15*WM-H*, 25*WM-H*, 25*WQH*$_1$-*H*, and *MOM-H* were the procedures that yielded "AVERAGE" values of 0.0434, 0.0374, 0.0620, and 0.0448 respectively, which were satisfying Bradley's robustness criterion and also proving to have a robust overall performance under a skewed normal tailed distribution.

4.2.1.4 Skewed heavy tailed distributions (*g* = 1; *h* = 0.5)

15*WHQ-H* was the only procedure considered robust according to the Bradley's robustness criterion under a skewed heavy tailed distribution (*g* = 1; *h* = 0.5) with balanced sample sizes and equal variances. The other procedures gave conservative Type I error rates under this condition. However, the Type I error rates for all procedures increased when the variances became unequal. In the balanced sample sizes with unequal variances condition, all the procedures were found to be robust except for 15*WM-H* (0.0228) and 25*WM-H* (0.0222), which yielded conservative Type I error rates.

Under unbalanced sample sizes with equal variances, only 15*WHQ-H* (0.0370) and the *t*-test (0.0272) were considered robust, whilst other procedures did not meet Bradley's robustness criterion. Under the unbalanced sample sizes with positive pairing condition, half of the procedures, which are 15*WM-H*, 25*WHQ-H*, 15*WHQ*$_1$-*H*, and *MOM-H* fulfilled the robustness criterion, whereas the other half fell outside the criterion interval of 0.025 to 0.075. All procedures were able to produce Type I error rates within the robust criterion interval except for 15*WM-H* and the *t*-test under negative pairing condition of skewed heavy tailed distribution.

Universiti Utara Malaysia

Under this extreme condition, the majority of the procedures performed well with providing "AVERAGE" Type I error rates from 0.0256 to 0.0560. However, three out of eight procedures, i.e. 15*WM-H*, 25*WM-H*, and 25*WHQ*$_1$-*H* failed to satisfy Bradley's robustness criterion with "AVERAGE" values of 0.0171, 0.0198, and 0.0242, respectively.

The overall performance of the procedures, based on the "GRAND AVERAGE" values, resulted in none of them failing Bradley's robustness criterion. All of the procedures were determined to be robust with "GRAND AVERAGE" values from the smallest value of 0.0346 (25*WM-H*) to the highest value of 0.0685 (25*WHQ-H*), all of which are still within the criterion interval of 0.025 to 0.075.

Table 4.1

The Type I error rates for J = 2

Type of Distribution	**Sample Size**		**Variance**		**Natural Pairing**	**15** ***WM-H***	**25** ***WM-H***	**15** ***WHQ-H***	**25** ***WHQ-H***	**15** ***WHQ$_1$-H***	**25** ***WHQ$_1$-H***	***MOM-H***	**Student's** ***t*-test**
g=0; *h*=0	20	20	1	1		**0.0520***	0.0524*	0.0566*	0.0546*	0.0560*	0.0538*	0.0410*	0.0528*
	20	20	1	36		0.0646*	0.0582*	0.0712*	0.0696*	0.0678*	0.0668*	**0.0498***	0.0618*
	15	25	1	1		**0.0498***	0.0488*	0.0588*	0.0532*	0.0588*	0.0520*	0.0366*	0.0490*
	15	25	1	36	+	0.0606*	0.0582*	0.0650*	0.0702*	0.0652*	0.0660*	**0.0496***	0.0198
	15	25	36	1	-	0.0536*	0.0546*	0.0672*	0.0602*	0.0672*	0.0586*	**0.0470***	0.1268
	AVERAGE					0.0561*	0.0544*	0.0638*	0.0616*	0.0630*	0.0594*	0.0448*	0.0620*
g=0; *h*=0.5	20	20	1	1		0.0220	0.0222	0.0292*	0.0220	0.0182	0.0220	0.0214	**0.0356***
	20	20	1	36		0.0276*	0.0304*	**0.0430***	0.0386*	0.0256*	0.0350*	0.0324*	0.0402*
	15	25	1	1		0.0144	0.0226	0.0300*	0.0226	0.0258*	0.0170	0.0232	**0.0374***
	15	25	1	36	+	0.0260*	0.0304*	**0.0424***	0.0330*	0.0252*	0.0242	0.0304*	0.0118
	15	25	36	1	-	0.0198	0.0288*	**0.0410***	0.0374*	0.0402*	0.0256*	0.0330*	0.0996
	AVERAGE					0.0220	0.0269*	0.0371*	0.0307*	0.0270*	0.0248	0.0281*	0.0449*
g=1; *h*=0	20	20	1	1		0.0280*	0.0280*	**0.0476***	0.0320*	0.0326*	0.0320*	0.0312*	0.0358*
	20	20	1	36		**0.0510***	0.0406*	0.1690	0.0962	0.0912	0.0944	0.0544*	0.1226
	15	25	1	1		0.0256*	0.0304*	**0.0524***	0.0430*	0.0454*	0.0302*	0.0300*	0.0382*
	15	25	1	36	+	0.0706*	0.0424*	0.1682	0.1356	0.1200	0.0870	**0.0506***	0.0370*
	15	25	36	1	-	0.0418*	**0.0458***	0.1310	0.0974	0.1322	0.0666*	0.0580*	0.2334
	AVERAGE					0.0434*	0.0374*	0.1136	0.0808	0.0843	0.0620*	0.0448*	0.0934
g=1; *h*=0.5	20	20	1	1		0.0104	0.0158	**0.0288***	0.0164	0.0140	0.0146	0.0200	0.0232
	20	20	1	36		0.0228	0.0222	0.0734*	**0.0472***	0.0348*	0.0384*	0.0398*	0.0434*
	15	25	1	1		0.0100	0.0148	**0.0370***	0.0214	0.0230	0.0134	0.0178	0.0272*
	15	25	1	36	+	0.0254*	0.0206	0.1040	**0.0554***	0.0416*	0.0246	0.0374*	0.0088
	15	25	36	1	-	0.0170	0.0256*	0.0540*	**0.0516***	0.0560*	0.0298*	0.0356*	0.1108
	AVERAGE					0.0171	0.0198	0.0594*	0.0384*	0.0339*	0.0242	0.0301*	0.0427*
GRAND AVERAGE						0.0347*	0.0346*	0.0685*	0.0529*	0.0520*	0.0426*	0.0370*	0.0608*

Notes: (*) liberal criterion; (**bold**) closest to nominal level

UUM Universiti Utara Malaysia

Table 4.2

The Type I error rates for J = 4

Type of Distribution	Sample Size				Variance				Natural Pairing	15 *WM-H*	25 *WM-H*	15 *WHQ-H*	25 *WHQ-H*	15 WHQ_1-*H*	25 WHQ_1-*H*	*MOM-H*	ANOVA *F*-test
g=0; *h*=0	20	20	20	20	1	1	1	1		0.0444*	0.0376*	**0.0510***	0.0472*	0.0522*	0.0462*	0.0256*	0.0518*
	20	20	20	20	1	1	1	36		0.0590*	**0.0540***	0.0668*	0.0638*	0.0614*	0.0626*	**0.0460***	0.1096
	10	15	25	30	1	1	1	1		0.0454*	0.0376*	0.0596*	0.0512*	0.0612*	0.0508*	0.0246	**0.0504***
	10	15	25	30	1	1	1	36	+	0.0580*	0.0572*	0.0640*	0.0682*	0.0618*	0.0626*	**0.0486***	0.0336*
	10	15	25	30	36	1	1	1	-	0.0738*	0.0596*	0.0870	0.0856	0.0916	0.0824	**0.0528***	0.2850
	AVERAGE									0.0561*	0.0492*	0.0657*	0.0632*	0.0656*	0.0609*	0.0395*	0.1061
g=0; *h*=0.5	20	20	20	20	1	1	1	1		0.0060	0.0098	0.0136	0.0062	0.0054	0.0060	0.0078	**0.0336***
	20	20	20	20	1	1	1	36		0.0210	0.0296*	**0.0344***	0.0308*	0.0214	0.0280*	0.0292*	0.0782
	10	15	25	30	1	1	1	1		0.0052	0.0074	0.0138	0.0098	0.0112	0.0074	0.0076	**0.0404***
	10	15	25	30	1	1	1	36	+	0.0220	0.0292*	**0.0408***	0.0378*	0.0254*	0.0248	0.0302*	0.0192
	10	15	25	30	36	1	1	1	-	0.0290*	0.0184	0.0440*	0.0452*	**0.0454***	0.0452*	0.0274*	0.2392
	AVERAGE									0.0166	0.0189	0.0293*	0.0260*	0.0218	0.0223	0.0204	0.0821
g=1; *h*=0	20	20	20	20	1	1	1	1		0.0112	0.0112	0.0368*	0.0166	0.0172	0.0158	0.0134	**0.0432***
	20	20	20	20	1	1	1	36		0.0484*	0.0384*	0.1712	0.0962	0.0912	0.0952	**0.0512***	0.2448
	10	15	25	30	1	1	1	1		0.0134	0.0116	0.0270*	0.0208	0.0262*	0.0178	0.0140	**0.0442***
	10	15	25	30	1	1	1	36	+	0.0686*	0.0390*	0.2068	0.1538	0.1594	0.1096	**0.0476***	0.1278
	10	15	25	30	36	1	1	1	-	**0.0574***	0.0368*	0.1100	0.1060	0.1086	0.1022	0.0616*	0.3804
	AVERAGE									0.0398*	0.0274*	0.1104	0.0787	0.0805	0.0681*	0.0376*	0.1681
g=1; *h*=0.5	20	20	20	20	1	1	1	1		0.0024	0.0040	0.0120	0.0040	0.0042	0.0040	0.0070	0.0226
	20	20	20	20	1	1	1	36		0.0146	0.0184	**0.0646***	0.0342*	0.0236	0.0278*	0.0336*	0.0918
	10	15	25	30	1	1	1	1		0.0034	0.0042	0.0128	0.0092	0.0098	0.0056	0.0052	**0.0376***
	10	15	25	30	1	1	1	36	+	0.0174	0.0182	0.1000	0.0578*	**0.0468***	0.0272*	0.0334*	0.0278*
	10	15	25	30	36	1	1	1	-	0.0188	0.0126	**0.0406***	0.0354*	0.0396*	0.0354*	0.0314*	0.2214
	AVERAGE									0.0113	0.0115	0.0460*	0.0281*	0.0248	0.0200	0.0221	0.0802
GRAND AVERAGE										0.0310*	0.0267*	0.0628*	0.0490*	0.0482*	0.0428*	0.0299*	0.1091

Notes: (*) liberal criterion; (**bold**) closest to nominal level

Universiti Utara Malaysia

4.2.2 Type I Error Rates for the Four-Group Test (*J* = 4)

Beside the Two-group test ($J = 2$), this study also focused on the Four-group test ($J = 4$) where the Type I error rate result is illustrated in Table 4.2. Similar to the Two-group test, four types of distribution with five conditions generated from the manipulation of sample sizes and variances were also used to evaluate the robustness of the proposed proposal in the Four-group test.

4.2.2.1 Standard normal distribution (*g* = 0; *h* = 0)

Under standard normal distribution ($g = 0$; $h = 0$), the "AVERAGE" values of all the procedures were found to fulfil Bradley's robustness criterion and were considered robust except for the ANOVA *F*-test, which yielded liberal "AVERAGE" values. Looking into each condition, all procedures were found robust under balanced sample sizes with equal variances. However, 15*WHQ-H* was determined to be the best procedure under this condition because it was able to provide a Type I error rate closest to the nominal level of $\alpha = 0.05$. All of the procedures yielded Type I error rates that fell within the robustness criterion interval except for the ANOVA *F*-test where its Type I error rates fell out of the criterion interval under unequal variances even with balanced sample sizes.

UUM Universiti Utara Malaysia

The *MOM-H* failed to fulfil the robustness criterion under the condition of unbalanced sample sizes with equal variances, whereas others procedure performed better with Type I error rates from 0.0376 to 0.0612. However, all of the procedures performed well under the positive pairing condition giving Type I error rates that fell within the criterion interval. 15*WHQ-H*, 25*WHQ-H*, 15*WHQ*$_1$*-H*, and 25*WHQ*$_1$*-H* showed slightly poor performances under the negative pairing condition yielding liberal Type I error rates. The ANOVA *F*-test performed the worst under this condition, providing a Type I error rate of 0.2850. Under this negative pairing condition, 15*WM-H*, 25*WM-H*, and *MOM-H* were found to be robust and were also able to yield Type I error rates within the interval of 0.025 to 0.075.

4.2.2.2 Symmetric heavy tailed distribution (*g* = 0; *h* = 0.5)

Only the classical procedure, the ANOVA *F*-test, was considered robust as its Type I error rates fell under the robustness criterion with symmetric heavy tailed distribution ($g = 0$; $h = 0.5$) and equal variances, either paired with balanced or unbalanced sample sizes. The rest of the procedures failed the criterion, providing conservative Type I error rates. When the variances became unequal, all of the procedures were still able to control their Type I error rates so that they fell within the interval of 0.025 and 0.075, except for 15*WM-H*, 15WHQ_1-*H*, and the ANOVA *F*-test under balanced sample sizes.

When unbalanced sample sizes were paired with unequal variances, all of the procedures were found to be robust, yielding robust Type I error rates except for 15*WM-H*, 25WHQ_1-*H*, and the ANOVA *F*-test that had slightly poor performances under the positive pairing condition. For the negative pairing condition, only 25*WM-H* and the ANOVA *F*-test did not perform well with poor Type I error rates. For "AVERAGE" values, only 15*WHQ-H* and 25*WHQ-H* performed well under a symmetric heavy tailed distribution, providing "AVERAGE" values within the robustness criterion interval. 15*WHQ-H* was also the one procedure, which had an "AVERAGE" Type I error rate closest to the significance level of $\alpha = 0.05$.

4.2.2.3 Skewed normal tailed distribution (*g* = 1; *h* = 0)

Under skewed normal tailed distribution ($g = 1$; $h = 0$), the Type I error rates for all procedures fell below the nominal level of $\alpha = 0.05$ under balanced sample sizes paired with equal variances. However, none of them fulfilled Bradley's robustness criterion except for 15*WHQ-H* and the ANOVA *F*-test with Type I error rates of 0.0368 and 0.0432, respectively. When the variances became unequal, both these procedures performed the worst and only 15*WM-H*, 25*WM-H*, and *MOM-H* performed well and were considered robust.

15*WHQ-H*, 15*WHQ*$_1$*-H*, and the ANOVA *F*-test yielded Type I error rates within the robustness criterion interval under unbalanced sample sizes paired with an equal variances condition, whereas the other procedures performed with conservative Type I error rates. The three above-mentioned procedures yielded poor robustness under unbalanced sample sizes paired with unequal variances regardless of positive or negative pairing. Under these positive and negative pairing conditions, 15*WM-H*, 25*WM-H*, and *MOM-H* performed well, producing Type I error rates from 0.0368 to 0.0616, which fulfilled the robustness criterion.

Universiti Utara Malaysia

Under this skewed normal tailed distribution, 15*WM-H*, 25*WM-H*, 25*WHQ*$_1$*-H*, and *MOM-H* performed well in controlling their "AVERAGE" Type I error rates and were found to be robust based on Bradley's robustness criterion. The "AVERAGE" Type I error rates for other procedures under this type of distribution performed liberally and poor robustness was noted.

4.2.2.4 Skewed heavy tailed distribution (*g* = 1; *h* = 0.5)

None of the procedures were found to be robust under skewed heavy tailed distribution (*g* = 1; *h* = 0.5) with balanced sample sizes and equal variances. Plus, all produced conservative Type I error rates. Only the ANOVA *F*-test fulfilled the robustness criterion and was considered robust when the sample sizes became unbalanced. Other procedures failed the criterion under unbalanced sample sizes paired with equal variances. 15*WHQ-H*, 25*WHQ-H*, 25*WHQ*$_1$-*H*, and *MOM-H* became robust under unequal variances paired with balanced sample sizes. The Type I error rates for other procedures fell out of the robustness criterion interval and were considered not robust under this condition.

All of the procedures were found to be robust under the positive paring conditions except for 15*WM-H*, 25*WM-H*, and 15*WHQ-H* where 15*WM-H* and 25*WM-H* provided conservative Type I error rates, whilst 15*WHQ-H* performed liberally in terms of Type I error rates. With negative paring, 15*WM-H*, 25*WM-H*, and the ANOVA *F*-test did not satisfy the robustness criterion, whereas other procedures were considered robust according to the criterion. From the "AVERAGE" values, only 15*WHQ-H* and 25*WHQ-H* were considered robust under a skewed heavy tailed distribution with Type I error rates of 0.0460 and 0.0581, respectively. . The other procedures gave conservative Type I error rates except for the ANOVA *F*-test, which yielded liberal Type I error rates.

Overall, all of the procedures except for the ANOVA *F*-test were found to be robust in terms of "GRAND AVERAGE" values, which is the average of all Type I error rates from each manipulated condition of distribution type, sample size, and group variance. All the "GRAND AVERAGE" Type I error rates ranged from 0.0267 to 0.0628, which still conformed to Bradley's robustness criterion interval except for the ANOVA *F*-test, which yielded a liberal Type I error rate (0.1091).

Universiti Utara Malaysia

4.3 Power Analysis

The analysis of power in statistical testing was another assessment considered in this study. Similar to the Type I error rates, this analysis uses the same condition designs. However, the setting of the central tendency measures was determined differently, as mentioned in Section 3.7 where these settings were based on the conditions of design, effect sizes, and pattern variability. For each design, three types of effect sizes (small, medium, and large) were chosen to evaluate the compared procedures. For the Two-group test, the three levels of selected effect sizes were small, $f = 0.20$, medium, $f = 0.50$, and large, $f = 0.80$. The small, $f = 0.10$, medium, $f = 0.25$, and large effect sizes, $f = 0.40$, were selected for the Four-group test. Regarding the pattern variability based on Cohen (1988), this study only focused on the maximum pattern variability.

UUM
Universiti Utara Malaysia

According to previous studies, a procedure is considered a high power procedure when it produces power of more than 80% and an accepted power performance must be at least more than 50% (Cohen, 1992a; Cohen, 1992b; Murphy, Myors & Wolach, 2008). Thus, these two criterions were chosen for the statistical power analysis of this study where the minimum accepted power is 50% marked with *, and a value more than 80% is considered high power and would be marked with bold font, as illustrated in Appendix A and Appendix B. However, the focus of this study is on Type I error rates rather than statistical power because the ability of the procedures to control Type I error rates is as important as the ability of the procedures to control the probability of rejecting the null hypothesis when it is true. The aim of the power analysis in this study is to assess the probability of rejecting null hypothesis when it is false, i.e. to assess the sensitivity of the procedures in detecting a statistically significant result (Cohen, 1992b). Thus, Figure 4.1 to Figure 4.8 only shows the power performance of the procedures that have satisfied Bradley's robustness criterion and considered robust under the designed test conditions.

4.3.1 Power of Two-Group Test (*J* = 2)

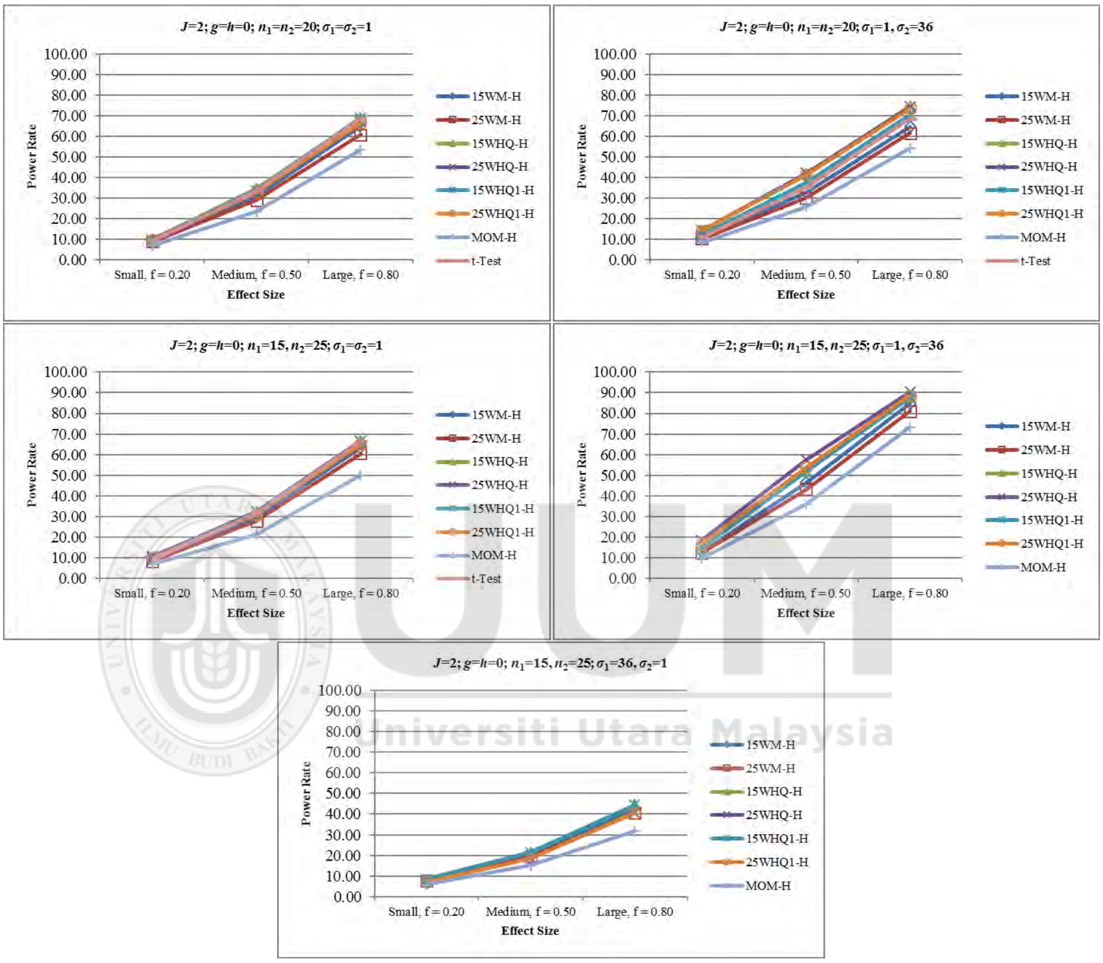
Figure 4.1 to Figure 4.4 shows the power rate performances for the Two-group test with four types of distribution and designed conditions that are similar to the assessment of Type I error rates (as outlined in Section 4.2).

4.3.1.1 Standard normal distribution (*g* = *h* = 0)

The power analysis of this study will only focus on the robust procedures, as mentioned above. Thus, Figure 4.1 only shows the statistical power for robust procedures that were evaluated under standard normal distribution (*g* = *h* = 0). Overall, the statistical power of all procedures increased with an increase in effect size from small to large, and all of the procedures were able to achieve the accepted power performance, i.e. 50%, when effect size was large, except for in the negative pairing condition. Under balanced sample sizes paired with equal variances condition, all procedures had comparable performances ranging from 65% to 69% under the large effect size, except for 25*WM-H*, with a slightly poor result of 61%. The poorest performance procedure with 53% was *MOM-H*. When variances became unequal, the majority of the procedures observed an improvement but still could not satisfy the high power criterion, as the highest obtainable power was only 75%, produced by 15*WHQ-H* under a large effect size. *MOM-H* was still able to provide the lowest power of 54% under this condition with a large effect size.

Universiti Utara Malaysia

The majority of the power results became slightly worse when sample sizes became unbalanced even when the variances were still homogenous. All of the procedures were only able to fulfil the accepted power criterion under a large effect size with the lowest power being 50% and the highest being 67% produced by *MOM-H* and 25*WHQ-H*, respectively.



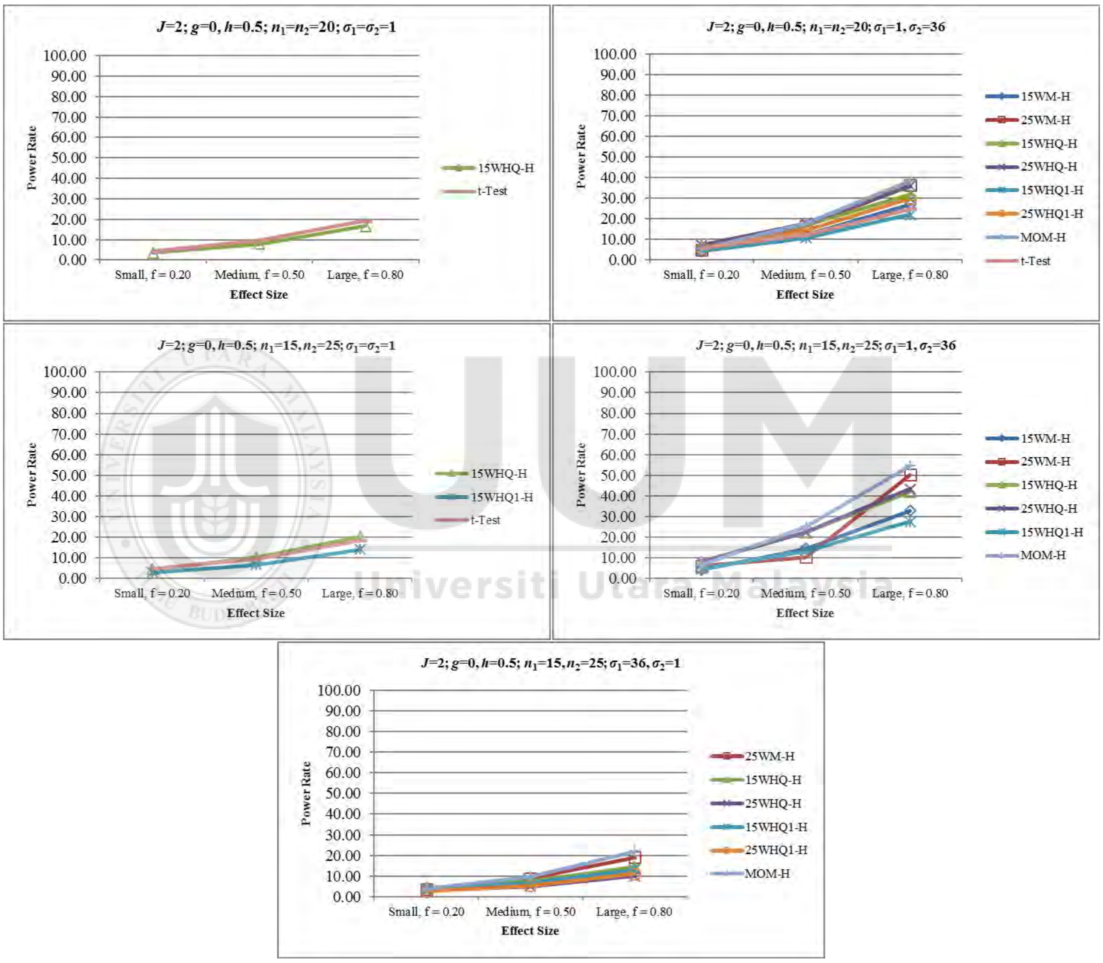
Universiti Utara Malaysia

Figure 4.1. The statistical power for $J = 2$, $g = h = 0$

All of the procedures observed better power performance under positive pairing. 15*WHQ-H*, 15*WHQ*1-*H*, 25*WHQ*1-*H*, and 25*WHQ-H* met the 50% power criterion ranging from 52% to 58% under a medium effect size. When the effect size became larger, all procedures were considered high power procedures with more than 80% power except for *MOM-H* with 73% power, albeit it was still considered accepted power procedure.

With negative pairing, all the procedures yielded the worst power performance under a standard normal distribution with none of the procedures achieving at least the accepted power performance. The highest obtainable power was a mere 45%, produced by 15*WHQ-H* and 15*WHQ*₁-*H*.

4.3.1.2 Symmetric heavy tailed distribution (*g* = 0; *h* = 0.5)



Universiti Utara Malaysia

Figure 4.2. The statistical power for $J = 2$, $g = 0$; $h = 0.5$

Under a symmetric heavy tailed distribution (*g* = 0; *h* = 0.5), all of the procedures performed poorly for all test conditions even under a large effect size, as shown in Figure 4.2. All of the procedures were considered to have low power, as none of them were able to provide at least 50% power, except for 25*WM-H* and *MOM-H*, which were able to produce the accepted power of 50% and 55%, respectively, under a positive pairing condition with a large effect size. A side observation of 25*WM-H* shows that this procedure achieved a power rate that was slightly increased from a small to medium effect size and then drastically increased with a large effect size.

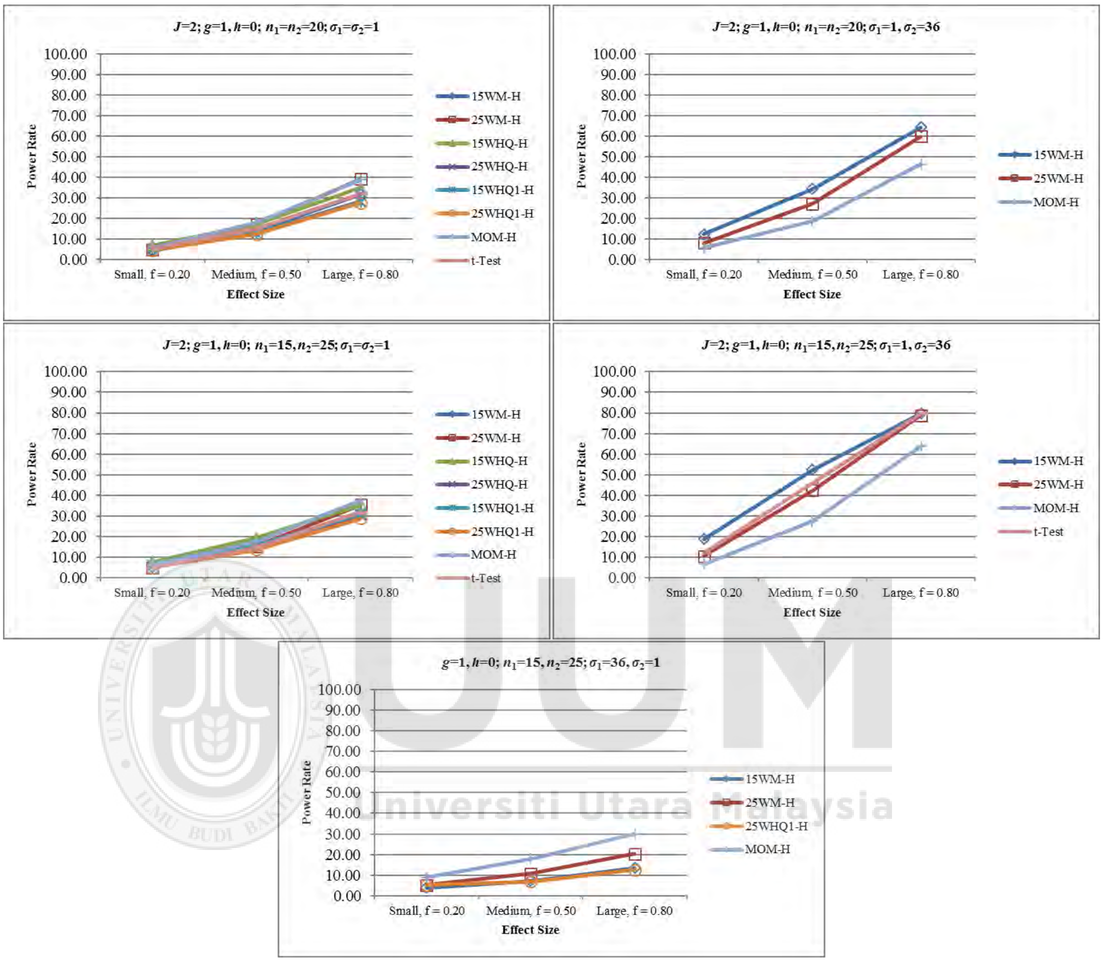
Beside the positive pairing condition, the best obtainable power was around 20%, produced by 15*WHQ-H* and the Student's *t*-test under equal variances regardless of balanced or unbalanced sample sizes. The best power with around 20% was also observed for 25*WM-H* and *MOM-H* under the negative pairing condition with large effect sizes. Under the condition of balanced sample sizes paired with unequal variances, the power performance was observed to be slightly better with a value close to 40% for 25*WM-H*, 25*WHQ-H*, and *MOM-H* under a large effect size.

4.3.1.3 Skewed normal tailed distribution (*g* = 1; *h* = 0)

Universiti Utara Malaysia

None of the compared procedures were able to fulfil the accepted power of 50% under a skewed normal tailed distribution (*g* = 1; *h* = 0) with balanced sample sizes and paired with equal variances, as shown in Figure 4.3. The best procedure, 25*WM-H*, gave 39% power under a large effect size and this was the highest power that could be obtained among the procedures. 15*WM-H* and 25*WM-H* were able to perform at least to a power of 50%, which is the accepted power under a large effect size when the variances were unequal even with the sample sizes still balanced. However, both procedures were still unable to satisfy the high power procedures, as the highest power obtained was only 64% (by 15*WM-H*).

When it comes to unbalanced sample sizes paired with equal variances and large effect sizes, the highest power obtained (by *MOM-H*) was only 38%, which is still considered a low power procedure. This means that all procedures were considered low power procedures as none of them were able to achieve the accepted power of 50% under this condition.



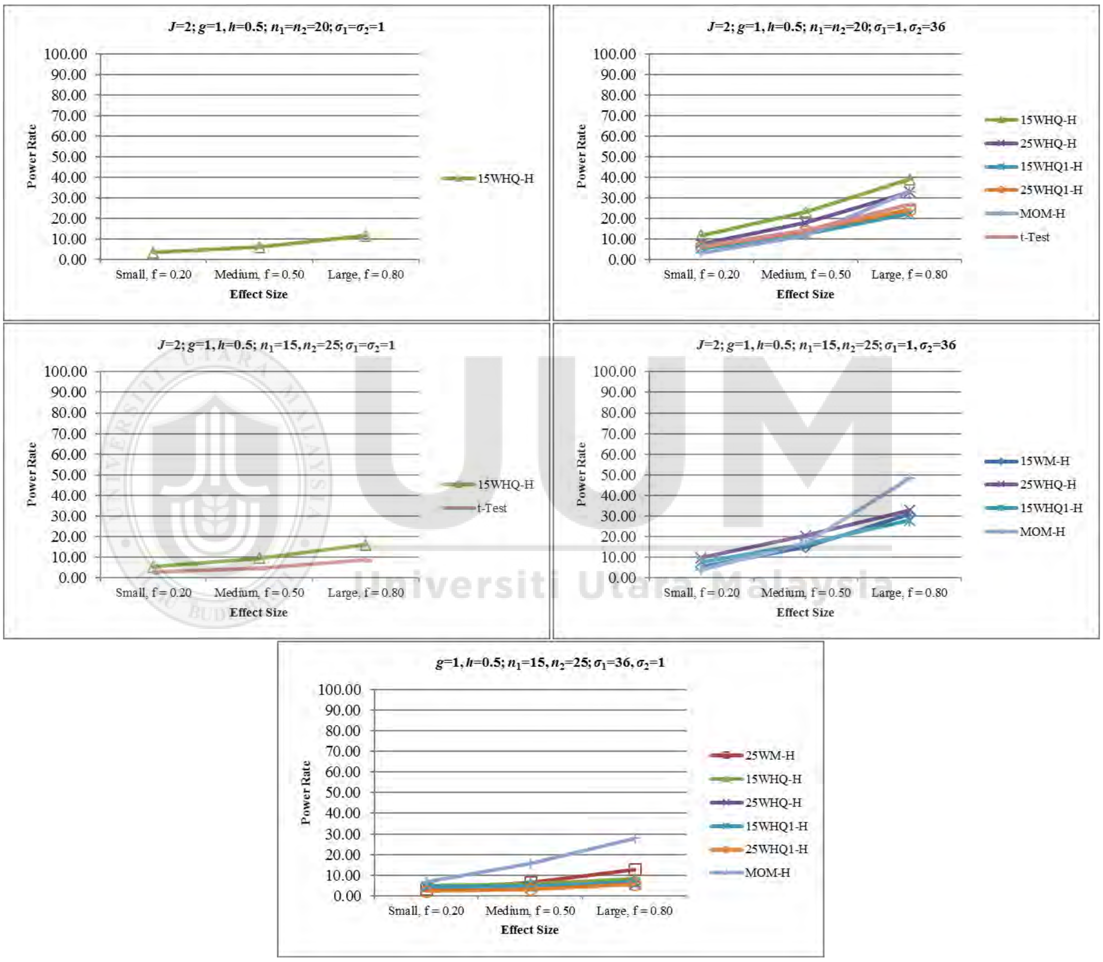
Universiti Utara Malaysia

Figure 4.3. The statistical power for J = 2, g = 1; h = 0

All of the procedures yielded acceptable but not high power procedures under the positive pairing condition and with a large effect size. However, three procedures that the 15*WM-H*, 25*WM-H* and *t*-test able to obtained the power rate as high as closed to the high power, 80% and the best procedure under this condition was 15*WM-H*. This procedure was able to fulfil the accepted power criterion with a medium effect size, giving 52% and 79% power, which is very close to the high power criterion of 80% with a large effect size.

Under the negative pairing condition, all procedures provided a poor performance with the highest power of 30%, provided by *MOM-H* under a large effect size. Under this condition, the trend of the *t*-test did not behave as per the usual trend where the power rate would usually increase with increasing effect size.

4.3.1.4 Skewed heavy tailed distributions (*g* = 1; *h* = 0.5)



Universiti Utara Malaysia

Figure 4.4. The statistical power for J = 2, g = 1; h = 0.5

From Figure 4.4, the power performances of the procedures were observed to be poor when the distribution became skewed with a heavy tailed (g = 1; h = 0.5) distribution compared to the skewed normal tailed distribution. 15*WHQ-H* was the only robust procedure in terms of Type I error rate under balanced sample sizes paired with equal variances condition, but its power performance was similarly poor, as it was only able to produce 11% power even under a large effect size. Its power became slightly better at 16% under a large effect size and was the best compared to other procedures when the sample sizes became unbalanced.

The highest power under the condition of balanced sample sizes and unequal variances was 12%, 23%, and 39% for small, medium, and large effect sizes, respectively, obtained by 15*WHQ-H*. However, even though these powers failed to satisfy the accepted power criterion of 50%, these were the best power performance compared to other procedures that had been evaluated under the same condition. Under this condition, all the procedures were considered low power procedures with the best power provided by 15*WHQ-H*, as discussed above.

UUM Universiti Utara Malaysia

Under the positive pairing condition, *MOM-H* was found to be the best procedure with 48% power under large effect size. However, this value still does not meet the accepted power criterion, but it is very close. *MOM-H* was also the best procedure, yielding 28% power, but is still considered a low power procedure under the negative pairing condition and a large effect size. The rest of the procedures performed poorly with approximately a 20% gap with *MOM-H* for both positive pairing and negative pairing conditions.

4.3.2 Power of Four-Group Test (J = 4)

The power analysis of the Four-group test (J = 4) was also considered in this study with similar designed conditions to the Two-group test (J = 2). The main difference between the Two-group test and the Four-group test was that the effect sizes used to determine the setting of the central tendency measure in the latter were f = 0.10, 0.25, and 0.40 for small, medium, and large effect sizes, respectively, as mentioned in

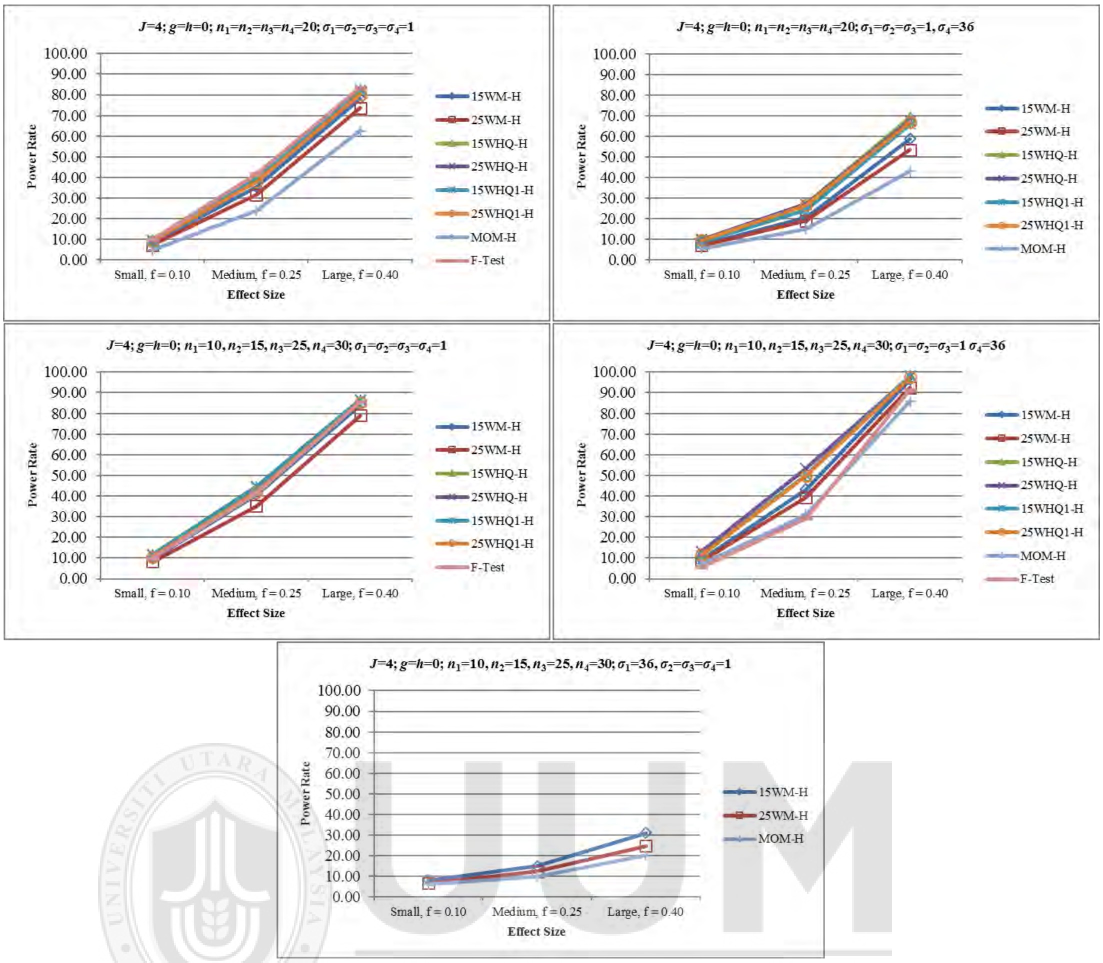
Section 3.7. The result of the power assessment for the Four-group test is illustrated in Figure 4.5 to Figure 4.8 where they are evaluated with a criterion that more than 80% correlates to high power and at least 50% is the accepted power performance (Cohen, 1992a; Cohen, 1992b; Murphy, Myors & Wolach, 2008).

4.3.2.1 Standard normal distribution (*g* = 0; *h* = 0)

Figure 4.5 illustrates the comparative power performances of the procedures under *g* = *h* = 0, a standard normal distribution with five designed conditions manipulated based on sample sizes and variances. Under balanced sample sizes and equal variances, most of the procedures were able to obtain a high power where the powers of the procedures were more than 80% at a large effect size. The procedures considered to have a high power were 15*WHQ-H*, 25*WHQ-H*, 15*WHQ*$_1$*-H*, 25*WHQ*$_1$*-H*, and the ANOVA *F*-test, whereas the rest of the procedures were considered to have acceptable power procedures, as they were still able to produce more than 50% power. The power performance of the procedures reduced when the variances became unequal. However, the procedures were still able to keep their power to at least more than the accepted power of 50% at a large effect size, except for *MOM-H*, which dropped in power to 43%, and was considered a low power procedure under balanced sample sizes paired with unequal variances.

UUM Universiti Utara Malaysia

When looking at the condition of unbalanced sample sizes paired with equal variances, most of the procedures were able to provide a high power from 84% to 87% except for 25*WM-H*, which yielded a power of 79% under a large effect size. Although 25*WM-H* did not satisfy the high power criterion, it was still considered an accepted power procedure as its power rate was more than 50% and was very close to the 80% high power criterion.



J=4; g=h=0; n_1=n_2=n_3=n_4=20; σ_1=σ_2=σ_3=σ_4=1

J=4; g=h=0; n_1=n_2=n_3=n_4=20; σ_1=σ_2=σ_3=1, σ_4=36

J=4; g=h=0; n_1=10, n_2=15, n_3=25, n_4=30; σ_1=σ_2=σ_3=σ_4=1

J=4; g=h=0; n_1=10, n_2=15, n_3=25, n_4=30; σ_1=σ_2=σ_3=1 σ_4=36

J=4; g=h=0; n_1=10, n_2=15, n_3=25, n_4=30; σ_1=36, σ_2=σ_3=σ_4=1

Figure 4.5. The statistical power for J = 4, g = h = 0

All of the procedures were able to meet the high power criterion under positive pairing with a large effect size. The lowest power, 86%, was produced by *MOM-H* and the highest power was produced by 15*WHQ*1-*H* at 98%. 15*WHQ-H*, 15*WHQ*1-*H*, and 25*WHQ-H* were also able to provide accepted powers of 50%, 50% and 53%, respectively, when the effect size was a medium. In this condition, 15*WHQ*1-*H* was considered the best procedure, as it was able to produce accepted power at a medium effect size, and also obtaining the highest power at a large effect size.

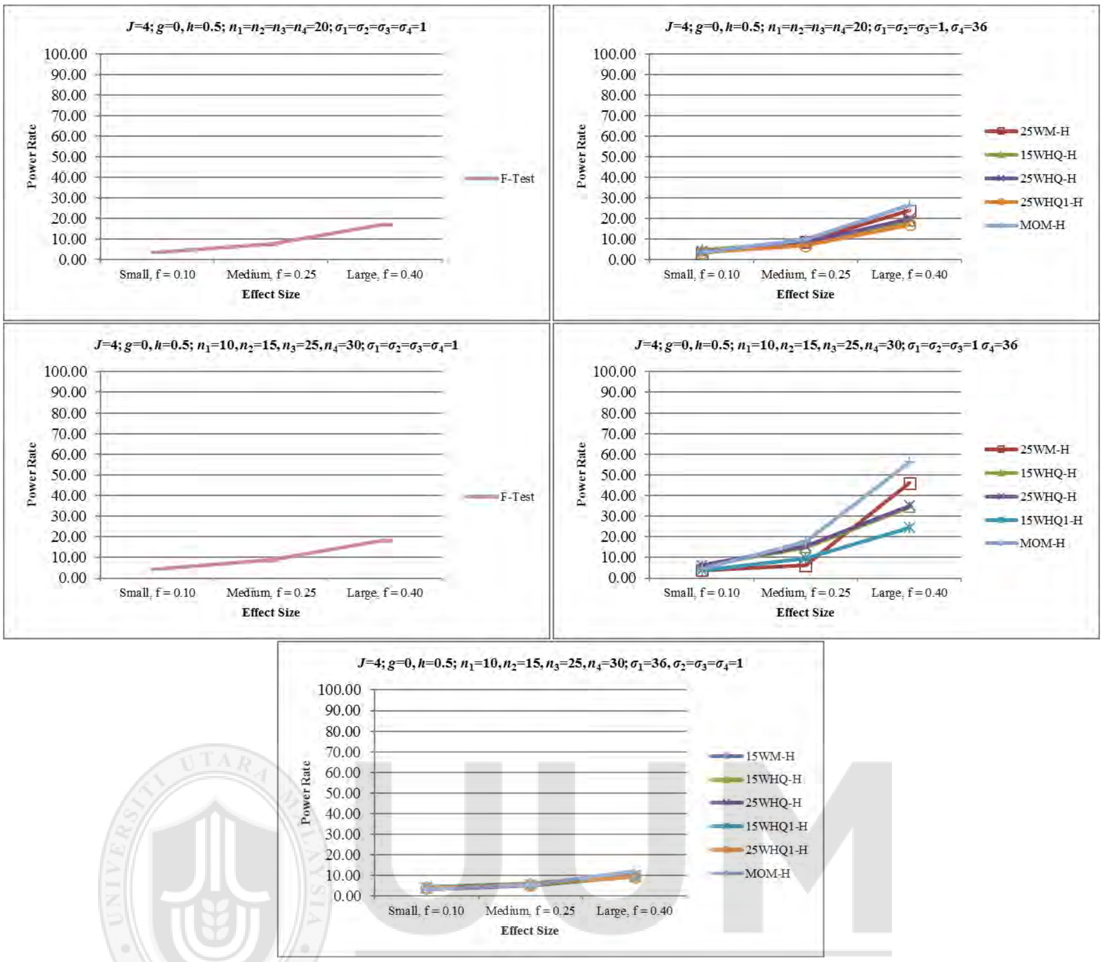
Under the negative pairing condition, only three procedures, 15*WM-H*, 25*WM-H*, and *MOM-H*, were found to be robust in terms of Type I error rates. However, all of the procedures were considered low power procedures, as the highest obtainable power was only 31%, given by 15*WM-H* under a large effect size.

4.3.2.2 Symmetric heavy tailed distribution (*g* = 0; *h* = 0.5)

Overall, it could be observed that the power performance of all the procedures were poor under a symmetric heavy tailed distribution, as shown in Figure 4.6. Under the condition where equal variances are paired with either balanced or unbalanced sample sizes, all procedures were non-robust with conservative Type I error rates except for the ANOVA *F*-test, but the highest power of ANOVA *F*-test was only 17% and 18% with a large effect size and is considered a low power rate under both equal variances conditions.

UUM Universiti Utara Malaysia

Under balanced sample sizes paired with unequal variances, the procedures gave low power, within the 17% to 27% range, under a large effect size. *MOM-H* was considered the best procedure with 27% power, which was the highest power among the procedures under this condition. Moving to the positive pairing condition, *MOM-H* was able to satisfy and give an accepted performance of 56% power and it was the only procedure able to achieve accepted power under a large effect size. Besides *MOM-H*, 25*WM-H* was observed to gain a drastic increment in power from a medium to large effect size, and producing 46% power, which is close to the accepted power criterion of 50% under a large effect size. However, none of the procedures could satisfy the accepted power criterion and were considered low power procedures under the negative pairing condition, providing approximately 10% power at a large effect size.



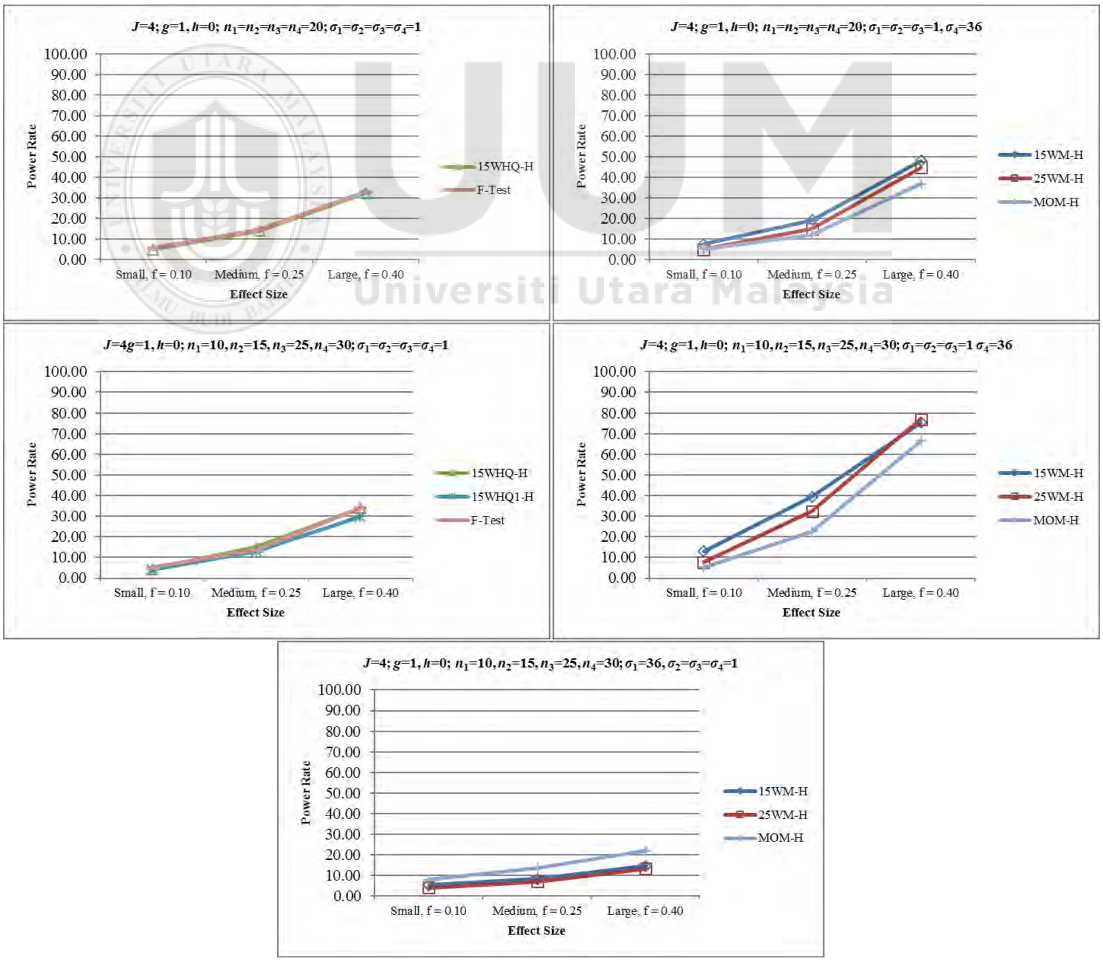
Universiti Utara Malaysia

Figure 4.6. The statistical power for $J = 4$, $g = 0$; $h = 0.5$

4.3.2.3 Skewed normal tailed distribution ($g = 1$; $h = 0$)

Under the skewed normal tailed distribution ($g = 1$; $h = 0$), the ANOVA F-test and 15*WHQ-H* were the only two procedures that were found to be robust in terms of Type I error rates and both had almost similar power performances, 33% and 32%, respectively, under balanced sample sizes paired with equal variances and a large effect size condition, as illustrated in Figure 4.7. However, both of the procedures were considered low power procedures, giving no more than 50% of accepted power. When the sample sizes became unbalanced, the power of both procedures were observed to slightly increase to 34% for the large effect size, but were still considered low power procedures.

For the unequal variances condition, regardless of balanced or unbalanced sample sizes, 15*WM-H*, 25*WM-H*, and *MOM-H* performed well in terms of Type I error rates and were considered robust. In terms of power performances, all of the them gave the poorest power under the negative pairing condition with the highest power of 22% given by *MOM-H* at a large effect size, but it is still considered a low power procedure. Even though it is still considered a low power procedure under balanced sample sizes, the powers of the procedures were observed to become better, giving power close to the accepted power of 50%, especially for 15*WM-H* and 25*WM-H*, which were able to produce 48% and 45% power, respectively, at a large effect size.



Universiti Utara Malaysia

Figure 4.7. The statistical power for J = 4, g = 1; h = 0

Under positive pairing and a large effect size, the powers of the procedures were observed to be the highest among other conditions under the skewed normal tailed distribution. The procedures were able to perform with more than 50%, which is the accepted power performance. 15*WM-H* and 25*WM-H* were the best procedures, which produced power close to but not quite 80% at a large effect size.

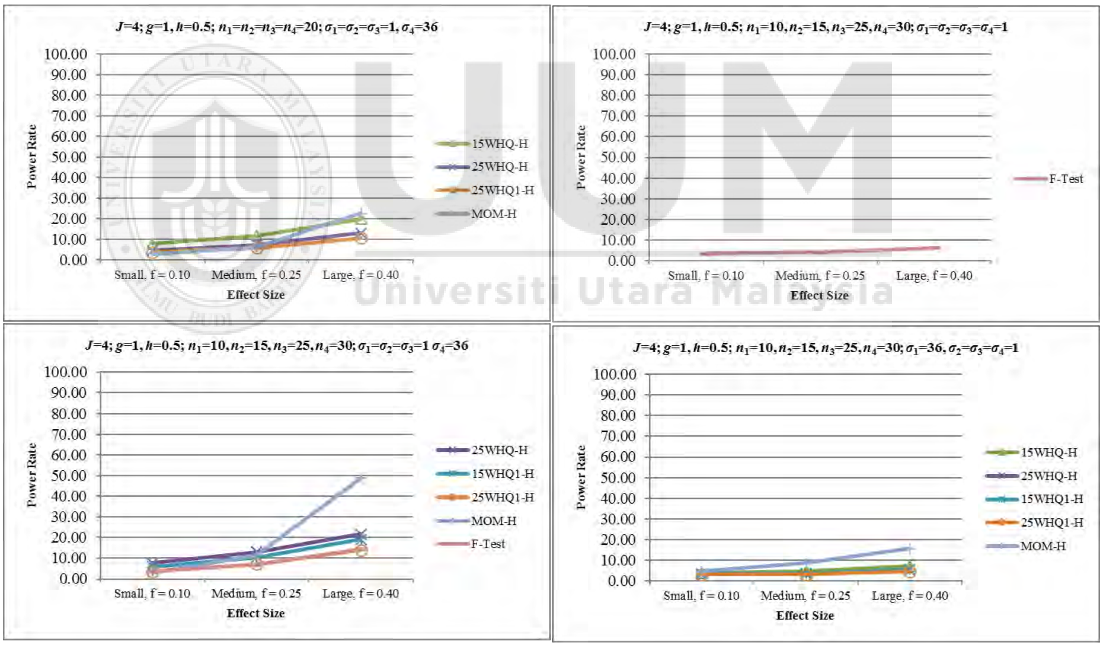
4.3.2.4 Skewed heavy tailed distributions (*g* = 1; *h* = 0.5)

Figure 4.8 shows the power performance of the procedures under a skewed heavy tailed distribution for all designed conditions except for the balanced sample sizes paired with equal variances condition. This is due to the focus of this study, which is the Type I error rates, as mentioned earlier in Section 4.3. Furthermore, all of the procedures provided conservative Type I error rates under balanced sample sizes paired with equal variances and were considered non-robust procedures according to Bradley's robustness criterion, as discussed in Section 4.2.2.4. Thus, Figure 4.8 does not show the balanced sample sizes paired with equal variances condition, as Figure 4.8 only shows the power performance of robust procedures.

Universiti Utara Malaysia

15*WHQ-H* and *MOM-H* were the best procedures even if they were considered low power procedures under balanced sample sizes paired with unequal variances and a large effect size condition with the highest power among the procedures of 20% and 23% power, respectively. Under the condition of unbalanced sample sizes paired with equal variances, only the ANOVA *F*-test was found to be robust but it was still a low power procedure, as it produced only 6% power even under large effect sizes.

MOM-H was the best procedure under both positive and negative pairing conditions, especially under the positive pairing condition. The power trend of other procedures increased gradually from a small to large effect size. On the other hand, the power trend of *MOM-H* drastically increased from a medium to large effect size and yielded the highest power out of all the procedures under a large effect size. However, it was still considered a low power procedure due to its power that was close but not more than the accepted power of 50%. Under the negative paring condition, the best procedure, *MOM-H*, was only able to produce 16% power at a large effect size and thus was also considered a low power procedure. In summary, none of the procedures shown in Figure 4.8 were able to satisfy at least the accepted power criterion and all were considered low power procedures under the skewed heavy tailed distribution.



Universiti Utara Malaysia

Figure 4.8. The statistical power for $J = 4$, $g = 1$; $h = 0.5$

4.4 Summary of Type I Error Rates and Power Analysis

For the overall Type I error rate performance, all of the proposed procedures were able to obtain robust AVERAGE Type I error rates within Bradley's robustness criterion regardless of the number of group size ($J = 2$) or ($J = 4$) under $g = 0$; $h = 0$ (standard normal). When the *g*- and –*h* distribution became $h = 0.5$ (heavy tailed) either with *g*

= 0 (normal) or g = 1 (skewed), only 15*WHQ-H* and 25*WHQ-H* were able to obtain AVERAGE Type I error rates within Bradley's robustness criterion and were considered robust for both J = 2 and J = 4. However, both of the procedures were unable to perform well under g = 1; h = 0 (skewed normal tailed). Under this distribution, the only procedures considered robust under J = 2 and J = 4 were 15*WM-H*, 25*WM-H*, and 25*WHQ*$_1$-*H*.

In regard to J = 2, referring to the Type I error rate under the test conditions, all the proposed procedures (*WM-H* and *AWM-H*) were found to be robust in all test conditions under g = 0; h = 0. 15*WHQ-H* was the only robust procedure in all test conditions when h = 0.5 with g = 0. Under g = 1; h = 0; only *WM-H* was able to perform well within the robust Type I error rates under all conditions considered. When it came to J = 4, only *WM-H* was able to control its Type I error rates well and was considered a robust procedure under g = 0; h = 0. The rest of the procedures were unable to control their Type I error rates within the robustness criterion interval for all test conditions but were still able to control their Type I error rate for certain conditions under any type of distribution. When compared to *MOM-H* and the classical procedures, the proposed procedures had comparable performances with *MOM-H* and performed better compared to the classical procedures when the distribution was non-normal.

Most of the Type I error rates for the proposed procedures including *MOM-H* decreased when the group size increased. The decrement in Type I error rates resulted in a few procedures not meeting the robustness criteria, giving conservative Type I error rates. However, an increment in Type I error rates resulted in certain procedures becoming liberal under certain test conditions and distributions. Most of the non-robust conditions of the procedures were observed to give conservative Type I error

Universiti Utara Malaysia

rates regardless of $J = 2$ or $J = 4$. Liberal Type I error rates were also observed and mostly occurred in conditions of unequal variances, especially under a skewed distribution.

In terms of power analysis, most of the robust procedures were able to achieve an accepted power of 50% under $g = 0$; $h = 0$ regardless of group sizes either two or four ($J = 2$ or $J = 4$) for all conditions except for negative pairing. Under this distribution, the proposed procedures were able to achieve a high power of 80% under the positive pairing condition for $J = 2$. When $J = 4$, the proposed procedures were also able to achieve a high power under equal variances regardless of balanced or unbalanced sample sizes and the positive pairing condition was the only condition in which all procedures could be considered high power procedures.

Under $g = 0$; $h = 0.5$, *MOM-H* was the only robust procedure that was able to achieve an accepted power of 50%, and it was also the procedure with the highest power under positive pairing for both $J = 2$ and $J = 4$. When the distribution became skewed, the robust procedures were able to achieve accepted power under the positive pairing condition for both $J = 2$ and $J = 4$ but only *WM-H* produced an accepted power under balanced sample sizes paired with unequal variances for $J = 2$. None of the robust procedures were able to obtain at least an accepted power of 50% under the extreme non-normal distribution where $g = 0$; $h = 0.5$ (skewed heavy tailed distribution).

4.5 Real data analysis

As mentioned in Section 1.3, real data analysis was done to verify the validity of the proposed procedures. The selected data was run through a crystallization process to improve the Mean Aperture (M.A.) of refined sugar from a sugar manufacturing plant.

Universiti Utara Malaysia

The M.A. is the sugar grain size, which is an important KPOV (Key Process Output Variable) for a sugar manufacturing plant. It is used to produce the customers' desired grain size. The different grain sizes will have different usage to customers and an incorrect production of sugar grain size will cause it to be reprocessed.

The M.A. obtained from an experiment with four conditions and 14 data had collected by systematic sampling throughout 23 tons of bulk production from each condition. It was observed that the means of M.A. hovered within the 0.9 to 1.04 mm range, as shown in Table 4.3. The mean is a statistical value, which is used to identify how the data would look like. Besides the mean, other important statistical values were also identified, as shown in the descriptive statistics of refined sugar M.A. in Table 4.3. The identification of the behaviour of the data is important prior to conducting any statistical test on the data. The selection of the statistical test for data analysis must be done carefully, as inadequate tests may impact the analysis results. Furthermore, the impacted results could be misleading and result in misinterpretation and the wrong decision being made.

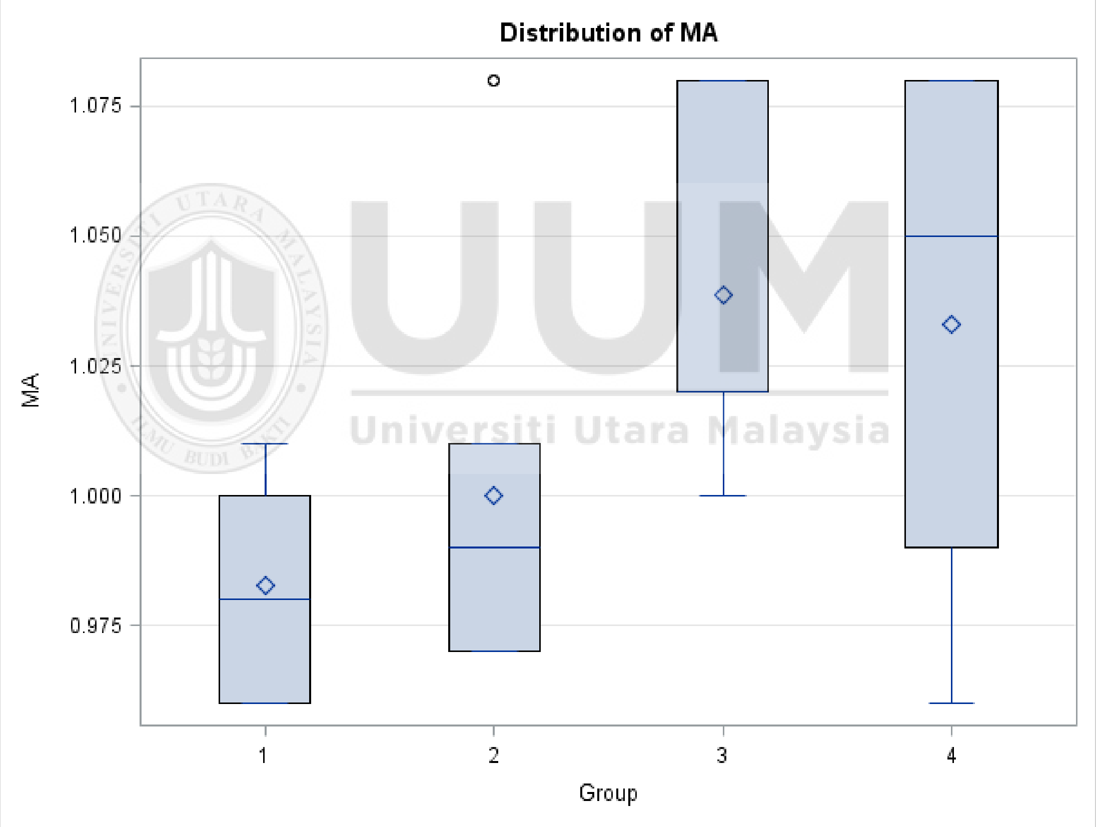
Universiti Utara Malaysia

Table 4.3

Some descriptive statistics on the M.A. of refined sugar

Group	*n*	Mean	Standard Deviation	Skewness	Kurtosis	Shapiro-Wilks
1	14	0.983	0.020	0.004	-1.771	0.0127*
2	14	1.000	0.039	1.288	0.877	0.0009*
3	14	1.039	0.033	0.499	-1.723	0.0010*
4	14	1.033	0.050	-0.572	-1.475	0.0050*
Total	56	12.175	16.081			

This dataset has a balanced sample size of n_1=n_2= n_3=n_4=14. The highest standard deviation in Group 3 shows that it has the widest range dispersion and has a negative skewness and kurtosis coefficient with values of -0.572 and -1.475, respectively. Group 2 is observed to have the highest positive coefficient of skewness compared to other groups, with a value of 1.288. Overall, each group in this dataset has a non-normal distribution where all of them produced a significant value based on the Shapiro-Wilks test with a p-value of less than 0.05.



UUM Universiti Utara Malaysia

Figure 4.9. The box plot of refined sugar M.A.

Figure 4.9 illustrates the behaviour of the data in box plots. The box plot of each group showed a single whisker or none at all. This means that the data of each group was not-normally distributed and skewed either to the left or to the right with regard to the whisker. There was no whisker in Group 2, with an observed outlier. This outlier caused the mean to shift to the right. Overall, the data was obviously different between the compared groups according to the properties of the M.A. of refined sugar, as illustrated via the box plots.

Besides the assumption of normality, another assumption concerning statistical testing is the homogeneity of variances. The data set was tested with Levene's Test for homogeneity of variances and was found to be significant with a *p*-value, of 0.0026, less than 0.05, as shown in Table 4.4.

Table 4.4

Levene's Test for Homogeneity of refined sugar M.A. Variances

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Group	3	0.00471	0.00157	5.4	0.0026*
Error	52	0.0151	0.000291		

Universiti Utara Malaysia

Table 4.5

The p-value of refined sugar M.A. testing

Statistical test	15 *WM-H*	25 *WM-H*	15 *WHQ-H*	25 *WHQ-H*	15 *WHQ*$_1$*-H*	25 *WHQ*$_1$*-H*	*MOM-H*	ANOVA *F*-test
p-value	0.0017*	0.0000*	0.0000*	0.0000*	0.0000*	0.0000*	0.2020	0.0004*

Based on the results of the Shapiro-Wilks and Levene's Test, the M.A. of refined sugar did not satisfy the assumption of normality and homogeneity of variances. Therefore, this data set was used to validate the performance of the compared procedures under violated assumptions; the *p*-values are outlined in Table 4.5. The *p*-values produced by the compared procedures showed a significant difference with 5% significance level for all procedures except for *MOM-H*, which was unable to detect the difference between the groups with a *p*-value of 0.2020.

Universiti Utara Malaysia

CHAPTER FIVE
DISCUSSIONS AND CONCLUSION

5.1 Introduction

The main objective of this study is to find alternative statistical methods that are able to test the equality of central tendency measures under non-normal data and heterogeneous of variances with higher accuracy. In real life, the non-normal dataset is commonly collected and the conventional statistical tests such as Students' *t*-test and ANOVA *F*-test are usually sensitive to the shape of the data distribution and making these tests unable to perform well, where these tests would be losing their control over the Type I error as well as reduce the power rate. Alternative methods such as non-parametric or transformation methods may be able to address the violation of the assumptions. However, these methods adopt the ranking values or transformed values instead of the original parametric values in the statistical testing which may lead to inaccurate results of the analysis (Siegel, 1957; Rasmussen, 1989). Another alternative method is the robust statistics. In statistical testing, the robust statistical method is considered a powerful method which is able to control its Type I error rate at nominal level and also obtain sufficient power rate, even if the dataset is non-normal or the variances are heterogeneous or both (Erceg-Hurn & Mirosevich, 2008).

Universiti Utara Malaysia

In dealing with the issue variance heterogeneity and non-normality, a few robust statistical methods were proposed, such as Welch test and *H*-statistic. The *H*-statistic was selected for this study due to its simple statistical calculation and ability to perform well under skewed distribution dataset. This robust statistic has better control of the Type I error rate and the statistical power when the central tendency measure of *H*-statistic is replaced by the modified one-step *M*-estimator (*MOM*) by Keselman, Wilcox, Othman, and Fradette (2002) denoted as *MOM-H* and further studied by Othman et al. (2004), and Syed Yahaya (2005).

The *MOM* estimator was one of the robust statistics which was used in *MOM-H* to improve the Type I error rate control and power rate. However, the *MOM* estimator used the trimming approach based on the outlier detection method. The trimming approach may cause the loss of important information especially when the sample size is small or when the outlier is very difficult to detect. Failure to detect outliers may influence the statistical test's power rate (Wilcox, 2003). Nonetheless, these two concerns can be addressed by using the Winsorized mean (*WM*) and adaptive Winsorized mean (*AWM*).

The *WM* winsorizes data symmetrically, while *AWM* winsorizes data assymmetrically based on the shape of the data distribution. Both methods require predetermined percentages of winsorization. However, the *AWM* uses the hinge estimator to identify how much to be winsorized on the left and right tails of the data (Keselman et al., 2007). In this study, the *AWM* adopts the hinge estimators of *HQ* and HQ_1 that was proposed by Reed and Stark (1996), due to their ability to provide better control of Type I error rate in the Welch test as recommended by Keselman et al. (2007).

As the *WM* and *AWM* requires the percentages of the winsorization proportion to be predetermined, there is concern as to how much data can be winsorized. In this study, 15% and 25% were used as the predetermined value of winsorizing due to both are common use as trimming percentages and recommended by various studies such as Huber (1972), Rosenberger and Gasko (1983), Mudholkar, Mudholkar and Srivastava (1991), Wilcox and Keselman (2003b). The 15% and 25% in trimming also provides

Universiti Utara Malaysia

better control of the Type I error rate and high statistical power according to the studies of Keselman et al. (2007), and Abdullah (2011).

Six robust procedures were proposed from the modification of the *H*-statistic by adopting the *WM* and *AWM* using hinge estimator of *HQ* and HQ_1. All of these six proposed procedures, named as 15*WM-H*, 25*WM-H*, 15*WHQ-H,* 25*WHQ-H,* 15WHQ_1-*H* and 25WHQ_1-*H,* were compared to the *MOM-H* and classical test (Student's *t*-test for two groups and ANOVA *F*-test for more than two groups test).

To assess the ability of the compared methods to control the Type I error rate and its power performance, 5000 simulated datasets were generated by the SAS generator RANNOR (SAS, 2011). These datasets were then bootstrapped 599 times using the percentile bootstrap method to test the hypothesis due to the intractability of the sampling distributions of the statistics. The simulated datasets were generated with manipulation of five variables. The five variables include the type of population distribution, the number of groups, sample size, the degree of variance heterogeneity, and the nature of pairing.

Universiti Utara Malaysia

Four types of distributions, generated by *g*- and -*h* distribution, had been used to test the compared methods for the effect of data distribution. The shape of distribution of the *g*- and -*h* distribution was controlled by the *g* and *h* values. Increasing the *g* and *h* values will increase the skewness and kurtosis of data. Therefore, the four conditions of *g* and *h* values used in this study includes $g = 0$; $h = 0$ (standard normal), $g = 0$; $h = 0.5$ (symmetry heavy tailed), $g = 1$; $h = 0$ (skewed normal tailed) and $g = 1$; $h = 0.5$ (skewed heavy tailed).

The heterogeneity of variances effect of the compared methods was also studied by using the unequal variances of 1:36 (1:1:1:36) ratio which was commented as an extreme variance heterogeneity condition in Othman et al. (2004), Syed Yahaya (2005), and Keselman et al. (2007). The sample sizes were also considered as one of the variable to study the effect of Type I error rate control and power rate. The nature pairings of sample sizes and variances effect were studied when the unbalanced sample sizes were paired with unequal variances. Two and four were the selected group sizes to study the performance of the compared methods under different group numbers.

To evaluate the performance of Type I error rate control, the robustness criterion proposed by Bradley (1978) was used. The method is considered robust when the Type I error rate, ρ meets the criterion $\rho \pm 0.025$ where the Type I error rate must be in between the interval of $0.025 \leq \rho \leq 0.075$ with significance level of 0.05. The best method will produce the Type I error rate closest to nominal value of 0.05. In terms of power performance evaluation, the criterion of minimum 50% and more than 80% was selected. The method that produces more than 50% will be considered as the accepted power rate, while the method that produces more than 80% will be considered as the high statistical power method.

Universiti Utara Malaysia

5.2 Type I Error Rates and Statistical Power Analysis

Table 5.1 shows the number of conditions for each procedure that meets the Bradley's robustness criterion (denoted as scores). The total conditions of each procedure is 40 regardless of the number of groups. Table 5.2 shows the number of conditions for the procedure which produce the closest Type I error rate to the nominal value, 0.05, and this procedure will be considered as the best. The number of conditions mentioned will be denoted as a score in both Table 5.1 and Table 5.2.

Table 5.1

Score of the procedures which were robust in terms of Type I error rates

Group Size	Type of Distribution	15 *WM-H*	25 *WM-H*	15 *WHQ-H*	25 *WHQ-H*	15 *WHQ*$_1$-*H*	25 *WHQ*$_1$-*H*	*MOM-H*	*t*-test / *F*-test
J=2	*g*=0; *h*=0	5	5	5	5	5	5	5	3
	g=0; *h*=0.5	2	3	5	3	4	2	3	3
	g=1; *h*=0	5	5	2	2	2	3	5	3
	g=1; *h*=0.5	1	1	4	3	3	2	3	2
	Score	13	14	16	13	14	12	16	11
J=4	*g*=0; *h*=0	5	5	4	4	4	4	4	3
	g=0; *h*=0.5	1	2	3	3	2	2	3	2
	g=1; *h*=0	3	3	2	0	1	0	3	2
	g=1; *h*=0.5	0	0	2	3	2	3	3	2
	Score	9	10	11	10	9	9	13	9
	Total Score	22	24	27	23	23	21	29	20

UUM
Universiti Utara Malaysia

Based on Table 5.1, all proposed procedures, except the Student's *t*-test, within the group size of two scored 5. This result suggests that the procedures are robust for all conditions, except the Student's *t*-test, when the distribution is normal (g = 0; h = 0). When the distribution becomes heavy tailed (g = 0; h = 0.5), only the 15*WHQ-H* obtained the score of 5, hence being the best among the compared procedures. However, under skewed distribution (g = 1; h = 0), the performance of the procedure deteriorated significantly with only 2 conditions meeting the Bradley's robustness criterion. The same outcome was observed for another few procedures including 25*WHQ-H*, 15*WHQ*$_1$-*H*, and 25*WHQ*$_1$-*H*. On the contrast,15*WM-H*, 25*WM-H* and *MOM-H* resulted in better performance under the skew distributed dataset with scores of 5. Nonetheless, the 15*WHQ-H* was also able to perform well under skewed heavy tailed (g = 1; h = 0.5) dataset although performance was slightly poor with scores of 4. Therefore, shows that it performs well as long as the data have heavy tail.

The performance of the procedures generally reduced when the group size was increased to 4. Under g = 0; h = 0, the *WM-H* regardless of the percentages of winsorization are the best as these procedures still scored 5. For other procedures, the performance was slightly poorer with scores of 4 and Student's t-test being lowest with a score of 3. The *WHQ-H* had comparable performance with *MOM-H* under g = 0; h = 0.5 with scores of 3. Under g = 1; h = 0 and g = 1; h = 0.5, the maximum score was 3. Under g = 1; h = 0, the best procedures were *WM-H* and *MOM-H* whereas *AWM-H* with 25% and *MOM-H* were the best under g = 1; h = 0.5.

Table 5.2

Score of the Type I error rates which were robust and closest to nominal level, 0.05 compared among the procedures

UUM Universiti Utara Malaysia

Group Size	Type of Distribution	15 *WM-H*	25 *WM-H*	15 *WHQ-H*	25 *WHQ-H*	15 WHQ_1-H	25 WHQ_1-H	*MOM-H*	t-test / F-test
J=2	g=0; h=0	2	0	0	0	0	0	3	0
	g=0; h=0.5	0	0	3	0	0	0	0	2
	g=1; h=0	1	1	2	0	0	0	1	0
	g=1; h=0.5	0	0	2	3	0	0	0	0
	Score	3	1	7	3	0	0	4	2
J=4	g=0; h=0	0	1	1	0	0	0	3	1
	g=0; h=0.5	0	0	2	0	1	0	0	2
	g=1; h=0	1	0	0	0	0	0	2	2
	g=1; h=0.5	0	0	2	0	1	0	0	1
	Score	1	1	5	0	2	0	5	6
Total Score		4	2	12	3	2	0	9	8

The total score of the proposed procedures were in between the scores of *MOM-H* and classical tests (Student's t-test and ANOVA F-test) with the 15*WHQ-H* being the procedure scoring closest to the *MOM-H* with total scores of 27 (for 15*WHQ-H*) and 29 (for *MOM-H*). However, the 15*WHQ-H* is the best procedure when it comes to the number of conditions that were robust and closest to nominal level, 0.05 as displayed in Table 5.2.

Besides Type I error rate, the statistical power was also evaluated as shown in Table 5.3 where the number of conditions (denoted as score) that met Bradley's robustness criterion and have accepted statistical power (at least 50%). For $J = 2$ case, one of the test conditions for all *AWM* had achieved at least 50% power rate since the effect size under $g = 0$; $h = 0$ were medium and all procedures, except Student's *t*-test, have scores of 4 when the effect size became large. Most of the procedures had low statistical power under non-normal distribution regardless of being skewed, heavy tailed or combination of both. Under $g = 0$; $h = 0.5$, only 25*WM-H* and *MOM-H* were able to obtain score of 1 when the effect size was large. The 15*WM-H* was able to achieve score of 1 when the effect size under $g = 1$; $h = 0$ was medium. It also obtained score of 2 when the effect size was large, similarly for 25*WM-H*. For $g = 1$; $h = 0.5$, none of the procedures were able to provide a power rate of at least 50% even under a large effect size. Overall, the *WM-H* was the best procedure under the group size of two, that scored the highest of 7 as compared to other procedures.

Under the group size of four, all the procedures showed the same performance as group size of two except 25HQ_1-*H* and *MOM-H* where the scores dropped when group size was increased from two to four under $g = 0$; $h = 0$. When the tail of distribution became heavy, only *MOM-H* was able to perform with a statistical power more than 50% with a score equal to 1. Under skewed distribution, only *WM-H* had the same performance as *MOM-H* with a score of 1. When all of the procedures were tested under $g = 1$; $h = 0.5$, similar outcomes were observed in both group sizes of two and four where none of them were able to provide the statistical power of at least 50%.

Table 5.3

Score of statistical power with robust condition more than 50%

Group Size	Type of Distribution	Effect Size	15 *WM-H*	25 *WM-H*	15 *WHQ-H*	25 *WHQ-H*	15 *WHQ*$_1$-*H*	25 *WHQ*$_1$-*H*	*MOM-H*	*t*-test / *F*-test
J=2	*g*=0; *h*=0	Small	0	0	0	0	0	0	0	0
		Medium	0	0	1	1	1	1	0	0
		Large	4	4	4	4	4	4	4	3
	g=0; *h*=0.5	Small	0	0	0	0	0	0	0	0
		Medium	0	0	0	0	0	0	0	0
		Large	0	1	0	0	0	0	1	0
	g=1; *h*=0	Small	0	0	0	0	0	0	0	0
		Medium	1	0	0	0	0	0	0	0
		Large	2	2	0	0	0	0	1	1
	g=1; *h*=0.5	Small	0	0	0	0	0	0	0	0
		Medium	0	0	0	0	0	0	0	0
		Large	0	0	0	0	0	0	0	0
	Score		7	7	5	5	5	5	6	4
J=4	*g*=0; *h*=0	Small	0	0	0	0	0	0	0	0
		Medium	0	0	1	1	1	0	0	0
		Large	4	4	4	4	4	4	2	3
	g=0; *h*=0.5	Small	0	0	0	0	0	0	0	0
		Medium	0	0	0	0	0	0	0	0
		Large	0	0	0	0	0	0	1	0
	g=1; *h*=0	Small	0	0	0	0	0	0	0	0
		Medium	0	0	0	0	0	0	0	0
		Large	1	1	0	0	0	0	1	0
	g=1; *h*=0.5	Small	0	0	0	0	0	0	0	0
		Medium	0	0	0	0	0	0	0	0
		Large	0	0	0	0	0	0	0	0
	Score		5	5	5	5	5	4	4	3
	Total Score		12	12	10	10	10	9	10	7

Universiti Utara Malaysia

Table 5.4

Score of statistical power with robust condition more than 80%

Group Size	Type of Distribution	Effect Size	15 *WM-H*	25 *WM-H*	15 *WHQ-H*	25 *WHQ-H*	15 WHQ_1-*H*	25 WHQ_1-*H*	*MOM-H*	*t*-test / *F*-test
J=2	*g*=0; *h*=0	Small	0	0	0	0	0	0	0	0
		Medium	0	0	0	0	0	0	0	0
		Large	1	1	1	1	1	1	0	0
	g=0; *h*=0.5	Small	0	0	0	0	0	0	0	0
		Medium	0	0	0	0	0	0	0	0
		Large	0	0	0	0	0	0	0	0
	g=1; *h*=0	Small	0	0	0	0	0	0	0	0
		Medium	0	0	0	0	0	0	0	0
		Large	0	0	0	0	0	0	0	0
	g=1; *h*=0.5	Small	0	0	0	0	0	0	0	0
		Medium	0	0	0	0	0	0	0	0
		Large	0	0	0	0	0	0	0	0
	Score		1	1	1	1	1	1	0	0
J=4	*g*=0; *h*=0	Small	0	0	0	0	0	0	0	0
		Medium	0	0	0	0	0	0	0	0
		Large	2	1	3	3	3	3	1	3
	g=0; *h*=0.5	Small	0	0	0	0	0	0	0	0
		Medium	0	0	0	0	0	0	0	0
		Large	0	0	0	0	0	0	0	0
	g=1; *h*=0	Small	0	0	0	0	0	0	0	0
		Medium	0	0	0	0	0	0	0	0
		Large	0	0	0	0	0	0	0	0
	g=1; *h*=0.5	Small	0	0	0	0	0	0	0	0
		Medium	0	0	0	0	0	0	0	0
		Large	0	0	0	0	0	0	0	0
	Score		2	1	3	3	3	3	1	3
	Total Score		3	2	4	4	4	4	1	3

Universiti Utara Malaysia

Comparing the total score between group sizes of four and two, the total score for all procedures in group size of four had lower scores except *WHQ-H* and 15*WHQ*$_1$-*H* that had the same score in both group sizes. According to the total score in Table 5.3, the best procedure was *WM-H* which had the highest score of 12. The *WHQ-H* and *WHQ*$_1$-*H* had comparable scores with *MOM-H* and the poorest was the classical procedures that only got scores of 7.

According to previous studies (Cohen, 1992a; Cohen, 1992b; Murphy, Myors & Wolach, 2008), the statistical power is considered as high when it achieves 80%. Table 5.4 illustrates the compared procedures' statistical power performance by the number of conditions (denoted as score) which were robust and have more than 80%. Most of the procedures were not high power regardless of the group size being two or four as shown in Table 5.4. For group size of two, the procedures were able to produce more than 80% only when $g = 0$; $h = 0$ and only one condition achieved the high power. The high power procedures were the proposed procedures of *WM-H*, *WHQ-H* and *WHQ*$_1$-*H*. When group size was four, the procedures were also able to achieve high power only when $g = 0$; $h = 0$ but with higher scores as compared to group size of two. In this case, the highest scoring procedures were *WHQ-H*, *WHQ*$_1$-*H* and ANOVA *F*-test. Overall, the *WHQ-H* and *WHQ*$_1$-*H* were the best two procedures which had the highest scores compared to others.

Universiti Utara Malaysia

5.3 Implications

As discussed in previous chapters, the main concerns were the non-normality and heterogeneous variances which may impact the Type I error rate and statistical power of classical methods of the hypothesis testing. Therefore, the main goal of this study is to find for alternative methods that can be applied under various data conditions without worrying about violations of the assumption. The *H*-statistic was modified using Winsorized Mean (*WM-H*) and Adaptive Winsorized Mean (*AWM-H*) as the central tendency measure. The results of this study showed that both proposed methods successfully improved the Type I error rate and statistical power even under violated assumption conditions.

Among all the compared procedures, none were considered as the best in all conditions. However, the 15*WHQ-H* can be considered as a good robust method that was evaluated under various test conditions as mentioned in previous chapter. It obtained better Type I error rate and achieved higher power than other procedures. It performed well for heavy tailed distribution especially for Two-group test. However, it is not recommended for skewed distribution with unequal variances conditions as stated in Teh, Abdullah, Syed Yahaya, and Md Yusof (2014). Besides, the others procedures from *AWM-H* are also not recommended for skewed distribution due to *AWM-H* unable to control the Type I error rate well under this distribution.

Universiti Utara Malaysia

For skewed distribution, the *WM-H* or existing *MOM-H* is recommended due to its ability to control the Type I error rate especially for heterogeneous variance conditions. Furthermore, the *WM-H* showed better statistical power as compared to the others. However, the *WM-H* and *MOM-H* have concerns when used for equal variances under Four-group test. These methods provided the conservative Type I error rate when the methods were evaluated under homogenous variances regardless of balanced or unbalanced conditions for Four-group test.

The results of this study showed that each procedure have their own strengths and weaknesses. While one procedure performs well in certain conditions, it may also be the worst in other conditions. As such, more research is required to find better alternatives for robust statistical methods. Nonetheless, the outcome of this study may give some ideas as a starting point.

5.4 Suggestion for Future Research

The modified *H*-statistic in this study was proven to improve the Type I error rate control and statistical power for the dataset with non-normal distribution and heterogeneous variances. However, it is not ideal in all conditions and few weaknesses was identified such as inability of *AWM-H* to perform well under skewed distributed dataset. Furthermore, the statistical power of the proposed procedures can still be improved as the procedures were unable to achieve the high power rate for most of the test conditions even when it produced better power rate compared to existing procedures.

Universiti Utara Malaysia

In this study, only the hinge estimator HQ and HQ_1 were used. However, Reed and Stark (1996) proposed 7 methods which are able to determine the shape of distribution either by tail length or skewness. The HQ and HQ_1 that was used in this study is the tail length based Hinge estimator and proved it is able to perform well under heavy tailed distribution but poor if the distribution is skewed. Therefore, further study on the entire hinge estimator would be able to help to identify the best Hinge estimator in *AWM-H*. Besides the *AWM* using hinge estimator to determine the shape and tailed to be winsorized, other alternative estimator which is able to handle both heavy tailed and skewed distribution is needed.

Another proposed procedure is *WM-H* which needs a predetermined value prior the winsorizing is conducted. The result of this study finds that the Type I error rates of the procedures performed conservatively for those failed to fulfilled the Bradley's robustness criterion. However, the predetermined value observed have impacted the *WM-H*'s Type I error rates. Therefore, a study on the predetermined value can be done to find the predetermined value which is able to optimise the performance of the proposed procedures.

The *WM* and *AWM* showed ability to improve the *H*-statistic performance in terms of Type I error rate and statistical power, but still failed to perform well in a few test conditions such as the unequal variances conditions. Therefore, the use of the *WM* and *AWM* in other statistical procedure is also believed to improve the overall performance of the procedure especially when the procedure is known to perform well under unequal variances conditions.

Further research in finding the best statistical method cannot be stopped and needs to continue. Beside the suggestions above, there are still many improvement opportunities that can be explored. The new and robust statistical methods may have been introduced and this may help in improving the accuracy of the statistical analysis and the right decision can be made accordingly.

.

Universiti Utara Malaysia

REFERENCE

Abdullah, S. (2011). *Kaedah Alexander-Govern menggunakan penganggar teguh dengan pendekatan pangkasan data: Satu kajian simulasi*. (Unpublished Doctoral thesis). Universiti Utara Malaysia, Sintok, Malaysia

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.

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*, 144-152.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

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.

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

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*. New York: Chapman & Hall Inc.

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods. *American Psychologist*, *63*(7)591-601. doi:10.1037/0003-066X.63.7.591

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

UUM Universiti Utara Malaysia

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: Wiley.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1983). *Understanding robust and exploratory data analysis*. New York: 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. (1972). Robust statistics: A review. *The Annals of Mathematical Statistics*, *43*(4), 1041-1067.

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:10.1037/1082989X.13.2.110.

Keselman, H. J., Othman, A. R., Wilcox, R. R., & Fradette, K. (2004). The new and improved two-sample *t* test. *Psychological Science, 15*(1), 47-51.

Keselman, H. J., Wilcox, R. R., Algina, J., Fradette, K., & Othman, A. R. (2004). A power comparison of robust test statistics based on adaptive estimators. *Journal of Modern Applied Statistical Methods*, *3*(1), 27-38.

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

Lix, L. M., and 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), 409-429. 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. (2012). Type I error rates of parametric, robust and nonparametric methods for two groups cases. *World Applied Sciences Journal*, *16*(12), 1815-1819.

UUM Universiti Utara Malaysia

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

Mudholkar, A., Mudholkar, G. S., & Srivastava, D. K. (1991). A construction and appraisal of pooled trimmed-*t* statistics. *Communications in Statistics: Theory and Methods*, *20*(4), 1345-1359. doi:10.1080/03610929108830569

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.

Rasmussen, J. L. (1989). Data transformation, Type I error rate and power. *British Journal of Mathematical and Statistical Psychology*, *42*(2), 203–213. doi:10.1111/j.2044-8317.1989.tb00910.x

Reed, J. F., & Stark, D. B. (1996). Hinge estimators of location: robust to asymmetry. *Computer Methods and Programs in Biomedicine*, *49*(1), 11-17. doi:10.1016/0169-2607(95)01708-9

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

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.

SAS Institute Inc. (2011). *SAS/IML User's Guide version 9.3*. Cary, NC: SAS Institute Inc.

Schrader, R. M., & Hettmansperger, T. P. (1980). Robust Analysis of Variance Based Upon a Likelihood Ratio Criterion. *Biometrika*, *67*(1), 93-101. doi:10.2307/2335321

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

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

Stigler, S. M. (1977). Do robust estimators work with real data? *The Annals of Statistics*, *5*(6), 1055-1098. doi:10.1214/aos/1176343997

Universiti Utara Malaysia

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, Pulau Pinang, 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.

Teh, K.W., Abdullah, S., Syed Yahaya, S. S., & Md Yusof, Z. (2014). Modified H-statistic with adaptive Winsorized mean in two groups test. *AIP Conference Proceedings*, *1602*(1), 1021-1025. doi: 10.1063/1.4882609

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

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. (1947). The generalization of Student's problem when several different population variances are involved. *Biometrika*, *34*(1/2), 28-35. doi:10.2307/2332510

Welch, B. L. (1951). On the comparison of several mean values: an alternative approach. *Biometrika*, *38*(3/4), 330-336. doi:10.2307/2332579

Wilcox, R. R. (1994). A one-way random effects model for trimmed means. *Psychometrika*, *59*(3), 289-306. doi:10.1007/BF02296126

Wilcox, R. R. (2003). *Applying contemporary statistical technique*. San Diego, CA: 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 one-way ANOVA based on a modified one-step M-estimator. *British Journal of Mathematical and Statistical Psychology*, *56*(1), 15-26. doi:10.1348/000711003321645313

UUM Universiti Utara Malaysia

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. *The British Psychological Society*, *53*(1), 69-82. doi:10.1348/000711000159187

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

UNIVERSITI UTARA MALAYSIA
ILMU BUDI BAKTI
UUM
Universiti Utara Malaysia

APPENDIX A

The statistical power for $J = 2$

Type of Distribution	Sample Size		Variance		Natural Pairing	Effect Size	15 *WM-H*	25 *WM-H*	15 *WHQ-H*	25 *WHQ-H*	15 WHQ_1*-H*	25 WHQ_1*-H*	*MOM-H*	Student's *t*-test
g=0; *h*=0	20	20	1	1		Small, *f* = 0.20	9.04	8.96	10.08	9.86	9.76	9.64	6.96	9.16
						Medium, *f* = 0.50	31.16	29.08	34.64	33.40	33.84	33.00	23.50	33.62
						Large, *f* = 0.80	64.92*	60.86*	68.84*	66.88*	68.90*	66.54*	53.44*	68.66*
	20	20	1	36		Small, *f* = 0.20	10.64	10.34	14.12	14.52	12.28	14.36	8.52	11.16
						Medium, *f* = 0.50	32.88	30.16	41.86	42.20	37.74	41.36	25.84	35.34
						Large, *f* = 0.80	64.68*	61.56*	74.48*	74.38*	70.54*	73.54*	54.16*	68.54*
	15	25	1	1		Small, *f* = 0.20	9.06	8.28	9.90	10.58	9.92	9.88	6.70	9.12
						Medium, *f* = 0.50	29.66	27.82	32.14	32.52	32.12	31.20	21.38	31.52
						Large, *f* = 0.80	63.36*	60.44*	66.50*	66.64*	66.52*	65.04*	50.02*	66.52*
	15	25	1	36	+	Small, *f* = 0.20	12.90	12.32	14.90	18.64	14.88	16.76	9.68	5.46
						Medium, *f* = 0.50	46.56	43.08	51.96*	57.56*	51.98*	53.88*	35.88	29.28
						Large, *f* = 0.80	**84.68***	**80.94***	**87.90***	**90.46***	**87.88***	**88.82***	73.26*	70.26*
	15	25	36	1	-	Small, *f* = 0.20	7.96	7.66	8.58	7.14	8.58	7.14	5.94	16.84
						Medium, *f* = 0.50	19.96	19.44	21.40	18.72	21.40	18.54	15.10	36.60
						Large, *f* = 0.80	43.02	40.58	44.58	40.58	44.56	40.72	31.72	62.86*
					AVERAGE		35.37	33.43	38.79	38.94	38.06	38.03	28.14	37.00

UUM Universiti Utara Malaysia

APPENDIX A

Type of Distribution	Sample Size		Variance		Natural Pairing	Effect Size	15 *WM-H*	25 *WM-H*	15 *WHQ-H*	25 *WHQ-H*	15 WHQ_1-H	25 WHQ_1-H	*MOM-H*	Student's *t*-test
g=0; h=0.5	20	20	1	1		Small, f = 0.20	3.04	4.18	3.58	3.12	2.56	2.92	4.50	4.34
						Medium, f = 0.50	8.14	13.06	7.94	7.30	5.72	7.00	14.34	9.72
						Large, f = 0.80	19.78	31.62	16.60	17.44	12.94	16.38	35.02	19.24
	20	20	1	36		Small, f = 0.20	4.38	4.96	7.18	7.26	4.56	5.78	5.58	5.94
						Medium, f = 0.50	12.24	16.58	16.94	17.82	10.74	14.50	17.76	12.24
						Large, f = 0.80	27.08	36.28	31.56	35.92	22.06	29.74	38.42	24.80
	15	25	1	1		Small, f = 0.20	1.84	3.30	4.34	2.94	2.96	2.36	3.60	4.68
						Medium, f = 0.50	6.20	10.24	10.20	6.88	6.52	6.74	12.72	9.58
						Large, f = 0.80	15.56	26.38	20.28	15.52	13.94	15.92	31.04	18.42
	15	25	1	36	+	Small, f = 0.20	4.20	6.00	8.30	7.90	4.52	5.70	6.54	2.10
						Medium, f = 0.50	14.32	10.24	22.30	22.36	13.18	17.04	25.14	7.12
						Large, f = 0.80	32.84	50.22*	41.96	43.30	27.44	35.54	54.54*	17.40
	15	25	36	1	-	Small, f = 0.20	2.56	3.78	4.18	3.20	3.98	2.68	3.92	11.64
						Medium, f = 0.50	5.82	8.88	7.60	5.00	6.78	5.34	9.82	18.38
						Large, f = 0.80	11.90	19.00	14.40	9.90	13.06	11.16	22.06	28.62
					AVERAGE		11.33	16.31	14.49	13.72	10.06	11.92	19.00	12.95

UUM Universiti Utara Malaysia

APPENDIX A

Type of Distribution	Sample Size		Variance		Natural Pairing	Effect Size	15 *WM-H*	25 *WM-H*	15 *WHQ-H*	25 *WHQ-H*	15 *WHQ_1-H*	25 *WHQ_1-H*	*MOM-H*	Student's *t*-test
g=1; *h*=0	20	20	1	1		Small, *f* = 0.20	4.42	4.82	7.16	4.82	4.84	4.62	6.02	5.50
						Medium, *f* = 0.50	13.92	17.18	17.42	12.78	12.74	12.44	18.16	15.14
						Large, *f* = 0.80	31.78	39.26	35.06	28.28	28.00	27.64	38.84	32.16
	20	20	1	36		Small, *f* = 0.20	12.46	8.14	33.18	21.12	19.56	20.26	5.88	25.48
						Medium, *f* = 0.50	34.48	27.24	66.72*	48.90	45.84	46.86	18.80	57.92*
						Large, *f* = 0.80	64.36*	59.82*	**90.24***	75.76*	70.48*	71.04*	46.44	**85.68***
	15	25	1	1		Small, *f* = 0.20	5.04	4.94	7.66	6.36	6.40	5.02	6.12	4.90
						Medium, *f* = 0.50	14.94	15.28	19.38	15.62	15.80	13.64	17.70	14.50
						Large, *f* = 0.80	31.32	35.50	35.46	30.28	30.74	28.68	37.62	31.84
	15	25	1	36	+	Small, *f* = 0.20	18.80	10.46	37.32	2882	26.66	20.54	6.54	12.10
						Medium, *f* = 0.50	52.32*	42.50	75.90*	61.14*	59.00*	49.68	27.52	46.16
						Large, *f* = 0.80	79.74*	78.72*	**95.04***	**81.22***	**80.56***	73.42*	63.90*	79.74*
	15	25	36	1	-	Small, *f* = 0.20	3.82	5.18	8.40	6.44	8.46	5.12	8.88	15.50
						Medium, *f* = 0.50	7.22	10.64	8.68	7.34	8.46	6.88	18.00	13.28
						Large, *f* = 0.80	13.58	20.42	13.58	12.60	13.26	12.60	30.00	18.24
					AVERAGE		25.88	25.34	36.75	29.43	28.72	26.56	23.36	30.54

Universiti Utara Malaysia

APPENDIX A

Type of Distribution	Sample Size		Variance		Natural Pairing	Effect Size	15 *WM-H*	25 *WM-H*	15 *WHQ-H*	25 *WHQ-H*	15 WHQ_1-*H*	25 WHQ_1-*H*	*MOM-H*	Student's *t*-test
.*g*=1; *h*=0.5	20	20	1	1		Small, *f* = 0.20	1.78	2.46	3.54	2.12	1.68	1.80	3.58	2.60
						Medium, *f* = 0.50	4.48	7.98	6.22	4.54	3.52	3.80	13.30	4.80
						Large, *f* = 0.80	10.40	19.72	11.76	8.74	6.80	7.30	31.78	9.32
	20	20	1	36		Small, *f* = 0.20	3.60	3.54	11.92	7.92	5.70	6.30	3.32	6.72
						Medium, *f* = 0.50	10.96	12.80	23.32	18.04	12.64	13.94	11.88	14.44
						Large, *f* = 0.80	24.12	30.74	39.00	32.88	22.48	24.48	33.06	26.84
	15	25	1	1		Small, *f* = 0.20	1.78	2.30	5.22	2.58	2.62	1.72	3.94	2.78
						Medium, *f* = 0.50	4.66	7.42	9.50	4.56	4.90	3.60	12.78	4.50
						Large, *f* = 0.80	9.84	17.26	16.06	8.82	9.00	7.72	30.26	8.54
	15	25	1	36	+	Small, *f* = 0.20	5.30	4.54	17.34	9.82	7.74	5.34	3.74	2.08
						Medium, *f* = 0.50	15.00	18.56	34.36	20.52	16.66	13.40	17.00	7.10
						Large, *f* = 0.80	30.74	43.94	50.74*	32.78	27.92	24.76	48.38	18.04
	15	25	36	1	-	Small, *f* = 0.20	1.82	3.18	4.84	3.92	4.44	2.32	6.58	9.22
						Medium, *f* = 0.50	3.40	6.40	5.70	3.96	4.60	3.24	15.46	9.88
						Large, *f* = 0.80	6.44	12.82	8.16	5.52	6.60	5.54	27.94	11.92
					AVERAGE		8.95	12.91	16.51	11.11	9.15	8.35	17.53	9.25
			GRAND AVERAGE				20.38	22.00	26.64	23.30	21.50	21.22	22.01	22.43

Notes: (*) more than 50%; (**bold**) more than 80%

UUM
Universiti Utara Malaysia

APPENDIX B

The statistical power rate for $J = 4$

Type of Distribution	Sample Size				Variance				Natural Pairing	Effect Size	15 *WM-H*	25 *WM-H*	15 *WHQ-H*	25 *WHQ-H*	15 *WHQ_1-H*	25 *WHQ_1-H*	*MOM-H*	ANOVA *F*-test
g=0; *h*=0	20	20	20	20	1	1	1	1		Small, *f* = 0.10	8.28	7.28	9.62	9.20	9.66	9.04	4.90	9.78
										Medium, *f* = 0.25	35.48	31.54	40.38	38.42	40.02	38.08	23.82	41.76
										Large, *f* = 0.40	78.30*	73.66*	**82.48***	**80.54***	**82.46***	**80.24***	**62.60***	**83.64***
	20	20	20	20	1	1	1	36		Small, *f* = 0.10	7.42	7.06	9.80	9.96	8.80	9.76	5.62	14.12
										Medium, *f* = 0.25	20.56	19.16	27.16	27.20	24.46	26.72	15.04	37.14
										Large, *f* = 0.40	58.86*	53.50*	69.18*	67.56*	65.78*	66.92*	43.12	**85.64***
	10	15	25	30	1	1	1	1		Small, *f* = 0.10	9.64	8.12	11.10	11.68	11.80	10.56	5.34	9.80
										Medium, *f* = 0.25	40.72	35.06	43.00	44.72	44.58	41.96	24.42	41.38
										Large, *f* = 0.40	**84.12***	79.08*	**85.80***	**86.52***	**86.54***	**85.40***	65.62*	**85.64***
	10	15	25	30	1	1	1	36	+	Small, *f* = 0.10	9.32	8.58	11.90	13.10	11.42	11.52	7.14	5.58
										Medium, *f* = 0.25	43.18	39.34	50.02*	53.48*	50.06*	49.74	31.12	29.16
										Large, *f* = 0.40	**96.38***	**92.54***	**98.00***	**97.80***	**98.22***	**97.40***	**85.88***	**91.38***
	10	15	25	30	36	1	1	1	-	Small, *f* = 0.10	8.04	6.66	9.00	9.00	9.76	8.56	6.10	32.22
										Medium, *f* = 0.25	14.98	12.32	15.16	15.42	16.90	14.92	9.74	52.44*
										Large, *f* = 0.40	31.04	24.76	31.18	31.70	34.00	30.64	20.16	**86.02***
					AVERAGE						36.42	33.24	39.59	39.75	39.63	38.76	27.37	47.05

Universiti Utara Malaysia

PPENDIX B

Type of Distribution	Sample Size				Variance				Natural Pairing	Effect Size	15 *WM-H*	25 *WM-H*	15 *WHQ-H*	25 *WHQ-H*	15 *WHQ$_1$-H*	25 *WHQ$_1$-H*	*MOM-H*	ANOVA *F*-test
										Small, *f* = 0.10	0.82	1.62	1.50	0.84	0.72	0.82	1.62	3.78
	20	20	20	20	1	1	1	1		Medium, *f* = 0.25	3.00	7.74	3.82	3.06	2.00	3.12	8.96	7.68
										Large, *f* = 0.40	10.56	25.44	9.74	9.12	5.72	9.00	29.64	17.26
										Small, *f* = 0.10	2.44	3.28	4.68	4.10	2.86	3.54	3.82	8.30
	20	20	20	20	1	1	1	36		Medium, *f* = 0.25	5.64	8.68	8.96	8.48	5.18	6.96	9.78	13.06
										Large, *f* = 0.40	15.04	23.88	18.90	20.08	11.98	16.98	26.50	24.74
										Small, *f* = 0.10	0.96	1.38	1.98	*g*=0; *h*=0.5	1.44	1.06	1.66	4.58
	10	15	25	30	1	1	1	1		Medium, *f* = 0.25	3.22	5.60	5.86	4.92	3.90	3.76	8.38	8.78
										Large, *f* = 0.40	11.10	19.34	15.42	13.02	10.50	11.08	29.78	18.38
										Small, *f* = 0.10	3.14	3.74	5.76	6.26	3.84	4.20	4.30	2.60
	10	15	25	30	1	1	1	36	+	Medium, *f* = 0.25	4.24	6.24	14.86	15.56	9.70	10.86	17.56	5.76
										Large, *f* = 0.40	27.76	46.04	34.36	35.20	24.46	28.44	56.20*	14.64
										Small, *f* = 0.10	3.04	2.04	4.26	4.28	4.14	4.06	3.20	25.56
	10	15	25	30	36	1	1	1	-	Medium, *f* = 0.25	5.16	3.88	6.02	5.62	5.68	5.38	5.52	31.14
										Large, *f* = 0.40	9.56	8.14	10.16	10.12	9.48	9.44	11.98	40.56
					AVERAGE						7.05	11.14	9.75	9.48	6.77	7.91	14.59	15.12

UUM Universiti Utara Malaysia

APPENDIX B

Type of Distribution	Sample Size				Variance				Natural Pairing	Effect Size	15 *WM-H*	25 *WM-H*	15 *WHQ-H*	25 *WHQ-H*	15 *WHQ$_1$-H*	25 *WHQ$_1$-H*	*MOM-H*	ANOVA *F*-test
g=1; *h*=0	20	20	20	20	1	1	1	1		Small, *f* = 0.10	2.00	2.26	5.08	2.28	2.44	2.26	2.50	5.78
										Medium, *f* = 0.25	7.76	11.32	14.04	8.12	8.34	8.10	11.34	14.86
										Large, *f* = 0.40	24.76	33.84	32.38	21.40	21.70	21.42	33.40	33.14
	20	20	20	20	1	1	1	36		Small, *f* = 0.10	7.64	5.12	24.72	14.02	13.42	13.76	5.24	33.08
										Medium, *f* = 0.25	19.36	15.10	47.46	29.54	28.84	29.28	12.28	58.82*
										Large, *f* = 0.40	48.10	44.82	78.42*	57.48*	56.20*	565.8*	36.76	**86.10***
	10	15	25	30	1	1	1	1		Small, *f* = 0.10	2.54	1.96	4.42	3.40	3.94	2.60	2.44	5.06
										Medium, *f* = 0.25	10.14	10.04	14.96	11.92	13.02	9.34	11.22	13.80
										Large, *f* = 0.40	27.48	31.36	33.98	28.32	29.74	24.90	35.40	34.10
	10	15	25	30	1	1	1	36	+	Small, *f* = 0.10	12.92	7.68	32.84	24.58	25.08	17.82	5.30	20.86
										Medium, *f* = 0.25	39.34	32.44	65.54*	52.90*	52.00*	41.52	22.54	48.70
										Large, *f* = 0.40	75.42*	76.96*	**92.54***	79.12*	78.14*	69.06*	66.42*	**82.52***
	10	15	25	30	36	1	1	1	-	Small, *f* = 0.10	5.40	4.12	9.60	9.56	9.78	8.84	7.96	35.80
										Medium, *f* = 0.25	8.26	7.06	11.24	10.78	11.16	10.12	13.62	40.24
										Large, *f* = 0.40	14.50	13.26	17.18	16.90	17.28	15.96	22.00	52.78*
					AVERAGE						20.37	19.82	32.29	24.69	24.74	22.10	19.23	37.71

Universiti Utara Malaysia

APPENDIX B

Type of Distribution	Sample Size				Variance				Natural Pairing	Effect Size	15 *WM-H*	25 *WM-H*	15 *WHQ-H*	25 *WHQ-H*	15 WHQ_1-*H*	25 WHQ_1-*H*	*MOM-H*	ANOVA *F*-test
g=1; *h*=0.5	20	20	20	20	1	1	1	1		Small, *f* = 0.10	0.32	0.54	1.26	0.48	0.46	0.48	1.38	2.34
										Medium, *f* = 0.25	1.38	2.64	2.60	1.04	0.88	0.96	6.14	3.84
										Large, *f* = 0.40	3036	9.60	5.04	2.32	2.10	2.26	21.92	6.46
	20	20	20	20	1	1	1	36		Small, *f* = 0.10	2.02	2.10	7.98	4.68	3.42	3.88	2.98	10.92
										Medium, *f* = 0.25	4.16	5.18	11.84	7.48	5.34	6.04	6.36	15.32
										Large, *f* = 0.40	9.94	16.60	19.94	13.20	9.62	10.80	22.84	24.18
	10	15	25	30	1	1	1	1		Small, *f* = 0.10	0.42	0.52	1.78	0.96	1.06	0.64	0.92	3.52
										Medium, *f* = 0.25	1.30	2.00	3.84	2.00	2.02	1.32	5.84	4.20
										Large, *f* = 0.40	3.92	7.76	8.14	4.12	4.04	2.62	21.66	6.18
	10	15	25	30	1	1	1	36	+	Small, *f* = 0.10	2.78	2.42	13.16	7.72	5.98	3.56	3.08	3.68
										Medium, *f* = 0.25	6.90	9.76	21.66	12.92	10.40	7.10	11.48	6.86
										Large, *f* = 0.40	17.74	31.04	35.42	21.68	19.16	13.88	48.82	14.48
	10	15	25	30	36	1	1	1	-	Small, *f* = 0.10	2.00	1.58	3.80	3.28	3.42	3.14	4.56	21.32
										Medium, *f* = 0.25	2.94	3.14	4.64	3.38	3.60	3.14	8.68	21.30
										Large, *f* = 0.40	4.92	5.84	7.06	5.08	5.36	4.46	15.58	24.28
					AVERAGE						4.27	6.71	9.88	6.02	5.12	04.29	12.15	11.26
			GRAND AVERAGE								17.03	17.73	22.88	19.99	19.07	18.27	18.34	27.78

Notes: (*) more than 50%; (**bold**) more than 80%

UUM Universiti Utara Malaysia