ALEXANDER-GOVERN TEST USING WINSORIZED MEANS

FARIDZAH BINTI JAMALUDDIN

MASTER OF SCIENCE (STATISTICS)

UNIVERSITI UTARA MALAYSIA

2015

Awang Had Salleh
Graduate School
of Arts And Sciences

Universiti Utara Malaysia

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

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

FARIDZAH JAMALUDDIN 812426

calon untuk Ijazah
(candidate for the degree of) MASTER

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

"ALEXANDER-GOVERN TEST USING WINSORIZED MEANS"

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 : ***23 September 2014.***
That the said thesis/dissertation is acceptable in form and content and displays a satisfactory knowledge of the field of study as demonstrated by the candidate through an oral examination held on: ***September 23, 2014.***

Pengerusi Viva: *(Chairman for VIVA)*	**Assoc. Prof. Dr. Maznah Mat Kasim**	Tandatangan *(Signature)*
Pemeriksa Luar: *(External Examiner)*	**Dr. Nora Muda**	Tandatangan *(Signature)*
Pemeriksa Dalam: *(Internal Examiner)*	**Dr. Nor Aishah Ahad**	Tandatangan *(Signature)*
Nama Penyelia/Penyelia-penyelia: *(Name of Supervisor/Supervisors)*	**Dr. Suhaida Abdullah**	Tandatangan *(Signature)*

Tarikh:
(Date) ***September 23, 2014***

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

Abstrak

Ujian klasik bagi menguji kesamaan kumpulan bebas yang berasaskan min aritmetik boleh menghasilkan keputusan yang tidak sah terutama apabila berurusan dengan data yang tidak normal dan varians heterogen (heteroskedastisiti). Bagi mengurangkan masalah ini, para penyelidik mengusahakan kaedah yang lebih sesuai dengan kondisi yang telah dinyatakan termasuk prosedur yang dikenali sebagai ujian Alexander-Govern. Prosedur ini adalah tidak sensitif terhadap kehadiran heteroskedastisiti di bawah taburan normal. Walau bagaimanapun, ujian yang menggunakan min aritmetik sebagai ukuran kecenderungan memusat adalah sensitif kepada data yang tidak normal. Ini adalah disebabkan oleh hakikat bahawa min aritmetik mudah dipengaruhi oleh bentuk taburan. Dalam kajian ini, min aritmetik digantikan dengan penganggar teguh, iaitu min *Winsor* atau min *Winsor* suai. Ujian Alexander-Govern yang dicadangkan dengan min *Winsor* dan dengan min *Winsor* suai masing-masing ditandakan sebagai *AGW* dan *AGAW*. Bagi tujuan perbandingan, peratusan pe*Winsoran* yang berbeza iaitu 5%, 10%, 15% dan 20% dipertimbangkan. Satu kajian simulasi telah dijalankan untuk mengkaji mengenai prestasi ujian berdasarkan kadar Ralat Jenis I dan kuasa. Empat pembolehubah; bentuk taburan, saiz sampel, tahap keheterogenan varians dan sifat pasangan dimanipulasi untuk mewujudkan keadaan yang boleh menyerlahkan kekuatan dan kelemahan setiap ujian. Prestasi ujian yang dicadangkan ini dibandingkan dengan kaedah parametrik lain yang setaraf iaitu, ujian-*t* dan *ANOVA*. Ujian yang dicadangkan menunjukkan peningkatan dari segi kawalan Ralat Jenis I dan kuasa yang semakin tinggi di bawah pengaruh heteroskedastisiti dan ketidaknormalan. Ujian *AGAW* menunjukkan prestasi terbaik dengan 10% pe*Winsoran* manakala ujian *AGW* menunjukkan prestasi terbaik dengan 5% pe*Winsoran*. Di bawah kebanyakan keadaan (74%), ujian *AGAW* mengatasi ujian *AGW*. Oleh yang demikian, min *Winsor* dan min *Winsor* suai berupaya meningkatkan prestasi asal ujian Alexander-Govern dengan berkesan. Prosedur yang dicadangkan ini memberi manfaat kepada pengamal statistik dalam menguji kesamaan kumpulan bebas walaupun di bawah pengaruh ketidaknormalan dan varians heterogen.

Kata Kunci: Min *Winsor*, Min *Winsor* suai, Ketaknormalan, Varians tak homogen

Abstract

Classical tests for testing the equality of independent groups which are based on arithmetic mean can produce invalid results especially when dealing with non-normal data and heterogeneous variances (heteroscedasticity). In alleviating the problem, researchers are working on methods that are more adapt to the aforementioned conditions which include a procedure known as Alexander-Govern test. This procedure is insensitive in the presence of heteroscedasticity under normal distribution. However, the test which employs the arithmetic mean as the central tendency measure is sensitive to non-normal data. This is due to the fact that the arithmetic mean is easily influenced by the shape of distribution. In this study, the arithmetic mean is replaced by robust estimators, namely the Winsorized mean or adaptive Winsorized mean. The proposed Alexander-Govern test with Winsorized mean and with adaptive Winsorized mean are denoted as *AGW* and *AGAW*, respectively. For the purpose of comparison, different Winsorization percentages of 5%, 10%, 15% and 20% are considered. A simulation study was conducted to investigate on the performance of the tests which is based on rate of Type I error and power. Four variables; shape of distribution, sample size, level of variance heterogeneity and nature of pairings are manipulated to create the conditions which could highlight the strengths and weaknesses of each test. The performance of the proposed tests is compared with their parametric counterparts, the *t*-test and *ANOVA.* The proposed tests show improvement in terms of controlling Type I Error and increasing power under the influence of heteroscedasticity and non-normality. The *AGAW* test performed best with 10% Winsorization while *AGW* test performed best with 5% Winsorization. Under most conditions (74%), *AGAW* tests outperform *AGW* tests. Therefore, the Winsorized mean and the adaptive Winsorized mean can significantly improve the performance of the original Alexander-Govern test. These proposed procedures are beneficial to statistical practitioners in testing the equality of independent groups even under the influence of non-normality and variance heterogeneity.

Keywords: Winsorized mean, Adaptive Winsorized mean, Non-normality, Heteroscedasticity

Acknowledgement

In the name of Allah, the Most Gracious and the Most Merciful. Thank you to Allah S.W.T for the gift of life and blessing that has enabled me to complete this research.

I would like to express my appreciation and acknowledgement to Dr. Suhaida Abdullah for her invaluable guidance, assistance and hard work in helping me throughout this research. Without her careful supervision and expertise, the completion of this research would not have been possible.

Special thanks to Assoc. Prof. Dr. Sharipah Soaad Syed Yahaya and Dr. Nor Aishah Ahad for their fruitful opinions and feedback to make this research a better piece of work during its initial stage.

Also special thanks to my father, Jamaluddin bin Hamid, my mother, Rozina binti Aziz and my brothers and sisters. With their love, patience, motivation, help and also their understanding, I have the emotional strength to complete this research.

Last but not least, I would also like to thank Universiti Utara Malaysia (UUM) for sponsoring my studies in Universiti Utara Malaysia.

Table of Contents

List of Tables

List of Figures

List of Appendices

List of Abbreviations

ANOVA	Analysis of variance
***A*-test**	Alexander-Govern test
AGW	Alexander-Govern test with Winsorized mean
AGW_5	Alexander-Govern test with 5% Winsorized mean
AGW_10	Alexander-Govern test with 10% Winsorized mean
AGW_15	Alexander-Govern test with 15% Winsorized mean
AGW_20	Alexander-Govern test with 20% Winsorized mean
AGAW	Alexander-Govern test with adaptive Winsorized mean
AGAW_5	Alexander-Govern test with 5% adaptive Winsorized mean
AGAW_10	Alexander-Govern test with 10% adaptive Winsorized mean
AGAW_15	Alexander-Govern test with 15% adaptive Winsorized mean
AGAW_20	Alexander-Govern test with 20% adaptive Winsorized mean

CHAPTER ONE

INTRODUCTION

Classical parametric tests, such as *t*-test and analysis of variance (*ANOVA*) *F* test are widely used by researchers in many disciplines. These tests are useful in comparing the equality of two or more treatment groups. A review conducted by Farcomeni and Ventura (2010) found that most of the studies in health sciences such as medicine and genetics, used classical test in comparing treatment groups. In addition, Erceg-Hurn and Mirosevich (2008) also mentioned the extensive usage of classical test in psychology studies.

The classical parametric tests are based on assumptions of normality and homoscedasticity. However, in dealing with real data, these assumptions are rarely met. For example, Micceri (1989) found that the majority of real data from the psychological and education literatures are skewed and heavy-tailed. Studies by Wilcox (1990) also found that most real data are often non-normal with the tendency to be either non-smooth, multi-modal, highly skewed or heavy-tailed. Besides that, comprehensive journal review conducted by Keselman et al. (1998) demonstrated that it is very hard to find homogeneous variances when dealing with education data as well as with data of child, clinical and experimental psychology. Another study by Erceg-Hurn and Marosevich (2008) claimed that it is usual for the homogeneous variances assumption to be violated when dealing with real data. The classical tests have been shown to have lack of robustness under the violation of the assumptions of normality and

homoscedasticity (Glass, Peckham & Sanders, 1972; Wilcox, 2002). Furthermore, the violation of both or either assumptions can reduce the power of test and can fail to control the Type I error rates (Keselman, Wilcox, Othman & Fradette, 2002; Wilcox, 2002).

There are numerous alternative methods that have been introduced to overcome the problems faced by the classical parametric tests such as nonparametric approach. For example, the Wilcoxon test serves as an alternative for *t*-test and the Kruskal-Wallis test for *ANOVA*. According to Zimmerman and Zumbo (1993), the Wilcoxon test is more powerful than *t*-test for several non-normal distributions but it exhibits a substantial power loss when faced with heterogeneous variances. The disadvantage of Kruskal-Wallis test is that it is very sensitive to the presence of heterogeneous variances (Oshima & Algina, 1992). Some researchers have developed approximate parametric tests as alternatives to the classical tests without the assumption of homogeneous variances. James test (James, 1951) and Welch test (Welch, 1951) are the parametric alternative tests under the violation of this assumption. Their ability to control Type I error rates and produce reasonable power under heterogeneous variances are the advantages of these tests. However, the results of these alternative tests could be seriously affected by non-normal data situation (Myers, 1998; Reed, 2005; Schneider & Penfield, 1997). The violation of normality assumption increased the probability of committing Type I error (Wilcox, 2002, Wilcox, 2005).

Alexander and Govern (1994) have proposed another approach in comparing treatment groups in dealing with heterogeneous variances called the Alexander-Govern test (*A*-

test). This method is identified as a good alternative to *ANOVA* compared to other alternative methods such are James test and Welch test (Fan & Hancock, 2012; Schneider & Penfield, 1997). This *A*-test is superior to the James test in terms of the simplicity of the computation of the test statistic (Myers, 1998; Schneider & Penfield, 1997). The test also has relative superiority in terms of both statistical power and Type I error rates under most of the experimental conditions considered in a study whose aim is to provide an alternative to *ANOVA* (Schneider & Penfield, 1997).

1.1 Problem Statement

The *A*-test has been recommended as a relatively good alternative method to the *ANOVA* under heterogeneous variances. No one can argue the robustness of the *A*-test under heterogeneous variances; however, this test has a limitation due to its sensitivity to the non-normal data (Myers, 1998; Schneider & Penfield, 1997). In addition, it produces the Type I error rates more toward the conservative value, 0.025 when distribution is heavy-tailed (Wilcox, 1997).

Studies have been done on the *A*-test focusing on how to improve this test so that it is robust under non-normal distribution. One research involves the modification of *A*-test with bootstrap procedure (Wilcox, 1997). However, the modification failed to improve the performance of the *A*-test under heavy-tailed distribution. Moreover, the modified test produces more conservative Type I error rates compared to the original test. Other studies adopted robust central tendency measure such as trimmed mean to the *A*-test (Lix & Keselman, 1998; Luh & Guo, 2005). The replacement of the arithmetic mean with trimmed mean is able to improve the *A*-test under skewed distribution but it failed to

control Type I error rates for extremely skewed distribution (Lix & Keselman, 1998). Since this modification has some limitations under certain experimental conditions, Luh and Guo (2005) proposed another approach by adopting the trimmed mean in conjunction with Hall's transformation into the *A*-test. In comparison to the *A*-test, the proposed approach only slightly improved the control of Type I error rates. Nevertheless, this improvement is considered too small. A further disadvantage is that the modification also renders the *A*-test to be no longer a simple method. Furthermore, this combination technique produces slightly lower power when dealing with normal data as compared with the original *A*-test.

Two modifications of the *A*-test by Abdullah (2011) adopted more flexible robust central tendency measures into the test which are the adaptive trimmed mean and the modified one-step-*M*-estimator (*MOM*). Both modified *A*-tests provide good control of Type I error rates under skewed distribution. However, both methods produce low power under heavy-tailed distribution. Besides that, the modification of *A*-test with modified one-step-*M*-estimator (*MOM*) failed to control Type I error rates when dealing with heavy-tailed distribution. Therefore, this modification still has limitation under heavy-tailed distribution.

Wilcox (2002) noted that the outliers tend to exist under heavy-tailed distribution. Thus, the use of arithmetic mean as a central tendency measure in *A*-test is seriously influenced by the existence of outliers. There are two common approaches that can be used when dealing with outliers which are trimming and Winsorization (Lusk, Halperin & Heiling, 2011; Moir, 1998). The existing studies on modifying *A*-test are using

trimming approach in dealing with outliers (Lix & Keselman, 1998; Luh & Guo, 2005; Abdullah 2011). Unfortunately, trimming is not always a feasible approach when dealing with outliers. According to Farrell-Singleton (2010), the outliers should not be trimmed or removed simply because they are extreme relative to the remaining of data values. Orr, Sackett and DuBois (1991) also agreed that no value should be removed simply because its magnitude is extreme in comparison to the other data values, instead it should be allowed to contribute to the results. Hawkins (1980) suggested using Winsorization instead of trimming when dealing with heavy-tailed distribution as mentioned in Lance (2011):

> If the observations are generated by the heavy-tailed distribution, and one wishes to estimate the parameters of this distribution, then the outliers represent valid observations. Thus one should be reluctant to discard them entirely, and hence prefer to use Winsorization, which is robust, but does make partial use of the outliers. (p. 5)

Farrell-Singleton (2010) has also proposed the Winsorization approach as a solution for treatment of data when outliers exist. In addition, the approach is also recommended as the preferred ways of treating outliers (Dhiren & Andrew, 2012; Thomas & Ward, 2006). The study conducted by Etzel et al. (2003) reveals that this approach works best when the distributions are leptokurtic, that is, distributions that are more peaked or taller than the normal distribution. The robust mean estimator based on Winsorization is recommended when dealing with heavy-tailed distribution (Fung & Rahman, 1980) and skewed distribution (Rivest, 1994). Therefore, instead of using the trimming approach, the Winsorization is a better alternative since it can deal with skewed and heavy-tailed situations as well as with the existence of outliers.

1.2 Objectives of the Study

The goal of this study is to modify the Alexander-Govern test to be robust under departure from normal distribution and homogeneous variances. The modified Alexander-Govern test should be able to control the Type I error rates and improve the power of a test. In achieving this goal, the following objectives need to be accomplished:

1. To modify the *A*-test using a central tendency measure based on Winsorization:
    i. Winsorized Mean
    ii. Adaptive Winsorized Mean
2. To compare the performance of modified *A*-test with *A*-test and classical tests in terms of Type I error rates and the power of a test.
3. To determine the best method under the violation of normality and homogeneous variances assumptions.
4. To evaluate the performance of the modified tests using real data.

1.3 Significance of the Study

This study is beneficial for statistical practitioners and also for applied researchers. The proposed central tendency measures and the robust modified tests would contribute to the area of statistics as well as for the application of statistical method. For statistical practitioners, the study serves as their reference or guide in analyzing the comparison of treatment groups. For applied researchers, the study provides them with a robust alternative method in comparing treatment groups under the violation of normality and homogeneous variances assumptions.

1.4 Organization of the Thesis

In Chapter 1, we have briefly introduced the alternative tests to the classical tests under the violation of normality and homoscedasticity assumptions. We also discussed the limitations of the modification of *A*-test done by previous researchers for the purpose of improving the *A*-test under non-normal distribution leading to the objectives of this research. The reviews of *A*-test and the proposed robust central tendency measure based on Winsorization are reported in Chapter 2. The definition of the Type I error and power of a test are also explained in the chapter. Chapter 3 describes how the empirical investigations are conducted. The discussion in this chapter includes the selection of the conditions being investigated for five different manipulated variables: number of groups, group sizes, group variances, nature of pairing and types of distributions, followed by data generation, the setting of central tendency measure for power analysis and we end the Chapter 3 with the procedure of proposed modified *A*-test using central tendency measures based on Winsorized mean and adaptive Winsorized mean. Chapter 4 discusses the analysis of results of Type I error rates, overall results and power of a test as well the performance of the proposed modified test on protoporphyrin dataset. Finally, the summary on the performance of the test, implication and limitation of the study as well as suggestion for future research are presented in Chapter 5.

CHAPTER TWO

LITERATURE REVIEW

This chapter provides reviews on the testing of the equality of means of independent groups, the classical tests and the Alexander-Govern test. The central tendency measures, trimming approach, Winsorization approach, Type I error rates and power of a test are also discussed in this chapter.

2.1 Testing the Equality of Means of Independent Groups

The classical tests that are commonly used for the purpose of testing the equality of means of independent groups are *t*-test and *ANOVA*. These tests are widely used in various fields. For example, Choi and Zhao (2014) use *t*-test to study whether there are any differences in degree of concerns about the ingredients in the food between the groups who cared about health issues and those who cared little about health issues when they eat out at restaurants.

Another study conducted by Ulusoy (2008) in manufacturing, used *ANOVA* to determine whether the image parameters of the three different mill products (ball, rod and autogenous) of talc mineral are statistically different from each other or not. The controlling and manipulated of particle shapes is essential since this particles shape affects the physical characteristics and behavior of talc mineral. The study by Bukat et al. (2008) on the influence of a particular element (Bi and Sb) added to Sn-Ag-Cu and Sn-Zn alloys on their surface and interfacial tension and wetting properties also uses

ANOVA. The additional of these two particular Bi and Sb are expected to decrease of the surface and interfacial tension and in turn improve the wettability of alloys.

Furthermore, Murari and Tater (2014) use *ANOVA* to measure the attitude of employees toward the adoption of information-technology-based (IT-based) banking service among four different Indian private banks namely the ICICI Bank, HDFC Bank, AXIS Bank and INDUSINS Bank. They have tested on the attitude of employees towards the IT-based banking services on the basis of relative advantage, complexities, potential risk, decision making process and innovation techniques used by the four different private banks employees in providing services to the customer.

These classical tests are powerful and valid for identifying treatment effects when certain assumptions are fulfilled. One of the important assumptions that underlie the classical tests is that the distribution of each population is normal. However, the normality assumption is difficult to achieve when dealing with real data. For example, analysis of real data of monthly index return stock market of Middle East and North African (MENA) region from a research done by Yu and Kabir Hassan (2010) are depicted in Figure 2.1.

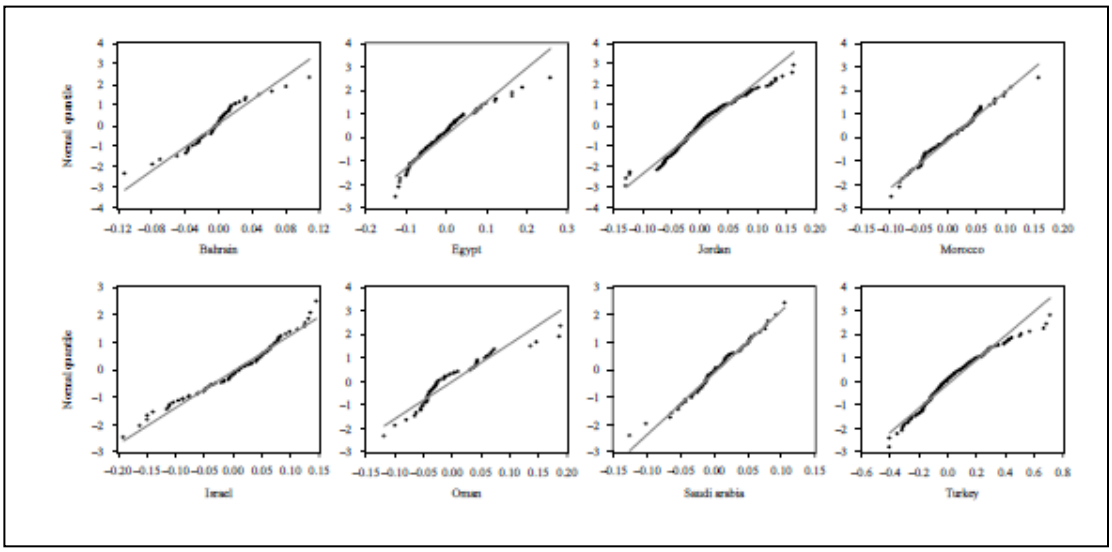

Figure 2.1. Normal QQ Plot of Middle East and North African (MENA) Stock Markets

It shows the normal QQ plot for stock market of MENA region and it can be observed that the monthly index returns for the MENA stock markets are far from being normally distributed for all countries except Morocco, Israel and Saudi Arabia.

Another important assumption underlying the classical tests is that the variances are homogeneous. In spite of that, it is well known that heterogeneous variances instead are more common in real data such as behavioral science data (Erceg-Hurn & Mirosevich, 2008) and clinical data (Grissom, 2000). This assumption can be explained in terms of variance ratio (VR). The VR is a ratio of the largest variance to the smallest variance. The ratios of VR greater than ten occur frequently (Grissom, 2000). Erceg-Hurn and Mirosevich (2008) have identified 28 studies from two recent issues of *Journal of Experimental Psychology: General* and *Journal of Experimental Psychology: Human Perception and Performance* in which data were analyzed using *ANOVA* with the presence of heterogeneous variances. The sample VRs between 2:1 and 4:1 were the common values in these studies. Not only that, the ratios values of 39:1, 59:1, 69:1 and

121:1 were also found. The presence of heterogeneous variances can have drastic effects on the reliability and validity of the test, especially when the group sample sizes are also unequal (Glass et al., 1972; Zimmerman, 2004). Moreover, such presence can jeopardize the validity of the result, by increasing the Type I error rates that can lead to invalid inferences (Nordstokke, Zumbo, Cairns & Saklofske, 2011). The test remains affected in the case of equal sample sizes with heterogeneous variances (Alexander & Govern, 1994)

The classical tests are valid and powerful tests. However, when the underlying assumptions are violated, their results are typically unreliable and invalid. The existence of non-normality and heterogeneous variances situation can produce unsatisfactory results in terms of both the Type I error rates and power (Wilcox, 1997).

Therefore, when the underlying assumptions of the classical test are violated, alternative methods such as nonparametric test, transforming data and heteroscedasticity parametric tests should be considered. Non parametric tests are well-known alternative procedures under non-normal data. These tests do not require parametric assumption since the interval/ratio data are transformed to rank-order data. According to Syed Yahaya (2005), this procedure possesses less power when compared with parametric test and requires large sample size to reject a false hypothesis. Furthermore, it is sensitive to the presence of heterogeneous variances either in balanced or unbalanced sample sizes (Lix, Keselman & Keselman, 1996).

Data could also be transformed to achieve normality and also to reduce heteroscedasticity. With regard to transformation of data, the main problem is that the interpretation of results of transformed data can be problematic and unclear (Grissom, 2000; Keselman, Wilcox, Lix, Algina & Fradette, 2007). In addition, means of the original data tend to be underestimated when both the assumptions are violated (Grissom 2000). Furthermore, data transformation may reduce the power when the distribution is heavy-tailed and outliers exist. Besides, it is hard to find a suitable transformation that will simultaneously deal with non-normality and heteroscedasticity (Wilcox & Keselman, 2003).

Other alternative method that researchers might consider is to use the heteroscedasticity parametric test. These approximate parametric tests are insensitive to heterogeneous variances. Alexander-Govern test, James test and Welch test are among the well-known heteroscedasticity parametric tests. They are able to control the Type I error rates and produce high power when group variances are heterogeneous and distributions are normal (Lix & Keselman, 1998; Schneider & Penfield, 1997). However, the literatures also point out that these tests fail to control the Type I error rates and are also less powerful when the data are both heterogeneous and non-normal especially when the sample sizes is unbalanced (Myers, 1998; Schneider & Penfield, 1997). Thus, these approximate parametric tests have limitations with regard to their sensitivity to the nature of the population distribution. The James test appears to be generally most accurate method under various conditions (Alexander & Govern, 1994; Myers, 1998). However, the major drawback of this test is the complexity of the computation (Myers, 1998). Previous researchers recommended using Alexander-Govern test instead of

James test under heterogeneous variances and normal data (Myers, 1998; Schneider & Penfield, 1997). The recommendation is due to the advantages of Alexander-Govern test in terms of controlling the Type I error rates and power (Schneider & Penfield, 1997). Besides that, another advantage is the simplicity of the computation of Alexander-Govern test statistic (Alexander & Govern, 1994; Myers, 1998; Schneider & Penfield, 1997; Wilcox, 1997). However, the Type I error rates become more conservative as the tails of distribution get heavier either with homogeneous or heterogeneous variances (Wilcox, 1997). The Pareto distribution, log-normal distribution, Weibull distribution and Cauchy distribution are the example of heavy-tailed distribution.

2.2 Classical Parametric test

The *t*-test and *ANOVA* are the most widely used statistical test in testing the equality of means. The *t*-test is developed by William Sealy Gosset in year 1908 and it is useful in testing the equality of two means. While, the *ANOVA* is useful in comparing three or more means and it was developed by Ronald A. Fisher in year 1930.

The classical tests are the most efficient tests and produce accurate result when both the assumption of normality and homogenous variances are satisfied (Erceg-Hurn & Mirosevich, 2008; Farcomeni & Ventura, 2010). However, the violation of normality and homoscedasticity assumptions contribute to the lack of robustness of the tests (Wilcox, 2002).

Therefore, the normality and homogenous variances assumptions need to be verified first before conducting any statistical test on the equality of independent groups.

Examining normality is done through graphical methods, numerical methods and formal testing.

One graphical method through which the researchers can gain an adequate perspective of the variable is using a histogram. This is the simplest diagnostic test for normality by comparing the observed data values to a distribution approximating the normal distribution. Although the histogram is the simplest diagnostic test for normality, this graphical method is problematic for smaller samples (Hair, Black, Babin, Anderson & Tatham, 2006).

Another reliable graphical method for assessing normality is the normal probability plot (Hair et al., 2006). This approach compares the cumulative distribution of actual data values with cumulative distribution of a normal distribution. The plotted data values are compared to the diagonal line which is the straight diagonal line representing the normal distribution. The distribution is assumed normal if the line representing the actual data distribution closely follows the diagonal line as shows in Figure 2.2.

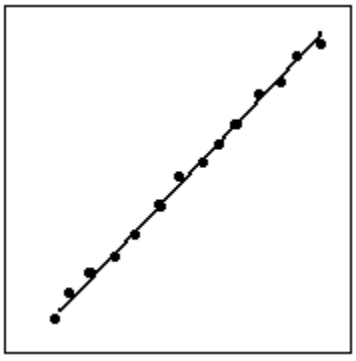

Figure 2.2. Normal Probability Plot of Normal Distribution

By using the normal probability plot, the distribution that departs from normal can be easily seen in terms of kurtosis and skewness. When the line falls below the diagonal, the distribution is flatter than expected. When it goes above the diagonal, the distribution is more peaked than normal curve. For example, in the normal probability plot of a peaked distribution as demonstrate in Figure 2.3.

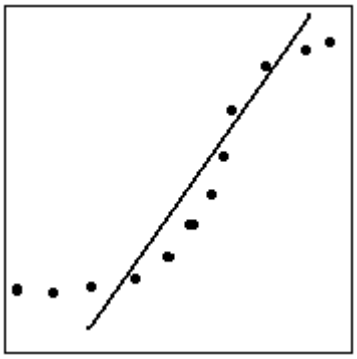

Figure 2.3. Normal Probability Plot of a Peaked Distribution

Initially the distribution is flatter and the plotted line falls below the diagonal. Then the peaked part of the distribution rapidly moves the plotted line above the diagonal, and eventually the line shifts to below the diagonal again as the distribution flattens (Hair et al., 2006).

While the skewed is most often represents as by a simple curve line, either below or above the diagonal. A negative skewed is indicated by curve line below the diagonal line as depicted in Figure 2.4.

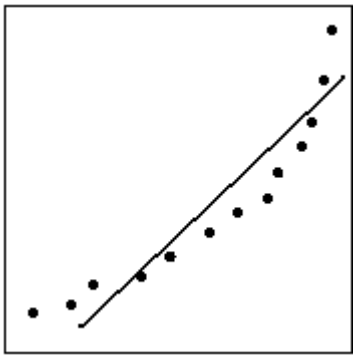

Figure 2.4. Normal Probability Plot of a Negative Skewed Distribution

Whereas, Figure 2.5 represent a positive skewness distribution where the curve line above the diagonal line.

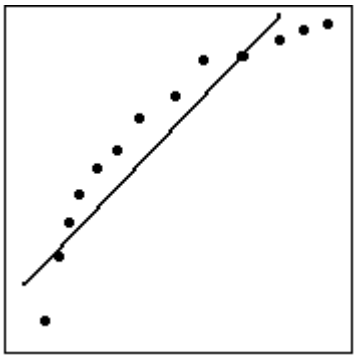

Figure 2.5. Normal Probability Plot of a Positive Skewed Distribution

Even though the normal probability plot can serve as a useful tool in checking normality, it is still not sufficient to provide conclusive evidence that the normal assumption holds. The more formal methods which are numerical methods such as skewness and kurtosis and formal testing test such as Shapiro-Wilks should be performed before making any conclusion about the normality of the data as an alternative to support the normal probability plot. The distribution is assumed normal when the kurtosis and skewness

values are zero. Therefore, the values above and below zero reflects departure from normal distribution.

A negative skewness indicates a distribution that is shifted to the right, whereas the positive skewness denotes a distribution shift to the left. While the positive kurtosis values reflect that the distributions are leptokurtic (peaked), the platykurtic (flatter) distributions are represented by the negative kurtosis values.

Later, using the values of skewness and kurtosis, the statistic value (*z* score) for both skewness and kurtosis can be calculated. If *z* score exceeds the critical value (± 1.96) which corresponds to the 0.05 significance level, then the distribution is consider non-normal with respect to that significance level. A Shapiro-Wilks is the preferred test because of its good power properties (Mendes & Pala, 2003). Moreover, this test is the most powerful normality test in comparison with Anderson-Darling test, Lilliefors test and Kolmogorov-Smirnov test (Razali & Wah, 2011). In this test, the distribution is considered non-normal when the *p*-value is less than the specified significance level.

In order to verify the second assumption, the homogeneity of variances, Levene's test is the most widely used test for testing the equality of variances. This test is widely used because of the availability of this test in most statistical software packages such as MINITAB and SPSS (Keyes & Levy, 1997). However, the violation of symmetry increases the Type I error rates of the Levene's test (Shoemaker, 2003; Zimmerman, 2004). Nordstokke et al. (2011) in their studies has found that the nonparametric Levene's test performs well in terms of maintenance of its nominal Type I error rates

and power when data are sampled from non-normal distribution. As described by Nordstokke and Zumbo (2010), the nonparametric Levene's test involves three steps:

i. pooling the data from each groups and ranking the score

ii. replacing the original values with rank values

iii. applying the Levene's test on the ranks

This test can be defined as in equation 2.1.

$$ANOVA(\left|R_{ij} - \overline{X_j}\right|) \tag{2.1}$$

where R_{ij} are the pooled values from each of the j^{th} groups, and $\overline{X_j}$ is the mean of the ranks for each group.

2.3 The Alexander-Govern Test

Ralph A. Alexander and Diane M. Govern have proposed an approximation test in testing the equality of independent groups under heterogeneous variances and it is known as Alexander-Govern test (*A*-test). This test is based on normalizing transformation (Hill transformation) of one-sample *t* statistic. It is a test proposed for testing the equality of *J* independent mean with null hypothesis of:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_J$$

where $\mu_1=\mu_2=\ldots=\mu_J$ are the mean of J independent groups. Every J groups with size n_j, has sample mean ($\bar{X}_j$) and each of the means will have a standard error (S_j). The S_j is derived as:

$$S_j = S_{\bar{X}_j} = \sqrt{\frac{\sum_{i=1}^{n_j}(X_i - \overline{X}_j)^2}{n_j(n_j-1)}} \tag{2.2}$$

Using the S_j, a weight (w_j) is calculated as:

$$w_j = \frac{1/S_j^2}{\sum_{j=1}^{J} 1/S_j^2} \tag{2.3}$$

such that $\sum w_j = 1$

Then the weighted mean (X^+) is computed as:

$$X^+ = \sum_{j=1}^{J} w_j \overline{X}_j \tag{2.4}$$

One-sample t statistic t_j is then calculated using weighted mean as follows:

$$t_j = \frac{\overline{X}_j - X^+}{S_j} \tag{2.5}$$

where each of the t_j will be distributed as t distribution with $v_j = n_j - 1$ degrees of freedom.

The procedure proceeds to obtain the z statistic from normalizing transformation (Hill transformation) for each of the t statistic value:

$$z_j = c + \frac{(c^3 + 3c)}{b} + \frac{(4c^7 + 33c^5 + 240c^3 + 855c)}{(10b^2 + 8bc^4 + 1000b)} \quad (2.6)$$

where c = $[a \ln(1 + t_j^2/v_j)]^{1/2}$; b = $48a^2$ and a = $v_j - 0.5$.

Finally the *A*- test statistic is obtained by the summation of the z_j^2 values:

$$A = \sum_{j=1}^{J} z_j^2 \quad (2.7)$$

where *A*-test is approximately distributed as $\mathcal{X}^2$ distribution with (*J*-1) degrees of freedom. The null hypothesis is rejected when the value of *A*-test is larger than $\mathcal{X}_{J-1}^2$.

2.4 Central Tendency Measures

Arithmetic mean is a well-known central tendency measure; however, it is very sensitive to the shape of data distribution. The performance of this estimator is only good under normal distribution but not for the case of non-normal distribution. As a result, any statistical method based on this estimator is unable to control Type I error rates and would produce low power when it comes to non-normal data (Keselman et al., 2007).

2.5 Robust Central Tendency Measure

The use of robust central tendency measure is encouraged in order to minimize the effect of non-normality (Wilcox, 1997). A robust central tendency measure is an estimator that is insensitive to small deviations from the assumptions. Thus, for over than 30 years, researchers have been considering different ways of estimating a value which represents the bulk of the observations (Keselman et al., 2007). Trimmed mean, adaptive trimmed mean and Winsorized mean are examples of robust central tendency measures. These

robust central tendency measures can be said as better than the commonly used mean, having the bounded influence function and higher breakdown point. In contrast, arithmetic mean has unbounded influence function and its breakdown point is 0 (Wilcox, 2005).

Influence function allows us to assess the relative influence of individual observations towards the value of an estimate (Huber, 2004). It can be either bounded or unbounded influence function. Estimator with bounded influence function is less affected by the changes in dataset. In contrast, estimators with unbounded influence function will greatly be affected by any slight occurrence in dataset and this might cause misleading of results interpretation. The breakdown point describes quantitatively how greatly small changes in the underlying distribution would change the distribution of an estimator (Huber, 2004). It can be defined as the minimum number of observations for which a functional goes to infinity (Wilcox, 2005) and unable to represent a true value. A good estimator should have bounded influence function and a high breakdown point (Huber, 2004).

2.5.1 Trimmed Mean and Adaptive trimmed mean

The trimmed mean and adaptive trimmed mean have a bounded influence function and their breakdown points are based on the percentage of trimming. Trimmed mean is a mean obtained after trim the data symmetrically on both tails of the distribution. This estimator is widely used by researchers in order to counter the effects of non-normality (Keselman, Wilcox, Algina, Fradette, & Othman, 2004). A research by Ozdemir, Wilcox and Yildiztepe (2013) compares two independent groups such as Yuen's test with

trimmed mean and percentile bootstrap method with trimmed mean as an alternative under the violation of normality and homogeneity of variances assumptions.

Although the previous studies found the advantages of using trimmed mean in improving the test under skewed distribution, it suffer from a practical concerns which is when the distribution is highly skewed to the right, it seems practical to trim more observations from the right tail instead of symmetrically trim on both sides of distribution (Wilcox, 2002). To overcome the problem faced by trimmed mean, Hogg (1974) has proposed an adaptive trimmed mean. This estimator utilizes the characteristic of the data to determine whether data should be trimmed symmetrically or asymmetrically. Asymmetrically trim means that the data is trim by different percentage of trimming on each tail of distribution.

2.5.2 Winsorized Mean

The Winsorized mean has a bounded influence function and its breakdown point is equal to the percentage of Winsorization. This mean is obtains by symmetrically winsorized the tail of distribution. The percentage of Winsorization (α %) is fixed in advance and mean is calculated based on the winsorized sample.

Let $x_1 < x_2 < \ldots < x_{n-1} < x_n$ represent the ordered observations in a sample. Let m be the number of observations to be winsorized and define $m = [\alpha n]$, where α represents the Winsorization percentage. Yuen (1971) defined α-Winsorized mean for sample as:

$$\bar{x}_w(\alpha) = \frac{(m+1)x_{m+1} + x_{m+2} + \ldots + x_{n-m-1} + (m+1)x_{n-m}}{n} \tag{2.8}$$

The sample Winsorized variance is given by

$$s_w^2(\alpha) = (n-1)^{-1}\{(m+1)[x_{m+1} - \bar{x}_w(\alpha)]^2 + [x_{m+2} - \bar{x}_w(\alpha)]^2 + \ldots + [x_{n-m-1} - \bar{x}_w(\alpha)]^2 + (m+1)[x_{n-m} - \bar{x}_w(\alpha)]^2\} \tag{2.9}$$

And the standard error of the Winsorized mean is

$$S_{w(\alpha)} = \sqrt{\frac{s_w^2(\alpha)}{n}} \tag{2.10}$$

2.6 Trimming Approach

Trimming is originally an approach for reducing the effects of the outliers in a sample (Wilcox, 2005). This approach removes the smallest and the largest observations in a given dataset. Trimming are done using symmetric or asymmetric trimming approach. For symmetric trimming, data are trimmed with equal percentage in both tails of distribution. However, when the data are asymmetric, trimming symmetrically is no longer appropriate. In such case, the data need to be trimmed in different amount of percentage on each tail. Unlike symmetric trimming, the percentage of data to be trimmed in asymmetric trimming need to be determined for each tail before the trimming approach is performed which is based on the characteristic of the data.

2.7 Winsorization Approach

Winsorization is an approach that gives less weight to values in the tails of distribution and pays more attention to those near the center (Wilcox, 2005). Basically this approach

replaces outlier in each tail of distribution with the closest values. This approach has been used for many years and still has important application in some areas.

For example, Dixon and Yuen (1974), indicated that the estimator based on Winsorization namely the Winsorized mean is more efficient than the trimmed mean for distributions with shapes close to Gaussian; however, the trimmed mean is more efficient for distributions that are far different from Gaussian. Furthermore, Fuller (1991) continues exploring the effects of the Winsorized mean in Weibull distribution which represents the right skewed distribution and discovers that this estimator is more efficient than the mean under this distribution. Recently, Mirtagioglu, Yigit, Mollaogullari, Genc and Mendes (2014) study on the influence of using Winsorized mean on Type I error rates in the comparison of the independent groups when the assumptions of *ANOVA* are violated.

Other researchers modified the regression analysis with Winsorization approach. Yale and Forsythe (1976) introduced and discussed on the winsorized regression as an alternative to linear regression models. The new approach proposed by Chen and Dixon (1972) in stratified and pooled procedure for linear regression analysis with Winsorization showed increased efficiency over the standard method. In a two-variable regression model setup, the Lien and Balakrishnan (2005) examined the effect of Winsorization on the regression estimates and the efficiency of the regression estimation.

Besides that, Fried (2004) has developed and testing a robust regression functional for local approximation of the trend in a moving time window and further investigate the outlier replacement by Winsorization based on robust scale estimator. The outlier replacement used for the purpose of improving the robustness of the procedure investigated. Scholze et al. (2001) improved the generalized least-squares method with Winsorization for the purpose of protecting the estimates against outliers. This modified approach was used for estimation of effect concentrations for continuous toxicity data.

A study on the comparison of the performance between the usual bootstrap and Winsorized bootstrap has argued on the performance of some resample (usual bootstrap) that may have a higher contamination level than the initial samples (Amado & Pires, 2004). In addition, Srivastava, Pan, Sarkar and Mudholkar (2009) have developed two variations of co-ordinatewise Winsorized-bootstrapped approach. These new methods are seen to provide significant improvement when the data are in the neighborhood of multivariate normal population without significant loss in performance.

The modification of Hotelling's T^2 with estimators of location and scale based on Winsorizing approach is proposed by Lix, Keselman and Hinds (2005). This modified test serves as an alternative test in testing the equality of means in two-group multivariate design under the violation of covariance homogeneity and normality assumptions. Furthermore, this approach is also suggested as the preferred outlier processing strategy and as a standard way of treating outliers (Dhiren & Andrew, 2012; Thomas & Ward, 2006). Besides that, Etzel et al. (2003) found that this approach work best for the non-normal leptokurtic distribution. This distribution has a more acute peak

around the mean and fatter tails. In addition, Singh, Dev and Mandelbaum (2014) also mention that this approach is a common approach employed in accounting research for handling outliers. Choi, Yoo, Kim and Kim (2014) use this approach to remove the extreme values present in financial data.

Many recent researches utilize this approach to ensure the statistical results are not heavily influenced by the presence of outliers. Ouyang and Wan (2014) have winsorized the differential timeliness ratio to mitigate the influence of outliers. A study conducted by Locorotondo, Dewaelheyns and Van Hulle (2014), for the purpose of examining the differences amount of cash holdings among business group affiliates and non-affiliates used the Winsorization approach in reducing the potential impact of outliers in amount of cash before continuing with the *t*-test.

Ferrara, Marsilli and Ortega (2014) extended Mixed Data Sampling (MIDAS) model in assessing the impact of financial volatility on output growth in three advanced economies namely US, UK and France. They evaluated the role of commodity and stock prices, two major financial variables, in their ability to anticipate the growth. The AR (1)-GARCH (1,1) is used to estimate the volatility of both financial variables. The GARCH process is not robust because it uses a standard maximum likelihood estimator and the researchers have smoothed out outliers from all daily returns of commodity and stock prices through Winsorization.

Another study by Lievenbruck and Schmid (2014) on national hedging volume, cost of goods sold, total debt, market capitalization and annualized inflation employs the

Winsorization approach to restrict the impact of outliers. The study on the cultural differences between countries helps in explaining firms' hedging decision. The cultural differences are measured in terms of the long-term versus short-term orientation, uncertainty avoidance, power distance and masculinity. Chen, Wang and Lin (2014) investigate on the governance role and network centrality of independent director in China. They have winsorized all observations in the top and bottom 1% for continuous variables such as controlling shareholders' tunneling, concentration of ownership, equality restriction, and performance. The use of this approach is to reduce the impact of outlier before continue with correlation and regression analysis.

The utilization of Winsorization approach to minimize the influence of extreme salivary cortisol values has been studied by Wong, Mailick, Greenberg, Hong and Coe (2014). They investigate on the impact of work stress on the awakening cortisol level in mothers of adolescent and adults with and without development disability. Cheng, Cullian and Zhang (2014) apply this approach to reduce the effect of outliers before continuing with correlation and regression analysis. The cash dividends, market to book ratio of equity, percentages of shares and earning divided by book value of equity are all winsorized by 1% in each tail. Nilsen et al. (2012) used Winsorization approach for reducing the effect of outlier in genomic copy number analysis. Very short segments of DNA with deviant copy numbers, technical aberrations or a combination are example of outliers in genomic copy number.

2.8 Type I Error

Rejecting a null hypothesis when in fact there is no difference between treatments can be referred as Type I error and denoted by alpha (α). The alpha value or nominal level should be small since it represents the chances of making an error in the decision. Statisticians generally agree on using these three nominal levels, 0.1, 0.05 and 0.01. The nominal level, 0.05 are widely used by robust statistics researchers (Ahad, Othman & Syed Yahaya, 2011; Keselman et al., 2007; Syed Yahaya, Othman & Keselman, 2006).

A robust test should able to control the Type I error rates near nominal level. In determining the robustness of the test, Bradley (1978) introduced liberal criterion of robustness. According to this criterion, a test is considered robust if its empirical Type I error rates falls within 0.5α to 1.5α, where α is a nominal level. The nominal level used in this study is 0.05. Therefore, the empirical Type I error rates that fall within 0.025 to 0.075 is considered robust. The test is known as not robust with conservative value when the Type I error rates fall below 0.025 and considered not robust with liberal value when the Type I error rates are above the 0.075. The closer the Type I error rates to nominal level, the more robust the procedure is (Syed Yahaya, 2005).

2.9 Power of a Test

The robustness of the test can also be assessed in terms of power. Power of a test is the probability of rejecting a false null hypothesis or the probability that will result in a conclusion that a phenomenon exists. It is denoted by 1- β, where β is the probability of making a Type II error. The Type II error occurs when it is concluded that there is no

treatment effect in the population, when in fact there is a real effect (fail to reject a false null hypothesis). The higher the power, the more sensitive the test is in detecting any difference between treatments. Since power measures the probability, the closer the value to 1, the better the test is in detecting that the phenomenon exists. There are three determinants of power: significance criterion, sample size and effect size (Aberson, 2010; Cohen, 1988).

2.9.1 Significance Criterion

Cohen (1988) defined significance criterion as the standard of proof that a phenomenon exists or a risk of mistakenly rejecting the null hypothesis. This risk is known as Type I error and denoted as alpha (α). The smaller the value of alpha, the more difficult it is to reject the null hypothesis and the harder it is to detect that the phenomenon exist. Setting the higher alpha level makes it easier to reject the null hypothesis as well as to increase the power (Cohen, 1988; Murphy & Myors, 1998). It is easier to reject null hypothesis when the significance level is 0.05 compare with 0.01. Thus, power of a statistical test will increase as the significance criterion increases.

2.9.2 Sample Size

Increase in sample size will increase the sensitivity of the test. Large sample size makes tests highly sensitive and at almost any specific point, the hypothesis can be rejected if test is sufficiently sensitive. In contrast, in small sample sizes, the test may not have enough power to reliably detect the effects (Murphy & Myors, 1998). Thus, increasing the sample size is an effective way to increase the power of the test.

2.9.3 Effect Size

Effect size is another determinant of power and can be defined as the degree to which the phenomenon under study is present in the population (Cohen, 1988). The presence of phenomenon in the population means that the null hypothesis is false. Furthermore, it serves as an index of degree of departure from the null hypothesis. Effect size index can be categorized into small, medium and large (Cohen, 1988). It is easier to detect the effect of a treatment if that effect is large. In contrast the treatments effects can be difficult to be reliably detected when treatments have very small effect.

2.9.3.1 Effect Size Index for Two-Group

For two-group, the effect size index, *d*, is the difference between two central tendency measures divided by the common within-population standard deviation (Cohen, 1988). Therefore, effect size index is defined as:

$$d = \frac{|\theta_1 - \theta_2|}{\sigma_0} \tag{2.11}$$

where

d = effect size index for two-group

θ_1 and θ_2 = the mean of group 1 and group 2, respectively, and

σ_0 = standard deviation of the either population (for condition of equal variances and pool variance for condition of unequal variances).

The pooled standard deviation is derived as in equation 2.12:

$$\sigma_0 = \sqrt{\frac{n_1\sigma_1^2 + n_2\sigma_2^2}{n_1 + n_2}} \tag{2.12}$$

where n_1 and n_2 are the first and second sample size while the σ_1^2 and σ_2^2 are the variance for each first and second groups. The effect size index is considered small when $d = 0.2$, medium when $d = 0.5$ and large when $d = 0.8$ (Cohen, 1988).

2.9.3.2 Effect Size Index for Four-Group

As the number of means increase beyond two ($J > 2$), the relationship between the effect size and the range of standardized means depends upon exactly how the means are dispersed over their range. The spread of the means is represented by the division of a standard deviation by the common standard deviation of the populations, as shown in equation 2.13.

$$f = \frac{\sigma_\theta}{\sigma_0} \tag{2.13}$$

where

σ_θ = the standard deviation of the population means expressed in original scale units, and

σ_0 = the standard deviation within the population.

Therefore, in the case of $J = 4$, d is no longer an effect size index, but it is referred to as the range of standardized means. With four means, the largest and the smallest of the means are used to define the range of standardized means. The range of standardized means, d, is defined as:

$$d = \frac{\theta_{\max} - \theta_{\min}}{\sigma_0} \tag{2.14}$$

θ_{max} = the largest mean

θ_{min} = the smallest mean, and

σ_0 = the (common) standard deviation within the population.

The spread of the means, *f*, is not uniquely determined but it depends on the specification of the pattern of the specification of the means. Cohen (1988) has identified three patterns of variability: minimum variability, intermediate variability and maximum variability. Below are the three patterns described by Cohen (1988) and the relationship between *f* and *d* for each pattern.

Pattern 1:

The minimum variability is defined as:

$$f = d\sqrt{\frac{1}{2J}} \tag{2.15}$$

Pattern 2:

The intermediate variability is defined as:

$$f = \frac{d}{2}\sqrt{\frac{J+1}{3(J-1)}} \tag{2.16}$$

Pattern 3:

The maximum variability is defined as:

$$f = \frac{1}{2} d$$ when J is even and

$$f = d \frac{\sqrt{J^2 - 1}}{2J}$$ when J is odd.

The effect size index is considered small, medium, and large when f = 0.1, 0.25, and 0.4, respectively. Table 2.1 shows the standard pattern variability set by Cohen (1988) for four-group when f = 0.1, 0.25 and 0.4.

Table 2.1

The Standard Pattern Variability for Four-Groups by Cohen (1988)

Effect Size	Pattern Variability
Small	-½ d, 0, 0, ½ d
Medium	-½ d, -¼ d, ¼ d, ½ d
Large	- ½ d,- ½ d, ½ d, ½ d

CHAPTER THREE

RESEARCH METHODOLOGY

This study focuses on modifying the Alexander-Govern test with robust central tendency measure based on Winsorization. The Winsorized mean and adaptive Winsorized mean are two robust central tendency measures used in modifying the Alexander-Govern test as replacements of the arithmetic mean.

There were five variables manipulated in this study in order to create the conditions which are capable to highlight the strength and the weaknesses of all the tests compared. The variables are number of groups, group sizes, group variances, nature of pairings and types of distributions. The robustness of the compared tests is evaluated in terms of their ability in controlling Type I error rates and power of a test.

3.1 Proposed Procedures

The aim of this study is to produce a robust *A*-test which able to be used not only under the violation of homogeneous variances, but also under non-normality. Therefore, in this study, some modification on the test has been done by replacing arithmetic mean with Winsorized mean and adaptive Winsorized mean.

One issue that is always discussed when it comes to Winsorization process is the percentage of Winsorization (α %). This study considers different percentage of

Winsorization (α %) due to recommendations of previous researchers. For example, Hill and Dixon (1982) have considered 10% and 5% for the percentages of trimming (as cited in Keselman et al., 2007), while the other researchers suggested the use of 20% trimming (Lix and Keselman, 1998; Wilcox, 1997). In addition, Abdullah (2011) found that 10% and 15% can be considered as sufficient percentages of trimming capable of producing good control of the Type I error rates under almost all the investigated conditions.

Based on the previous research on the percentage of trimming, this study employs 5%, 10%, 15% and 20% as the percentages of Winsorization resulting in eight modified tests. All the eight modified tests are showed in Figure 3.1.

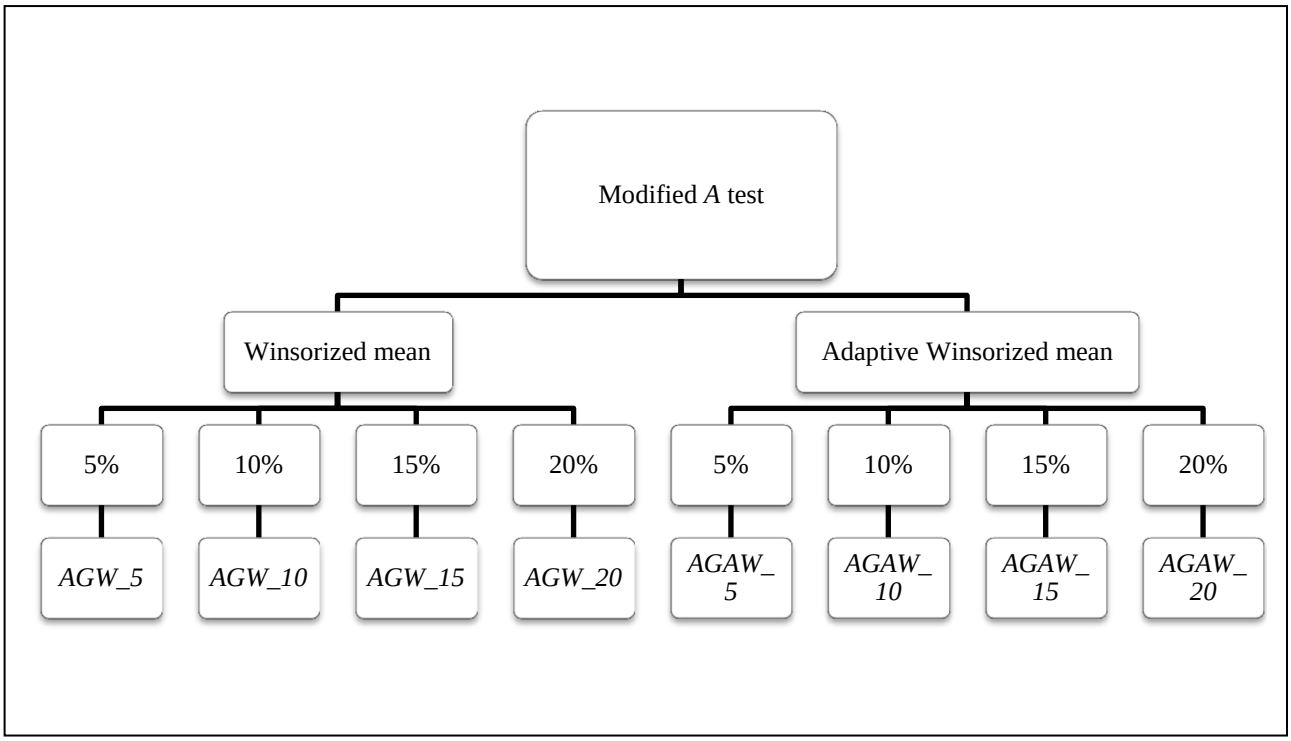

Figure 3.1. The Modified *A*-test with Winsorized mean and adaptive Winsorized mean

The corresponding modified *A*-tests with their total amount of percentages are denoted as *AGW_5*, *AGW_10*, *AGW_15*, *AGW_20*, *AGAW_5*, *AGAW_10*, *AGAW_15* and *AGAW_20* respectively.

The original Alexander-Govern test statistic is testing the equality of mean and assumes it distribution approximately the chi-square distribution. However, the distribution of the test statistic might differ from chi-square as the replacement of central tendency measures with Winsorized mean and adaptive Winsorized mean. Therefore, this study attempts to approximate the distribution of both test statistics; *AGW* and *AGAW*.

For that purpose, a total of 50 *z*-statistics of *AGW* and *AGAW* are generated using simulated data for two-group case and four-group case. Then, the distribution of the *z*-statistics is compared with normal distribution where the null and alternative hypotheses as follow:

H_0: The *z*-statistic of test is follow a normal distribution

H_1: The *z*-statistic of test is not follow a normal distribution

If the z-statistics are independent standard normal variables, then if we create a new variable as the sum of *J* number of squared z-statistics, then a new variable will follow a chi-square distribution with *J*-1 degree of freedom.

The result of the *p*-values of the *z*-statistics for both the *AGW* test and *AGAW* test for two-group case and four-group case are reported in Table 3.1.

Table 3.1

The p-value of the z-statistic

Test	Case	Group	Kolmogorov-Smirnow (Sig.)
AGW	Two-group case (*J*=2)	1	0.200
		2	0.200
	Four-group case (*J*=4)	1	0.200
		2	0.200
		3	0.200
		4	0.200
AGAW	Two-group case (*J*=2)	1	0.200
		2	0.200
	Four-group case (*J*=4)	1	0.200
		2	0.200
		3	0.200
		4	0.200

It is demonstrate that all the *p*-values are greater than nominal level, 0.1 which indicate that the distribution of all *z*-statistics is follow normal distribution. Since the z-statistic of *AGW* and *AGAW* are standard normal variables, therefore we can said that the distribution of *AGW* and *AGAW* test statistic are follow chi-square distribution with *J* -1 degree of freedom.

3.2 Manipulations of Variables

There are five variables manipulated: number of groups, group sizes, group variances, nature of pairings and types of distributions.

3.2.1 Number of Groups

Wilcox, Charlin and Thompson (1986) has found that the *ANOVA* seems to become increasingly sensitive to unequal variances as the number of treatment groups increased and thus it become less robust (as cited in Abdullah, 2011). It is apparent that the

number of groups affected the robustness of the test. Therefore, the current study investigates the two-group and four-group case which has also been studied by previous researchers on the Alexander-Govern test (Abdullah, 2011; Lix & Keselman, 1998; Luh & Guo, 2005).

3.2.2 Group Sizes

The Type I error rates produced by *ANOVA* are inflated for the group with balanced sample sizes and heterogeneous variances. However, it becomes more serious when the group sizes are unbalanced (Lix et al., 1996). Therefore, the present study is considering both the balance and unbalanced sample sizes in examining the robustness of the test.

A studied by Othman, Keselman, Padmanabhan, Wilcox and Fradette (2004) using group sizes of 70 and 90 and the Type I error rates produced are close to nominal level, $\alpha = 0.05$. Furthermore, it can be inferred that the group sizes of any value within the 70 and 90 should produce reasonably good type I error rates (Syed Yahaya, 2005). Therefore, the total group sizes for the case of two-group and four-group used in this study are $N = 40$ and 80 respectively. The same number of total sample sizes also has been used by previous researchers (Abdullah, 2011; Syed Yahaya et al., 2006). For balanced sample size, the number of observation for each group is pegged at 20 which considers $n_1 = n_2 = 20$ for each of two-group case and $n_1 = n_2 = n_3 = n_4 = 20$ for four-group case.

Meanwhile, for unbalanced sample sizes, each of the group is assigned different sample size. The smallest sample size is 10 while the largest is 30. The sample size is assigned

based on proportion 3:5 for two-group case which produced n_1 = 15 and n_2 = 25. For four-group case, it follows the proportion of 2:3:5:6 where the sample size considered are n_1 = 10, n_2 =15, n_3 =25 and n_4 = 30. This proportion for the two-group and four-group has been used by Abdullah (2011) and Syed Yahaya et al. (2006).

3.2.3 Group Variances

The homogeneity of variance is one of the assumptions that need to be satisfied in order to use the classical test for comparing treatment groups. The *ANOVA* has been known to be lack of robustness under the violation of homogeneity of variance (Myers, 1998). Since the heterogeneity of variance might influence the control of Type I error rates and power of a test, it is important to examine the effect of this variable in terms of the Type I error rates and power.

The degree of variance heterogeneity is divided into two categories which are of equal and unequal variances. For equal variances, any value can be used as long as it is equal. In this study, the value of 1 is chosen so that for equal variances for two-group case, the values are 1:1 and for four-group case, the variances are 1:1:1:1. For the case of unequal variances, this study only considers an extreme degree of unequal variances with values of 1:36 for two-group case and 1:1:1:36 for four-group case. Although the ratio of 1:36 and 1:1:1:36 appears extreme, it is actually reasonable in order to see how well the tests perform under extreme condition (Keselman et al., 2007). The underlying idea is that if the test is able to work under an extreme degree of heterogeneity, then it would also be likely to work under most conditions of heterogeneity to be encountered by researchers (Syed Yahaya, 2005).

3.2.4 Nature of Pairings

The nature of pairings is another variable considered in this study which considers positive and negative pairings. Positive pairing is the situation when the larger variance is associated with the larger sample size or the smaller variance is associated with the smaller sample size. Meanwhile the negative pairing is the situation of the larger variance being associated with the smaller sample size or the smaller variance associated with the larger sample size.

Previous researchers has indicated that the Type I error rates declined below the nominal level when the nature of pairing is positive and in contrast, the Type I error rates increased above the nominal level when the nature of pairing is negative (Zimmerman, 2004). Since the nature of pairing has an effect on the Type I error rates, this study considers examining the compared tests with respect to the nature of pairings.

3.2.5 Types of Distributions

The original *A*-test is found to be not robust under several non-normal distributions (Myers, 1998; Schneider & Penfield, 1997). Therefore, the type of distribution is one of the important factors that need to be investigated in evaluating the performance of the test. For that reason, the family of the *g* and *h* distribution are generated to represent four types of distribution.

The four types of distributions are standard normal, symmetric heavy-tailed, skewed normal-tailed and skewed heavy-tailed. In the *g*-and-*h* distribution, parameter *g* controls the degree of skewness while parameter *h* controls the kurtosis. As parameter *g* is

incremented, the degree of skewness will also increase. Meanwhile, for the *h* value, the larger the value, the heavier are the tails of the distribution. Therefore, the types of distributions are generated by controlling the *g* and *h* values where for standard normal, $g = h = 0$, for symmetric heavy-tailed, $g = 0$ and $h = 0.5$, for skewed normal-tailed, $g = 0.5$ and $h = 0$ and for skewed heavy-tailed: $g = 0.5$ and $h = 0.5$.

In addition to these four types of *g*-and-*h* distributions, the log normal distribution is also used. Wilcox (2005) has made criticism of the four *g*-and-*h* distributions that does not represent a large enough departure from normality in term of the skewness of the distribution and suggested to consider the log normal distribution. According to Lix and Keselman (1998), the $g = 1$ and $h = 0$ represents the extremely skewed distribution. Therefore, log normal distribution used in this study is represented by $g = 1$and $h = 0$.

The corresponding skewness and kurtosis of five types of *g*-and-*h* distributions are summarized in Table 3.2.

Table 3.2

Some Properties of the g-and-h Distribution

g	*h*	Skewness	Kurtosis	Distribution shapes
0	0	0	3	Normal
0	0.5	0	Undefined	Symmetric heavy-tailed
0.5	0	1.75	8.9	Skewed normal-tailed
0.5	0.5	Undefined	Undefined	Skewed heavy-tailed
1	0	6.2	114	Extremely skewed normal tailed

(Source: Wilcox (2005))

3.3 Design Specification

To examine the robustness of the proposed tests, the following conditions are created and showed in Table 3.3 and Table 3.4 for two-group case and four-group case, respectively.

Table 3.3

Design Specification for J = 2

Distribution	Group size	Group variance	Nature of pairing
$g = 0, h = 0$	20, 20	1: 1	
	20, 20	1: 36	
	15, 25	1: 1	
	15, 25	1: 36	Positive
	15, 25	36:1	Negative
$g = 0, h = 0.5$	20, 20	1: 1	
	20, 20	1: 36	
	15, 25	1: 1	
	15, 25	1: 36	Positive
	15, 25	36:1	Negative
$g = 0.5, h = 0$	20, 20	1: 1	
	20, 20	1: 36	
	15, 25	1: 1	
	15, 25	1: 36	Positive
	15, 25	36:1	Negative
$g = 0.5, h = 0.5$	20, 20	1: 1	
	20, 20	1: 36	
	15, 25	1: 1	
	15, 25	1: 36	Positive
	15, 25	36:1	Negative
$g = 1, h = 0$	20, 20	1: 1	
	20, 20	1: 36	
	15, 25	1: 1	
	15, 25	1: 36	Positive
	15, 25	36:1	Negative

Table 3.4

Design Specification for J = 4

Distribution	Group size	Group variance	Nature of pairing
$g = 0, h = 0$	20, 20, 20, 20	1: 1: 1: 1	
	20, 20, 20, 20	1: 1: 1: 36	
	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
$g = 0, h = 0.5$	20, 20, 20, 20	1: 1: 1: 1	
	20, 20, 20, 20	1: 1: 1: 36	
	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
$g = 0.5, h = 0$	20, 20, 20, 20	1: 1: 1: 1	
	20, 20, 20, 20	1: 1: 1: 36	
	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
$g = 0.5, h = 0.5$	20, 20, 20, 20	1: 1: 1: 1	
	20, 20, 20, 20	1: 1: 1: 36	
	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
$g = 1, h = 0$	20, 20, 20, 20	1: 1: 1: 1	
	20, 20, 20, 20	1: 1: 1: 36	
	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

3.4 Data Generation

This study requires simulation process to perform the evaluation on the robustness of the tests. The simulation is run using *SAS* generator *RANNOR* function (SAS Institute, 2009). Data from *g*-and-*h* distribution are generated by transforming the standard normal variates, Z_{ij}, using the following equations 3.1 and 3.2.

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

Generally, when dealing with skewed distribution, the central tendency measures such as the Winsorized mean and adaptive Winsorized mean have values unequal to zero. To make certain that the null hypothesis remains true, the observations, Y_{ij}, from each simulated skewed distributions are shifted by subtracting the population central tendency parameter, θ from the observations such that,

$$\boldsymbol{X}_{ij} = \boldsymbol{Y}_{ij} - \boldsymbol{\theta} \quad (3.3)$$

where θ is the Winsorized mean or adaptive Winsorized mean.

The values of θ are determined by computing $\hat{\theta}$ from one million observations generated from the studied distribution. Therefore, when working with Winsorized mean or adaptive Winsorized mean, the population Winsorized mean or adaptive Winsorized mean should be subtracted from Y_{ij} to ensure that the null hypothesis for equal population Winsorized mean or adaptive Winsorized mean remains true.

Based on one million observations generated, the population Winsorized mean and population adaptive Winsorized mean corresponding to the percentage of Winsorization for each of the distributions are listed in Table 3.5.

Table 3.5

Location Parameters with Respect to Distributions

	Location Parameter, ϑ							
	Winsorized mean				Adaptive Winsorized mean			
Type of distributions	5%	10%	15%	20%	5%	10%	15%	20%
$g = 0, h = 0$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$g = 0, h = 0.5$	0.01	0.01	0.01	0.00	0.00	0.00	0.30	0.30
$g = 0.5, h = 0$	0.27	0.18	0.18	0.12	0.27	0.27	0.28	0.30
$g = 0.5, h = 0.5$	0.80	0.31	0.31	0.16	0.80	0.80	0.96	0.96
$g = 1, h = 0$	0.65	0.41	0.41	0.25	0.65	0.65	0.66	0.67

The shifted variates are then transformed to suit the experimental conditions. For the case of heterogeneous variances, each of X_{ij} from equation (3.3) is multiplied by the square root σ_j in order to obtain a distribution with a standard deviation, σ_j as follows,

$$X_{ij} = (Y_{ij} - \theta) \times \sqrt{\sigma_j^2} \tag{3.4}$$

For example, in the case of heterogeneous variances for four-group case, if σ_1^2 = 36, σ_2^2 = 1, σ_3^2 = 1 and σ_4^2 = 1, first, every observation need to be standardized by subtracting each of them with the population's parameter being investigated. Then the standardized observation is multiplied by the standard deviation. For this example, multiplication by 6 is applied to the first group and multiplication by 1 is applied to the remaining three groups. The example of a SAS programming of the modified tests with 5% Winsorized mean and adaptive Winsorized mean under specific condition examined are presented in Appendix B and Appendix C respectively.

In the analysis of the Type I error rates, the groups central tendency measures are set to zero. In contrast, for the power analysis, the group central tendency measure values depend on the setting of the central tendency measures as discussed in Section 3.5. Both Type I error and power rates are obtained by adding the values of central tendency measure to equation (3.4) such that,

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

For each of the design investigated, 5000 dataset are simulated to obtain the Type I error rates and power. The nominal level used is $\alpha = 0.05$.

3.5 The Setting of Central Tendency Measures for Power Analysis

To analyze the power of a test, the group's central tendency measures cannot be set to zero. The values of the alternative hypotheses are determined based on calculation provided by Cohen (1988).

3.5.1 Two-Group Case

According to Cohen (1988), the effect size index, *d*, for two-group case was calculated using the equation (2.11). For this study, the values of θ_1 and θ_2 in equation (2.11) are replaced by the central tendency measure of Winsorized mean or adaptive Winsorized mean.

The effect sizes index are categorized into small for $d = 0.2$, medium for $d = 0.5$ and large for $d = 0.8$. Effect size index is considered small when $d = 0.2$, medium when $d =$

0.5 and large when d = 0.8. Therefore, the values of θ_2 can be determined by equation (2.11) by setting the value of θ_1 = 1 as follows:

$$\theta_2 = d(\sigma_0) + 1 \tag{3.6}$$

The standard deviation (σ_0) depends on whether the population variances are homogeneous or heterogeneous. The calculation of the setting of the central tendency measures are explained as follows:

3.5.1.1 For the case of $\sigma_1 = \sigma_2$, $n_1 = n_2$

Under this condition, the homogeneous variances are set as $\sigma^2_1 = \sigma^2_2 = 1$ and the sample sizes are set as $n_1 = n_2 = 20$ as shown in Table 3.3. Since the variances are homogeneous, then the value of σ_0 in equation (3.6) is equal to 1. The values of θ_2 are calculated using the equation (3.6) for each of effect size index:

d = 0.2: θ_2 = 0.2 (1) + 1 = 1.2

d = 0.5: θ_2 = 0.5 (1) + 1 = 1.5

d = 0.8: θ_2 = 0.8 (1) + 1 = 1.8

The setting of the central tendency measures for this condition is displayed in Table 3.6.

Table 3.6

The Setting of the Central Tendency Measures for Case of $\sigma_1 = \sigma_2$, $n_1 = n_2$

Effect size index, *d*	(θ_1, θ_2)
0.2	(1,1.2)
0.5	(1,1.5)
0.8	(1,1.8)

3.5.1.2 For the case of $\sigma_1 \neq \sigma_2$, $n_1 = n_2$

Under this condition, the variances are set as $\sigma^2_1 = 1$, $\sigma^2_2 = 36$ and the sample sizes are set as $n_1 = n_2 = 20$. For the case of heterogeneous variances with balanced sample size, the value of σ_0 is obtained by taking the square root of the sum of means of two variances, σ^2_1 and σ^2_2 (Cohen, 1988) as follows:

$$\sigma_0 = \sqrt{\frac{\sigma_1^2 + \sigma_2^2}{2}}$$

$$\sigma_0 = \sqrt{\frac{1+36}{2}}$$

$$\sigma_0 = 4.3$$

By inserting $\sigma_0 = 4.3$ and $\theta_1 = 1$, the values of θ_2 are obtained for each effect size index as follows:

$d = 0.2$: $\theta_2 = 0.2\ (4.3) + 1 = 1.86$

$d = 0.5$: $\theta_2 = 0.5\ (4.3) + 1 = 2.15$

$d = 0.8$: $\theta_2 = 0.8\ (4.3) + 1 = 3.44$

Table 3.7 displays the setting of central tendency measure for the case of heterogeneous variances with balanced sample size.

Table 3.7

The Setting of the Central Tendency Measures for Case of $\sigma_1 \neq \sigma_2$, $n_1 = n_2$

Effect size index, *d*	(θ_1, θ_2)
0.2	(1,1.86)
0.5	(1,2.15)
0.8	(1,3.44)

3.5.1.3 For the case of $\sigma_1 = \sigma_2$, $n_1 \neq n_2$

Under this condition, the homogeneous variances are set as $\sigma^2_1 = \sigma^2_2 = 1$ and unbalanced sample sizes are set to be $n_1 = 15$ and $n_2 = 25$. The values of θ_1 and θ_2 are the same as in Table 3.6 since the calculation of σ_0 is similar as the case in **Section 3.5.1.1**.

3.5.1.4 For the case of $\sigma_1 \neq \sigma_2$, $n_1 = n_2$

As mentioned in section (3.2.4) in the case of heterogeneous variances, there are two types of nature pairings (positive and negative). For positive pairing, the variances and sample sizes are paired such that $\sigma_1^2 = 1$ is paired with $n_1 = 15$ and $\sigma_2^2 = 36$ is paired with $n_2 = 25$.

Meanwhile for negative pairing the values of variances and sample sizes are set such that $\sigma_1^2 = 36$ is paired with $n_1 = 15$ and $\sigma_2^2 = 1$ is paired with $n_2 = 25$. In this situation, variance is needed as the value of σ_0 and it is calculated as in equation (2.12).

Therefore, for the situation of positive pairing, the σ_0 is obtained as:

$$\sigma_0 = \sqrt{\frac{(15 \times 1) + (25 \times 36)}{15 + 25}}$$
$$\sigma_0 = 4.78$$

The value of σ_0 for negative pairing is obtained as follows:

$$\sigma_0 = \sqrt{\frac{(15 \times 36) + (25 \times 1)}{15 + 25}}$$
$$\sigma_0 = 3.76$$

The values of θ_2 for positive and negative pairing are calculated using equation (3.6) for each of effect size index:

i. Positive pairing

$d = 0.2$: $\theta_2 = 0.2\ (4.78) + 1 = 1.96$

$d = 0.5$: $\theta_2 = 0.5\ (4.78) + 1 = 3.39$

$d = 0.8$: $\theta_2 = 0.8\ (4.78) + 1 = 4.82$

ii. Negative pairing

$d = 0.2$: $\theta_2 = 0.2\ (3.76) + 1 = 1.75$

$d = 0.5$: $\theta_2 = 0.5\ (3.76) + 1 = 2.88$

$d = 0.8$: $\theta_2 = 0.8\ (3.76) + 1 = 3.00$

The setting of the central tendency measures for the positive and negative pairings condition is displayed in Table 3.8.

Table 3.8

The Setting of the Central Tendency Measures for Case of $\sigma_1 \neq \sigma_2$, $n_1 \neq n_2$

Effect size index, *d*	(θ_1, θ_2)	
	Positive pairing	Negative pairing
0.2	(1,1.96)	(1, 1.75)
0.5	(1,3.39)	(1, 2.88)
0.8	(1,4.82)	(1, 3.00)

3.5.2 Four-Group Case

The setting of central tendency measure for more than two-group case is different from the case of two-group. In this case, *d* is no longer as a difference of standardized mean, but it refers to the range of standardized means as in equation (2.14).

In this study, the values of central tendency measure are replaced by the Winsorized mean and adaptive Winsorized mean instead of the mean. The effect size index for the case of more than two groups, *f*, is obtained based on the pattern of variability. Cohen (1988) has identified three patterns of variability: minimum variability, intermediate variability and maximum variability. The relationship between *f* and *d* depend on these three patterns of variability as discussed in detail in **Section 2.9.3.2**.

The values of *f* can be categorized as small, medium and large with regard to the values of 0.1, 0.25 and 0.4 respectively (Cohen, 1988). Maximum pattern of variability is considered in determining the values of central tendency measures. Thus, for this pattern of variability, the setting of the central tendency measure is displayed in Table 3.9.

Table 3.9

The Setting of the Central Tendency Measures for Four-Group Case.

f	*d*	(θ_1, θ_2, θ_3, θ_4)
0.10	0.2	(-0.1, -0.1, 0.1, 0.1)
0.25	0.5	(-0.25, -0.25, 0.25, 0.25)
0.40	0.8	(-0.4, -0.4, 0.4, 0.4)

3.6 Modified Alexander-Govern Test with Winsorized Mean

The modification of Alexander-Govern test using Winsorized mean, testing the equality of Winsorized mean instead of the mean. The following is the null hypothesis used in this study:

$$H_0 : \mu_w(\alpha)_1 = \mu_w(\alpha)_2 = ... = \mu_w(\alpha)_J$$

where $\mu_w(\alpha)_1 = \mu_w(\alpha)_2 = ... = \mu_w(\alpha)_J$ are the Winsorized mean of *J* independent groups. For each J^{th} group, with size n_j, and the sample Winsorized mean $\bar{x}_w(\alpha)_j$ is calculated using equation (2.8).

Using standard error, $S_{w(\alpha)_j}$ as defined in equation (2.10), a weight, w_j is calculated as:

$$w_j = \frac{1/S^2_{w(\alpha)_j}}{\sum_{j=1}^{J} 1/S^2_{w(\alpha)_j}} \tag{3.7}$$

such that $\sum w_j = 1$

Then the weighted Winsorized mean (x_w^+) is estimated as:

$$x_w^+ = \sum_{j=1}^{J} w_j \bar{x}_w(\alpha)_j \tag{3.8}$$

One-sample *t* statistic is calculated using weighted Winsorized mean as follows:

$$t_{wj} = \frac{\bar{x}_w(\alpha)_j - x_w^+}{S_{w(\alpha)_j}} \tag{3.9}$$

where each of the t_{wj} will be distributed as *t* distribution with $v_j = n_j - 1$ degrees of freedom.

The procedure proceeds to obtain the *z* statistic from normalizing transformation for each of the t_{wj} statistic value by using equation (2.6).

Finally, the value of *AGW* test statistic is obtained by total up the z_j^2 values:

$$AGW = \sum_{j=1}^{J} z_j^2 \tag{3.10}$$

where *AGW* is approximately distributed as $\mathcal{X}^2$ with (*J*-1) degrees of freedom. The null hypothesis is rejected when the value of *AGW* is larger than $\mathcal{X}^2_{J-1}$.

3.7 Adaptive Winsorized Mean

The non-normal data situation appears quite often in real data. Therefore, there is a need to have a central tendency measure which is able to consider the characteristic of the distribution first before winsorized it. For example, if the distribution is skewed to the right, instead of Winsorizing both tails equally, it makes sense to have more observations from the right tail to be winsorized than the left tail of the distribution.

Thus, in this study, the adaptive Winsorized mean is proposed which is based on the adaptive Winsorization approach. This approach gives different percentage of Winsorization for each tail of distribution by examining the shape of the distribution. The HQ_1 hinge estimator is used in order to determine the lower and upper percentage of Winsorization. The lower and upper percentages of Winsorization are representing the percentage of Winsorization for left and right tail of distribution respectively. The adaptive Winsorized mean is obtained by following these five steps:

1. Set the total percentage of Winsorization.
2. Determined the lower and upper Winsorization based on the HQ_1 hinge estimator.
3. Calculate the number of observations to be winsorized from each tail.
4. Winsorized the sample.
5. Calculate the mean on the winsorized sample.

In the first step, we set the different values of Winsorization percentage α = 5%, 10%, 15% and 20%. For the second step, the HQ_1 hinge estimator is used in determining the

lower and upper Winsorization of distribution since this hinge estimator is the best controlled of Type I errors and highly recommended by Keselman et al. (2007). This hinge estimator is based on the measure of tail-length for a given set of *n* observations. Reed and Stark (1996) in their study adopt the notation of Hogg (1974) to define this HQ_1. Based on the sample with the ordered value, let $L_{(\gamma)}$ be the mean of the smallest $[\gamma n]$ observations, where $[\gamma n]$ is rounded down into the nearest integer and $U_{(\gamma)}$ be the largest $[\gamma n]$ observations. For example when γ = 0.2 where γ is the proportion of the observation and therefore, $L_{(.2)}$ is the mean of the smallest 0.2*n* observations.

$$Q_1 = \frac{U_{.2} - L_{.2}}{U_{.5} - L_{.5}} \tag{3.11}$$

The value of Q_1can classified into three categories of tail-length distribution where the value of $Q_1 < 1.81$ implies a light-tailed, $1.81 \leq Q_1 \leq 1.87$ a medium-tailed distribution and $Q_1 > 1.87$ as heavy-tailed distribution. Then, the lower Winsorization percentage is calculated by:

$$\alpha_l = \alpha[HQ_1] \tag{3.12}$$

where $HQ_1$is defined as in Equation 3.13:

$$HQ_1 = \frac{UW_{Q_1}}{UW_{Q_1} + LW_{Q_1}} \tag{3.13}$$

The UW_x and LW_x are the numerator and denominator portions of *x* statistic, in this case the *x* can be refer to the Q_1. While the UW_{Q1} and LW_{Q1} can be given by:

$$UW_{Q_1} = U_{0.2} - L_{0.2}$$

$$LW_{Q_1} = U_{0.5} - L_{0.5}$$

Thus, the upper Winsorization percentage is given by:

$$= \alpha_u = \alpha - \alpha_l \tag{3.14}$$

where α_l and α_u would be α_1 and α_2 respectively.

Then, we calculate the mean based on the sample values that have been winsorized. The adaptive Winsorized mean is defined by:

$$\bar{x}_{aw}(\alpha_1, \alpha_2) = \frac{(m_1+1)x_{m_1+1} + x_{m_1+2} + \ldots + x_{n-m_2-1} + (m_2+1)x_{n-m_2}}{n} \tag{3.15}$$

where $m_1 = [n_j\alpha_1]$ and $m_2 = [n_j\alpha_2]$ while α_1 and α_2 are the lower and upper percentages of Winsorization, respectively.

The standard error of $\bar{x}_{aw}(\alpha_1, \alpha_2)$ can be estimated by

$$s_{\bar{x}_{aw}(\alpha_1,\alpha_2)} = \sqrt{\frac{s^2_{aw}(\alpha_1, \alpha_2)}{n(n-1)}} \tag{3.16}$$

where $s^2_{aw}(\alpha_1, \alpha_2)$ can be estimated as:

$$\begin{aligned} s^2_{aw}(\alpha_1, \alpha_2) = {} & (m_1+1)[x_{m_1+1} - \bar{x}_{aw}(\alpha_1, \alpha_2)]^2 + [x_{m_1+2} - \bar{x}_{aw}(\alpha_1, \alpha_2)]^2 + \ldots + \\ & [x_{n-m_2-1} - \bar{x}_{aw}(\alpha_1, \alpha_2)]^2 + (m_2+1)[x_{n-m_2} - \bar{x}_{aw}(\alpha_1, \alpha_2)]^2 - \\ & \{m_1[x_{m_1+1} - \bar{x}_{aw}(\alpha_1, \alpha_2)] + m_2[x_{n-m_2} - \bar{x}_{aw}(\alpha_1, \alpha_2)]\}^2 / n \end{aligned} \tag{3.17}$$

A manual calculation for adaptive Winsorized mean is presented in Appendix A.

3.8 Modified Alexander-Govern Test with Adaptive Winsorized Mean

The modified *A*-test using adaptive Winsorized mean is tested for the equality of adaptive Winsorized mean instead the mean. The following is the null hypothesis used in this study:

$$H_0 : \mu_{aw}(\alpha_1,\alpha_2)_1 = \mu_{aw}(\alpha_1,\alpha_2)_2 = ... = \mu_{aw}(\alpha_1,\alpha_2)_J$$

where $\mu_{aw}(\alpha_1,\alpha_2)_1 = \mu_{aw}(\alpha_1,\alpha_2)_2 = ... = \mu_{aw}(\alpha_1,\alpha_2)_J$ are the adaptive Winsorized mean of *J* independent groups. For each J^{th} groups with size n_j and the sample adaptive Winsorized mean, $\bar{x}_{aw}(\alpha_1,\alpha_2)_j$ is calculated using equation (3.15):

Using standard error, $S_{aw(\alpha_1,\alpha_2)_j}$ as defined in equation (3.16), a weight w_j is calculated as:

$$w_j = \frac{1/S^2_{aw(\alpha_1,\alpha_2)_j}}{\sum_{j=1}^{J} 1/S^2_{aw(\alpha_1,\alpha_2)_j}} \tag{3.18}$$

such that $\sum w_j = 1$

Then, the weighted adaptive Winsorized mean (x_w^+) is computed as:

$$x_w^+ = \sum_{j=1}^{J} w_j \bar{x}_{aw}(\alpha_1,\alpha_2)_j \tag{3.19}$$

One-sample *t* statistic calculated using weighted adaptive Winsorized mean is estimated as:

$$t_{awj} = \frac{\bar{x}_{aw}(\alpha_1,\alpha_2)_j - x_w^+}{S_{aw(\alpha_1,\alpha_2)_j}} \tag{3.20}$$

where each of the t_{awj} will be distributed as *t* distribution with $v_j = n_j - 1$ degrees of freedom.

The procedure proceeds to obtain the *z* statistic from normalizing transformation for each of the t_{awj}statistic value by using equation (2.6).

Finally, the *AGAW* test statistic is obtained by totalling the z_j^2 values:

$$AGAW = \sum_{j=1}^{J} z_j^2 \tag{3.21}$$

where *AGAW* is approximately distributed as $\mathcal{X}^2$ with (*J*-1) degrees of freedom. The null hypothesis is rejected when the value of *AGAW* is larger than $\mathcal{X}^2_{J-1}$.

3.9 Application on Real Data

The purpose of analysis on real data is to validate the performance of the modified tests in testing the equality of means of three independent groups. The secondary data used for this study is obtained from the study by Ali and Sweeney (1974). The groups of subjects are Group 1, the group of normal healthy laboratory workers, Group II, the group of patients admitted with acute alcoholism with ring sideroblast in bone marrow and Group III, the group of patients admitted with acute alcoholism without ring sideroblast in bone marrow. Using the levels of protoporphyrin in 15 subjects in Group I, 11 subjects in Group II, and 15 subjects in Group III, the question is whether it can be

concluded that there exist differences among the three groups with respect to the protoporphyrin levels.

CHAPTER FOUR

FINDINGS AND DISCUSSIONS

This study emphasizes on the modified *A*-test since this test is not robust under small departure from normal distribution. The modification is done by adopting the existing robust central tendency measures, namely, the Winsorized mean and a proposed central tendency measure called the adaptive Winsorized mean. The performance of all modified tests are evaluated in terms of Type I error rates and power of the tests. To accentuate the strengths and weaknesses of these modified tests, each of them is examined under various experimental conditions such as the number of groups, sample sizes, degree of variance heterogeneity, nature of pairings and types of distributions.

4.1 Type I Error Rates

Performances of the proposed tests in terms of Type I error rates are evaluated using Bradley's liberal criterion of robustness where a test is considered robust if its empirical Type I error rates falls within 0.5α to 1.5α, where α is a nominal level. Since the nominal level used in this study is 0.05, the test is robust if the empirical Type I error rates fall within 0.025 to 0.075. When the Type I error rate is below 0.025, the test is considered not robust with conservative value and when the error rate is above 0.075, it is considered not robust with liberal value. In comparing robustness between tests, the test with a smaller difference between the error rates and the nominal level is considered to be more robust.

Tables 4.1 to 4.8 present the results of all compared tests in terms of Type I error rates according to the types of distribution used and the nature of pairings. The modified *A*-tests with Winsorized mean and adaptive Winsorized mean employ four different percentages of Winsorizing which are 5%, 10%, 15% and 20%. With respect to these Winsorizing percentages, the modified *A*-tests with Winsorized means are denoted as *AGW_5*, *AGW_10*, *AGW_15* and *AGW_20*, while the modified *A*-tests with adaptive Winsorized means are represented as *AGAW_5*, *AGAW_10*, *AGAW_15* and *AGAW_20*. In each table, the Type I error rate values that lie within the liberal criterion of robustness (from 0.025 to 0.075) are written in bold.

4.1.1 Balanced Sample Sizes and Homogeneous Variances

For the condition of balanced sample sizes and homogeneous variances, the results of the analysis on the Type I error rates for $J = 2$ and $J = 4$ are presented in Tables 4.1 and 4.2, respectively.

4.1.1.1 Two-Group Case

In $J = 2$, it is clear that the *AGAW_5* test, *AGAW_10* test, *A*-test and *t*-test are robust regardless of the types of distributions. The *A*-test and *AGAW_5* test performed well under normal distribution producing 0.0008, the smallest difference value from nominal level. The *AGAW_10* test outperformed the other tests with the Type I error rates nearest to nominal level under three distributions: skewed normal-tailed, skewed heavy-tailed and symmetric heavy-tailed while for the extremely skewed normal-tailed, the *t*-test produced the Type I error rate closest to the nominal level.

4.1.1.2 Four-Group Case

The empirical Type I error rates for $J = 4$ under the condition of balanced sample sizes and homogeneous variances are reported in Table 4.2. As demonstrated, only *AGAW_5* test, *AGAW_10* test, *A*-test and *ANOVA* produced the Type I error rates within robust interval across all five distributions. For normal and skewed normal-tailed distributions, the Type I error rates produced by *AGAW_5* test and *A*-test are the nearest to the nominal level compared to the others. The *AGAW_10* test performs better than the other tests under heavy-tailed of either symmetric or skewed distribution. As for the extremely skewed normal-tailed, the *ANOVA* produces the smallest difference of Type I error rate to nominal level of about 0.007.

Table 4.1

The Empirical Type I Error Rates for J = 2 under Balanced Sample Sizes and Homogeneous Variances.

	Sample sizes = (20,20), Variances = (1:1)									
Distributions	*AGW_5*	*AGW_10*	*AGW_15*	*AGW_20*	*AGAW_5*	*AGAW_10*	*AGAW_15*	*AGAW_20*	*A*-test	*t*-test
$g = 0, h = 0$	0.0798	0.1132	0.1662	0.2388	**0.0508**	**0.0638**	0.0910	0.0954	**0.0508**	**0.0528**
$g = 0, h = 0.5$	0.1028	0.1640	0.2166	0.2892	**0.0336**	**0.0578**	0.1284	0.1296	**0.0336**	**0.0356**
$g = 0.5, h = 0$	0.0810	0.1246	0.1724	0.2456	**0.0450**	**0.0500**	0.0874	0.0908	**0.0450**	**0.0474**
$g = 0.5, h = 0.5$	0.0960	0.1610	0.2238	0.2968	**0.0264**	**0.0356**	0.1040	0.1056	**0.0264**	**0.0288**
$g = 1, h = 0$	0.0844	0.1406	0.2006	0.2720	**0.0332**	**0.0340**	0.0874	0.0880	**0.0332**	**0.0358**

Note: Bold value indicates the Type I error rates within [0.025, 0.075]

Table 4.2

The Empirical Type I Error Rates for J = 4 under Balanced Sample Sizes and Homogeneous Variances.

	Sample sizes = (20, 20, 20, 20), Variances = (1:1:1:1)									
Distributions	*AGW_5*	*AGW_10*	*AGW_15*	*AGW_20*	*AGAW_5*	*AGAW_10*	*AGAW_15*	*AGAW_20*	*A*-test	*ANOVA*
$g = 0, h = 0$	0.0962	0.1742	0.2794	0.4142	**0.0518**	**0.0728**	0.1214	0.1342	**0.0518**	**0.0518**
$g = 0, h = 0.5$	0.1246	0.2402	0.3614	0.4956	**0.0280**	**0.0602**	0.1720	0.1724	**0.0280**	**0.0336**
$g = 0.5, h = 0$	0.1112	0.1942	0.2972	0.4336	**0.0522**	**0.0604**	0.1250	0.1300	**0.0522**	**0.0550**
$g = 0.5, h = 0.5$	0.1336	0.2593	0.3798	0.5082	**0.0322**	**0.0508**	0.1672	0.1706	**0.0322**	**0.0290**
$g = 1, h = 0$	0.1440	0.2462	0.3580	0.4860	**0.0578**	**0.0610**	0.1534	0.1536	**0.0578**	**0.0430**

Note: Bold value indicates the Type I error rates within [0.025, 0.075].

4.1.2 Balanced Sample Sizes and Heterogeneous Variances

Table 4.3 and Table 4.4 reported the Type I error rates for $J = 2$ and $J = 4$ respectively under balanced sample sizes and heterogeneous variances.

4.1.2.1 Two-Group Case

The robustness of all the compared tests decreases with the presence of heterogeneous variances. The *AGAW_5* test, *AGAW_10* test, *A*-test and *t*-test are robust under normal and symmetric heavy-tailed distributions while for skewed normal-tailed, the *AGAW_5* test, *AGAW_10* test and *A*-test maintain their robustness. It is noted that the *AGAW_5* test and *A*-test produce the Type I error rates closest to nominal level compared to others under normal and skewed normal-tailed distributions while the *AGAW_10* test is the best test when distribution is symmetric heavy-tailed. None of the tests are robust when data are of skewed heavy-tailed and extremely skewed normal-tailed distributions.

4.1.2.2 Four-Group Case

Similar with $J = 2$, the robustness of the tests in $J = 4$ decreases as variances becomes heterogeneous as displayed in Table 4.4. The tests that are robust under normal distribution are the *AGAW_5* test and *A*-test. For symmetric heavy-tailed and skewed normal-tailed distributions, the *AGAW_5* test, *AGAW_10* test and *A*-test produce the Type I error rates within the robust interval. Under symmetric heavy-tailed distribution, the Type I error rates of both *AGAW_5* test and *A*-test are close to the conservative value, while the error rate of *AGAW_10* test is closer to the nominal level. Unfortunately none of the tests are robust for skewed heavy-tailed and extremely skewed normal-tailed distributions.

Table 4.3

The Empirical Type I Error Rates for J = 2 under Balanced Sample Sizes and Heterogeneous Variances.

	Sample sizes = (20,20), Variances = (1:36)									
Distributions	*AGW_5*	*AGW_10*	*AGW_15*	*AGW_20*	*AGAW_5*	*AGAW_10*	*AGAW_15*	*AGAW_20*	*A*-test	*t*-test
$g = 0, h = 0$	0.0840	0.1202	0.1704	0.2412	**0.0562**	**0.0704**	0.0990	0.1046	**0.0562**	**0.0618**
$g = 0, h = 0.5$	0.0946	0.1512	0.2144	0.2888	**0.0340**	**0.0560**	0.1912	0.1922	**0.0348**	**0.0412**
$g = 0.5, h = 0$	0.1206	0.1416	0.2064	0.2642	**0.0710**	**0.0732**	0.1272	0.1396	**0.0710**	0.0788
$g = 0.5, h = 0.5$	0.3390	0.2116	0.2910	0.3146	0.1764	0.1948	0.4108	0.4164	0.1812	0.1916
$g = 1, h = 0$	0.2280	0.1988	0.2910	0.3046	0.1222	0.1224	0.2282	0.2332	0.1222	0.1292

Note: Bold value indicates the Type I error rates within [0.025, 0.075].

Table 4.4

The Empirical Type I Error Rates for J = 4 under Balanced Sample Sizes and Heterogeneous Variances.

	Sample sizes = (20, 20, 20, 20), Variances = (1:1:1:36)									
Distributions	*AGW_5*	*AGW_10*	*AGW_15*	*AGW_20*	*AGAW_5*	*AGAW_10*	*AGAW_15*	*AGAW_20*	*A*-test	*ANOVA*
$g = 0, h = 0$	0.1000	0.1746	0.2736	0.4136	**0.0522**	0.0756	0.1240	0.1390	**0.0522**	0.1096
$g = 0, h = 0.5$	0.1260	0.2394	0.3630	0.4944	**0.0282**	**0.0596**	0.2152	0.2164	**0.0280**	0.0784
$g = 0.5, h = 0$	0.1302	0.1952	0.3074	0.4308	**0.0642**	**0.0704**	0.1396	0.1490	**0.0642**	0.1276
$g = 0.5, h = 0.5$	0.3200	0.2820	0.4248	0.5172	0.1266	0.1564	0.4056	0.4110	0.1318	0.2400
$g = 1, h = 0$	0.2316	0.2636	0.4036	0.5014	0.1050	0.1056	0.2398	0.2436	0.1050	0.1760

Note: Bold value indicates the Type I error rates within [0.025, 0.075].

4.1.3 Unbalanced Sample Sizes and Homogeneous Variances

In the case of unbalanced sample sizes with homogeneous variances, the Type I error rates are displayed in Table 4.5 for $J = 2$ and Table 4.6 for $J = 4$.

4.1.3.1 Two-Group Case

Under the conditions of unbalanced sample sizes with homogeneous variances, the number of tests which are considered robust increases in comparison with the test under the conditions of balanced sample sizes. This is shown by the Type I error rates as listed in Table 4.5 which is slightly different from Table 4.1. Under balanced condition, there are four tests: *AGAW_5* test, *AGAW_10* test, *A*-test and *t*-test found to be robust under all distributions. However, under unbalanced condition, the number is six with the additional of two more tests, *AGW_5* and *AGAW_15*, to the four tests mentioned earlier.

The classical *t*-test produces the Type I error rates nearest to nominal level compared to the other tests under normal distribution. For symmetric heavy-tailed and skewed normal-tailed distributions, the *AGAW_10* test produces the Type I error rates closest to the nominal level and followed by the *AGW_5* test. Under skewed heavy-tailed as well as under extremely skewed normal-tailed distributions, the *AGW_5* test produces error rates closest to the nominal level and followed by the *AGAW_15* test and *AGAW_10* test.

4.1.3.2 Four-Group Case

The performance of the tests under unbalanced sample sizes and homogeneous variances is depicted in Table 4.6. The *AGW_5* test, *AGAW_5* test, *AGAW_10* test and *A*-test are robust for all distributions except under extremely skewed distribution. However,

ANOVA performed well with all the Type I error rates within the robust interval. In addition, this test also produces Type I error rates nearest to the nominal level with values of 0.0504, 0.0512 and 0.0442 under normal, skewed normal-tailed and extremely skewed normal-tailed distributions, respectively. Meanwhile, for heavy-tailed distribution of either symmetric or skewed, the *AGAW_10* test produces the Type I error rates closest to nominal level.

Table 4.5

The Empirical Type I Error Rates for J = 2 under Unbalanced Sample Sizes and Homogeneous Variances.

Distributions	Sample sizes = (15,25), Variances = (1:1)									
	AGW_5	*AGW_10*	*AGW_15*	*AGW_20*	*AGAW_5*	*AGAW_10*	*AGAW_15*	*AGAW_20*	*A*-test	*t*-test
$g = 0, h = 0$	**0.0538**	0.0864	0.1392	0.2316	**0.0468**	**0.0546**	**0.0692**	0.0932	**0.0468**	**0.0490**
$g = 0, h = 0.5$	**0.0462**	0.1172	0.1828	0.2744	**0.0284**	**0.0510**	**0.0738**	0.1080	**0.0284**	**0.0374**
$g = 0.5, h = 0$	**0.0518**	0.0938	0.1468	0.2376	**0.0476**	**0.0512**	**0.0612**	0.0954	**0.0476**	**0.0468**
$g = 0.5, h = 0.5$	**0.0490**	0.1204	0.1916	0.2828	**0.0286**	**0.0380**	**0.0580**	0.0870	**0.0286**	**0.0324**
$g = 1, h = 0$	**0.0520**	0.1080	0.1720	0.2646	**0.0396**	**0.0458**	**0.0546**	0.1010	**0.0396**	**0.0382**

Note: Bold value indicates the Type I error rates within [0.025, 0.075].

Table 4.6

The Empirical Type I Error Rates for J = 4 under Unbalanced Sample Sizes and Homogeneous Variances.

Distributions	Sample sizes = (10, 15, 25, 30), Variances = (1:1:1:1)									
	AGW_5	*AGW_10*	*AGW_15*	*AGW_20*	*AGAW_5*	*AGAW_10*	*AGAW_15*	*AGAW_20*	*A*-test	*ANOVA*
$g = 0, h = 0$	**0.0692**	0.1558	0.2184	0.4338	**0.0526**	**0.0692**	0.1002	0.1304	**0.0526**	**0.0504**
$g = 0, h = 0.5$	**0.0614**	0.2008	0.2830	0.4966	**0.0254**	**0.0512**	0.1092	0.1448	**0.0254**	**0.0404**
$g = 0.5, h = 0$	**0.0746**	0.1750	0.2412	0.4562	**0.0580**	**0.0722**	0.0892	0.1272	**0.0580**	**0.0512**
$g = 0.5, h = 0.5$	**0.0696**	0.2176	0.3070	0.5122	**0.0302**	**0.0520**	0.1070	0.1374	**0.0302**	**0.0416**
$g = 1, h = 0$	0.1002	0.2314	0.3066	0.5002	0.0796	0.0882	0.1070	0.1530	0.0796	**0.0442**

Note: Bold value indicates the Type I error rates within [0.025, 0.075].

4.1.4 Unbalanced Sample Sizes and Heterogeneous Variances

Table 4.7 and Table 4.8 depict the empirical Type I error rates for the condition of unbalanced sample sizes and heterogeneous variances.

4.1.4.1 Two-Group Case

As shown in Table 4.7, under normal distribution, the Type I error rates produced by *AGW_5* test, *AGAW_5* test, *AGAW_10* test and *A*-test are robust for both pairings while the *AGAW_15* test is robust only for negative pairing. The Type I error rates closest to nominal level are produced by the *AGAW_5* test and *A*-test for positive pairing and by the *AGW_5* test for negative pairing. Under symmetric heavy-tailed distribution, the *AGAW_5* test, *AGAW_10* test and *A*-test produce the Type I error rates within the robust interval for both nature of pairings with *AGAW_10* test having the closest value to the nominal level.

Under skewed normal-tailed distribution, the *AGAW_5* test and *A*-test are the only tests that are robust for both natures of pairings while the *AGW_5* test and *AGAW_10* test are found to be robust in negative pairing and *t*-test is robust for positive pairing. There are no tests that can be considered robust under skewed heavy tailed distribution for both pairings. The *t*-test is the only test that produces robust Type I error rate for positive pairing under extremely skewed normal-tailed.

4.1.4.2 Four-Group Case

As observed in Table 4.8, when distributions are normal and symmetric heavy-tailed, the *AGW_5* test, *AGAW_5* test, *AGAW_10* test and *A*-test are robust for both natures of

pairing. The *ANOVA* test produces robust Type I error rates only for positive pairing under normal distribution. The *AGAW_5* test and *A*-test produce the Type I error rates closest to nominal level regardless of the type of pairings under normal distribution.

Meanwhile for symmetric heavy-tailed distribution, the *AGAW_10* test produces the Type I error rate nearest to nominal level compared to the *AGAW_5* test and *A*-test. The *AGAW_5* test and *A*-test are the only tests that are robust under skewed normal-tailed distribution in both natures of pairing while the *ANOVA* is only robust in positive pairing with the Type I error rate closest to the nominal level. None of the tests can be considered robust under both skewed heavy-tailed and extremely skewed normal tailed distributions.

Table 4.7

The Empirical Type I Error Rates for J = 2 under Unbalanced Sample Sizes and Heterogeneous Variances.

		Sample sizes = (15,25), Variances = (1:36)									
Distributions	Nature of pairing	*AGW_5*	*AGW_10*	*AGW_15*	*AGW_20*	*AGAW_5*	*AGAW_10*	*AGAW_15*	*AGAW_20*	*A*-test	*t*-test
$g = 0, h = 0$	Positive	**0.0744**	0.1026	0.1396	0.2382	**0.0560**	**0.0734**	0.0898	0.1126	**0.0560**	0.0198
	Negative	**0.0486**	0.0852	0.1440	0.2396	**0.0478**	**0.0532**	**0.0674**	0.0856	**0.0478**	0.1268
g = 0, h = 0.5	Positive	0.0838	0.1374	0.1814	0.2832	**0.0394**	**0.0590**	0.1790	0.1674	**0.0380**	0.0110
	Negative	**0.0322**	0.0904	0.1712	0.2716	**0.0296**	**0.0504**	0.0976	0.1626	**0.0298**	0.0998
$g = 0.5, h = 0$	Positive	0.1058	0.1188	0.1616	0.2574	**0.0712**	0.0952	0.1126	0.1220	**0.0712**	**0.0324**
	Negative	**0.0732**	0.1022	0.1718	0.2490	**0.0744**	**0.0738**	0.0804	0.1452	**0.0744**	0.1486
$g = 0.5, h = 0.5$	Positive	0.3000	0.1754	0.2350	0.3038	0.1644	0.1864	0.3848	0.3750	0.1714	0.1142
	Negative	0.1806	0.1374	0.2400	0.2864	0.1780	0.1990	0.2558	0.3542	0.1848	0.2698
$g = 1, h = 0$	Positive	0.1934	0.1634	0.2236	0.2926	0.1108	0.1416	0.1980	0.1954	0.1108	**0.0722**
	Negative	0.1278	0.1436	0.2488	0.2934	0.1328	0.1300	0.1306	0.2480	0.1328	0.2054

Note: Bold value indicates the Type I error rates within [0.025, 0.075].

Table 4.8

The Empirical Type I Error Rates for J = 4 under Unbalanced Sample Sizes and Heterogeneous Variances.

Distribution	Nature of pairing	Sample sizes = (10,15,25,30), Variances = (1:1:1:36)									
		AGW_5	*AGW_10*	*AGW_15*	*AGW_20*	*AGAW_5*	*AGAW_10*	*AGAW_15*	*AGAW_20*	*A*-test	*ANOVA*
$g = 0, h = 0$	Positive	**0.0682**	0.1600	0.2212	0.4250	**0.0534**	**0.0682**	0.0998	0.1322	**0.0534**	**0.0336**
	Negative	**0.0674**	0.1538	0.2128	0.4362	**0.0510**	**0.0676**	0.0976	0.1290	**0.0510**	0.2850
g = 0, h = 0.5	Positive	**0.0630**	0.2096	0.2886	0.4888	**0.0252**	**0.0570**	0.1636	0.2198	**0.0256**	0.0202
	Negative	**0.0616**	0.2002	0.2780	0.4986	**0.0264**	**0.0538**	0.1260	0.1600	**0.0264**	0.2410
$g = 0.5, h = 0$	Positive	0.0862	0.1838	0.2530	0.4444	**0.0602**	0.0816	0.1016	0.1598	**0.0602**	**0.0446**
	Negative	0.0824	0.1830	0.2458	0.4602	**0.0694**	0.0796	0.0970	0.1362	**0.0694**	0.3092
$g = 0.5, h = 0.5$	Positive	0.2480	0.2584	0.3470	0.5166	0.1192	0.1384	0.3428	0.4164	0.1258	0.1550
	Negative	0.1556	0.2476	0.3278	0.5186	0.1208	0.1376	0.2518	0.2764	0.1252	0.3582
$g = 1, h = 0$	Positive	0.1580	0.2492	0.3366	0.4992	0.0934	0.1142	0.1638	0.2650	0.0934	0.0888
	Negative	0.1506	0.2694	0.3364	0.5188	0.1314	0.1406	0.1578	0.1994	0.1314	0.3422

Note: Bold value indicates the Type I error rates within [0.025, 0.075].

4.2 Discussion on Type I Error Rates

The discussion on the comparing the capability among two proposed tests, *AGW* test and *AGAW* test with the *A*-test and classical test is discuss in **Section 4.2.1**. For the *AGW* test and *AGAW* test, they are said as capable to control the Type I error rate under condition examined if at least one of their modified tests is robust either using 5%, 10%, 15% or 20% of Winsorization. This section also discussed on the capability of the proposed test with respect to different types of distributions, group variances and percentages of Winsorization.

4.2.1 Comparison of *AGW* test, *AGAW* test, *A*-test and Classical test.

In this study, the combination of the five variables manipulated: number of groups, group sizes, group variances, nature of pairings and types of distributions, produces a total of 50 conditions as shown in Table 4.1 to Table 4.8. For example, as displayed in Table 4.1, the *AGW* test is incapable to control the Type I error rates under all five conditions examined. In contrast the *AGAW* test, *A*-test and *t*-test are capable to control the Type I error rates under all five conditions examined. Table 4.9 compared the capability of the proposed tests, *AGW* test and *AGAW* test, as well as the *A*-test and classical test under 50 conditions examined. It is shown that the *AGAW* test and *A*-test outperforms the *AGW* test and classical test with their ability to cater a higher number of conditions. Both the *AGAW* test and *A*-test are more capable of controlling the error than the *AGW* test and classical test where they are capable to cater 37 out of 50 conditions or 74% of the conditions are robust, whereas the *AGW* test and classical test can only cater 34% and 52% of the conditions respectively.

Table 4.9

Capability of the Compared Tests

Procedure	Number of conditions	Percentage of capability
AGW	17	34%
AGAW	37	74%
A-test	37	74%
Classical test	26	52%

Note: Total number of conditions is 50

The percentage of capability of the *AGW* test and *AGAW* test with the *A*-test and classical test under various types of distributions is depicted in Table 4.10. Under these three of the different distributions examined: normal, symmetric heavy-tailed and skewed normal-tailed, the *AGAW* test and *A*-test are more capable in controlling the error than the *AGW* test and classical test. Moreover, the *AGAW* test and *A*-test are able to cater all conditions presented under normal and skewed normal-tailed distributions.

It can be observed that under symmetric heavy-tailed, the performance of the *AGAW* test is better than the *A*-test. However the robustness of the all test decreases as distribution become skewed heavy-tailed and extremely skewed normal-tailed distribution, with decreasing percentages of capabilities except for the classical test. The classical test performs well under extremely skewed normal-tailed distribution, where it is capable to cater 50% of the conditions investigated.

Table 4.10

Capability of Compared Tests under Distribution Condition

Type of distributions	Procedure			
	AGW (number of conditions /percentage of capability)	*AGAW* (number of conditions /percentage of capability)	*A*-test (number of conditions /percentage of capability)	Classical test (number of conditions /percentage of capability)
Normal	6 60%	10 100%	10 100%	6 60%
Symmetric heavy-tailed	5 50%	10 100%	9 90%	5 50%
Skewed normal-tailed	3 30%	10 100%	10 100%	6 60%
Skewed heavy-tailed	2 20%	4 40%	4 40%	4 40%
Extremely skewed normal-tailed	1 10%	3 30%	3 30%	5 50%

Note: Total number of conditions is 10 for each of distribution

While the *AGW* test performs quite well under normal and symmetric heavy-tailed distribution compared to the remaining skewed distributions, it performs rather well under normal and symmetric heavy-tailed distribution where the *AGW* test is capable of controlling the Type I error rates catering for 60% and 50% of the conditions, respectively. However, the robustness of the *AGW* test decreases as distribution becomes skewed and worst when distribution is extremely skewed where it can cater only 10% of the conditions.

It is demonstrated that the classical test outperforms the remaining tests: *AGW* test, *AGAW* test and *A*-test with the ability to cater more conditions when variances are homogeneous as shown in Table 4.11. The classical test is more capable of controlling

Type I error rates where it is capable to cater all conditions investigated when variances are homogeneous. This is followed by the *AGAW* test and *A*-test where they are each capable to cater 95% of the conditions. Meanwhile, the *AGW* test produces the lowest percentage of capability when variances are homogeneous with the capability to cater only 45% of the conditions.

Table 4.11

Capability of Compared Tests under Different Group Variances

Group Variances	Test	Number of conditions	Percentage of capability
Homogeneous	*AGW*	9	45%
	AGAW	19	95%
	A-test	19	95%
	Classical test	20	100%
Heterogeneous	*AGW*	8	27%
	AGAW	18	60%
	A-test	18	60%
	Classical test	6	20%

Note: Total number of conditions with homogeneous variances is 20
Note: Total number of conditions with heterogeneous variances is 30

On the other hand, the robustness of the classical test decreases when variances are heterogeneous where the percentage of capability drastically drops from 100% to 20% as stated in Table 4.11. This shows that the violation of assumption of homogeneous variances jeopardized the robustness of the classical test. The *AGAW* test and *A*-test are more capable of controlling the error than the *AGW* test and classical test where they are capable to cater 60% of the conditions when variances are heterogeneous.

In term of the percentage of Winsorization used, the 5% Winsorization is capable of catering 34% of the conditions. Among these four percentages of Winsorization used in *AGW* test, the 5% Winsorization is the only percentage which is able to control the Type I error rates. The remaining percentages of Winsorization: 10%, 15% and 20% of *AGW* test fail to control the Type I error rates under all conditions examined. From the results in Table 4.12, we can conclude that the proposed *AGW* test works well under the small percentage of 5% of Winsorization.

Table 4.12

Capability of Proposed Tests with respect to Different Percentages of Winsorization

Winsorization Percentage	*AGW* Number of conditions	Percentages of capability (%)	*AGAW* Number of conditions	Percentages of capability (%)
5	17	34	37	74
10	0	0	33	66
15	0	0	6	12
20	0	0	0	0

Note: Total number of conditions is 50

The percentages of the capability for *AGAW* test under various percentage of Winsorization are presented in the fifth column of Table 4.12. It is observed that the *AGAW* test performed better under the small percentages of Winsorization, which are 5% and 10%. For the *AGAW* test, the 5% Winsorization is the best amount of Winsorization, capable of catering 37 out of 50 conditions which equivalent to 74%.

Besides 5% Winsorization, the 10% Winsorization in *AGAW* test is also good in controlling the error rates, capable of catering 66% of the conditions. The 5% and 10%

are adequate in order to produce the *AGAW* test with good control of Type I error rates. On the other hand, when the percentage of Winsorization is incremented to 20%, both the *AGAW* test and *AGW* test become not robust.

4.3 Power of a Test

Power of a test refers to the probability that it will lead researchers to reject the null hypothesis, when that hypothesis is in fact false. According to Murphy and Myors (1998), the amount of 50% in power is judged to be an adequate and it is considered high when its value is above 80%. The power of test are assessed only for those tests which are robust as suggested in robust procedures for testing a directional alternative hypothesis of comparing treatment and control groups on multiple outcomes (Lix, Deering, Fouladi & Manivong, 2009). In this study, the assessment of power of test refers to the results written in bold in Tables 4.1- 4.8. The results of the analysis on power for $J = 2$ and $J = 4$ are illustrated in Figure 4.1 until Figure 4.8.

4.3.1 Balanced Sample Sizes and Homogeneous Variances

The results for power of test for $J = 2$ and $J = 4$ are displayed in Figures 4.1 and 4.2, respectively.

4.3.1.1 Two-Group Case

It is obvious that the tests produce adequate power of 50% under normal and skewed normal-tailed distributions when the effect size is greater than 0.5. As demonstrated, the *AGAW_10* test produces the highest power under three types of distributions: normal, symmetric heavy-tailed and skewed heavy-tailed. Under skewed or extremely skewed

normal-tailed distributions, the power values produced by the robust tests are quite similar to each other.

4.3.1.2 Four-Group Case

The power produced by all the investigated tests is high under normal distribution with values above 80% when the effect size is 0.8. Under skewed normal-tailed distribution, the power values are considered adequate with values of more than 50% when the effect size is above 0.5. It is clear that the *AGAW_10* test produces the highest power value regardless of the types of distributions with the exception of the extremely skewed normal-tailed distribution. Under extremely skewed normal-tailed distribution, the power produced by the *AGAW_5* test, *AGAW_10* test and *A*-test are comparable to each other.

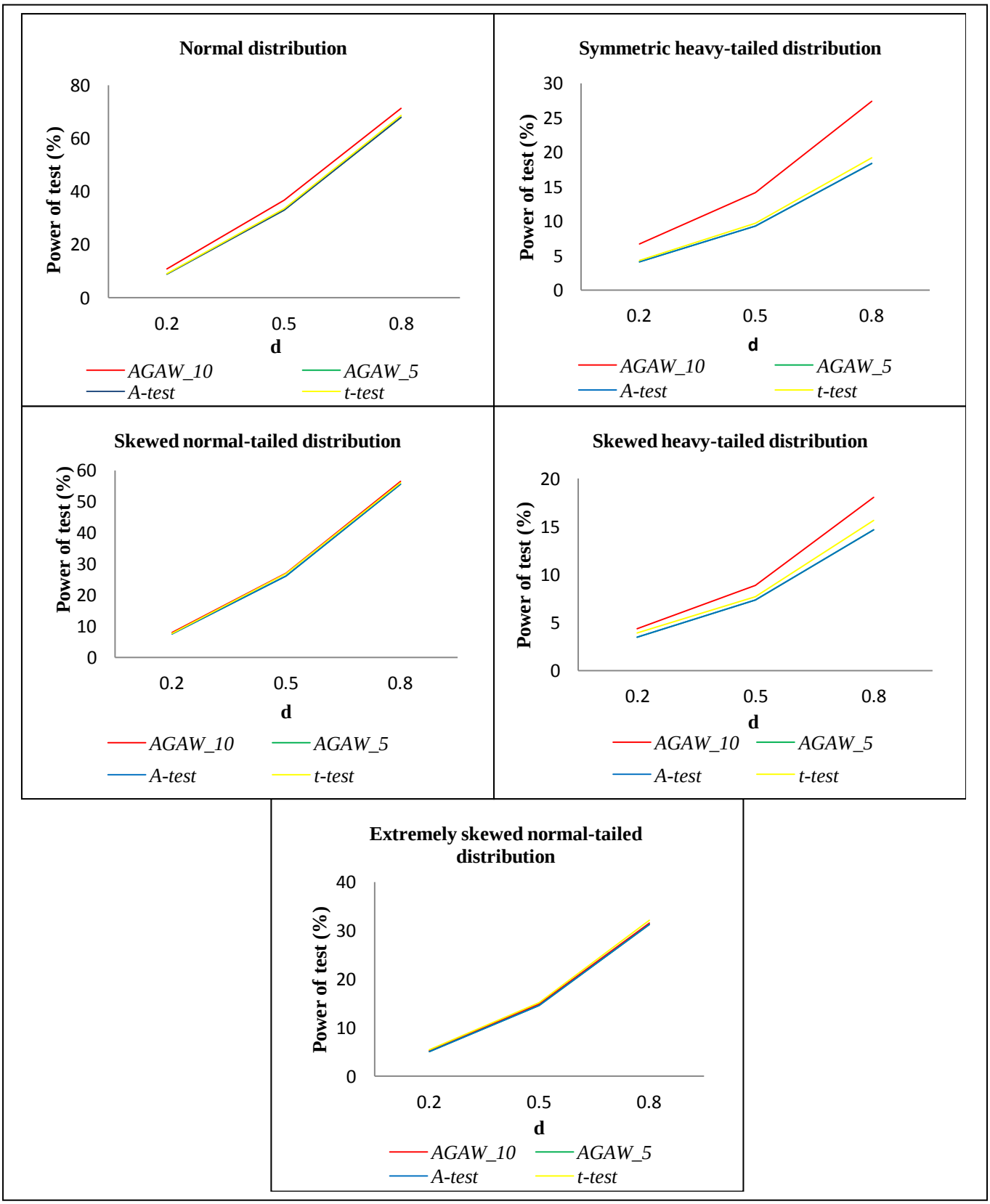

Figure 4.1. Power of Test for $J = 2$ under Balanced Sample Sizes and Homogeneous Variances

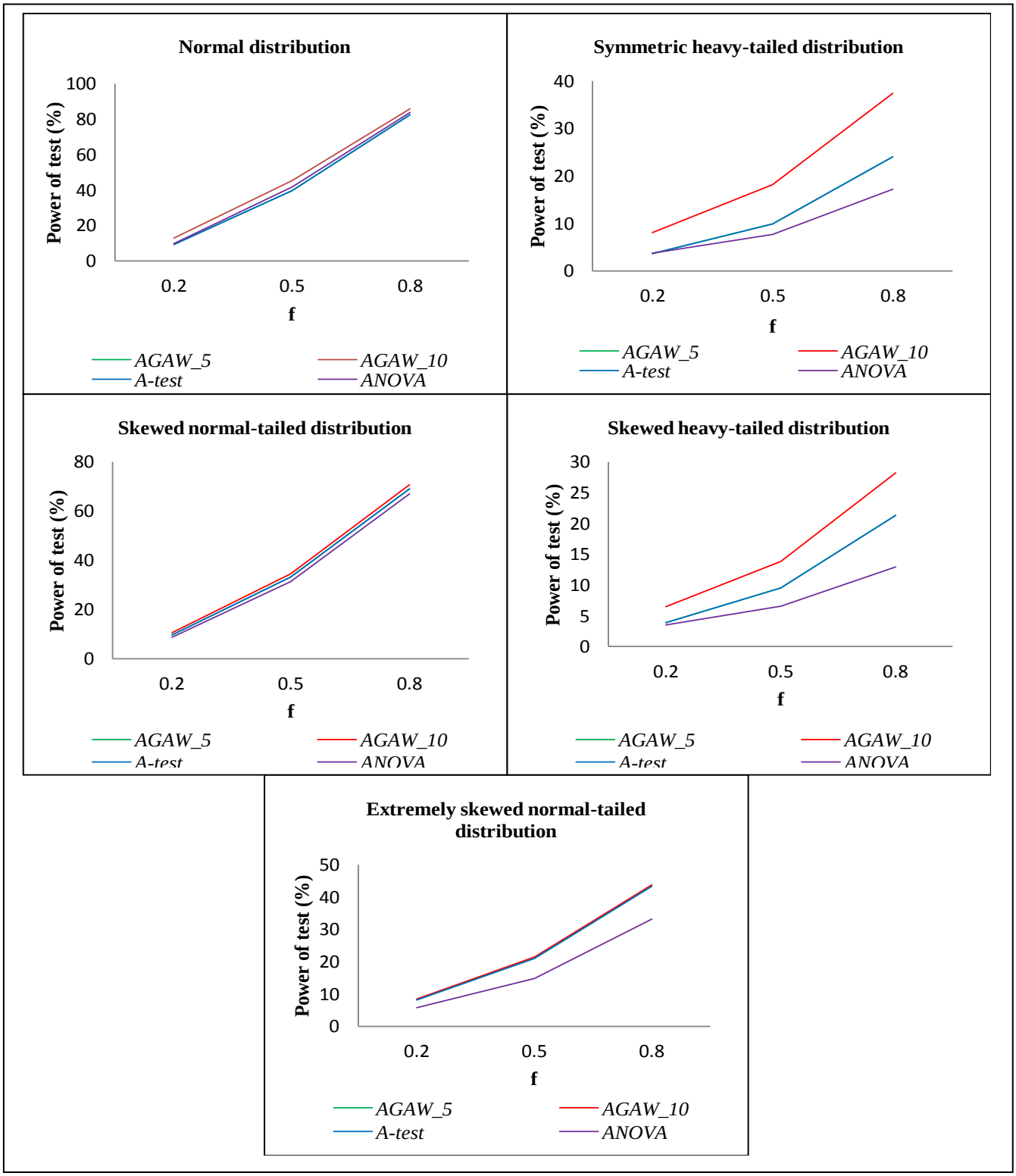

Figure 4.2. Power of Test for $J = 4$ under Balanced Sample Sizes and Homogeneous Variances

4.3.2 Balanced Sample Sizes and Heterogeneous Variances

Figures 4.3 and 4.4 present the resulting power of test for $J = 2$ and $J = 4$ respectively for the condition of balanced sample sizes and heterogeneous variances.

4.3.2.1 Two-Group Case

As can be observed in Figure 4.3, none of the tests produce adequate power. However, the *AGAW_10* test performs the best where it produces the highest power value under these three types of distributions: normal, symmetric heavy-tailed and skewed normal-tailed.

4.3.2.2 Four-Group Case

All the robust tests produce adequate power value for normal distribution where they reached more than 60% when the effect size is 0.8. Meanwhile, under skewed normal-tailed distribution, an adequate power value is obtained when the effect size is 0.8. The *AGAW_10* test produces the highest power values for symmetric heavy-tailed and skewed normal-tailed distributions.

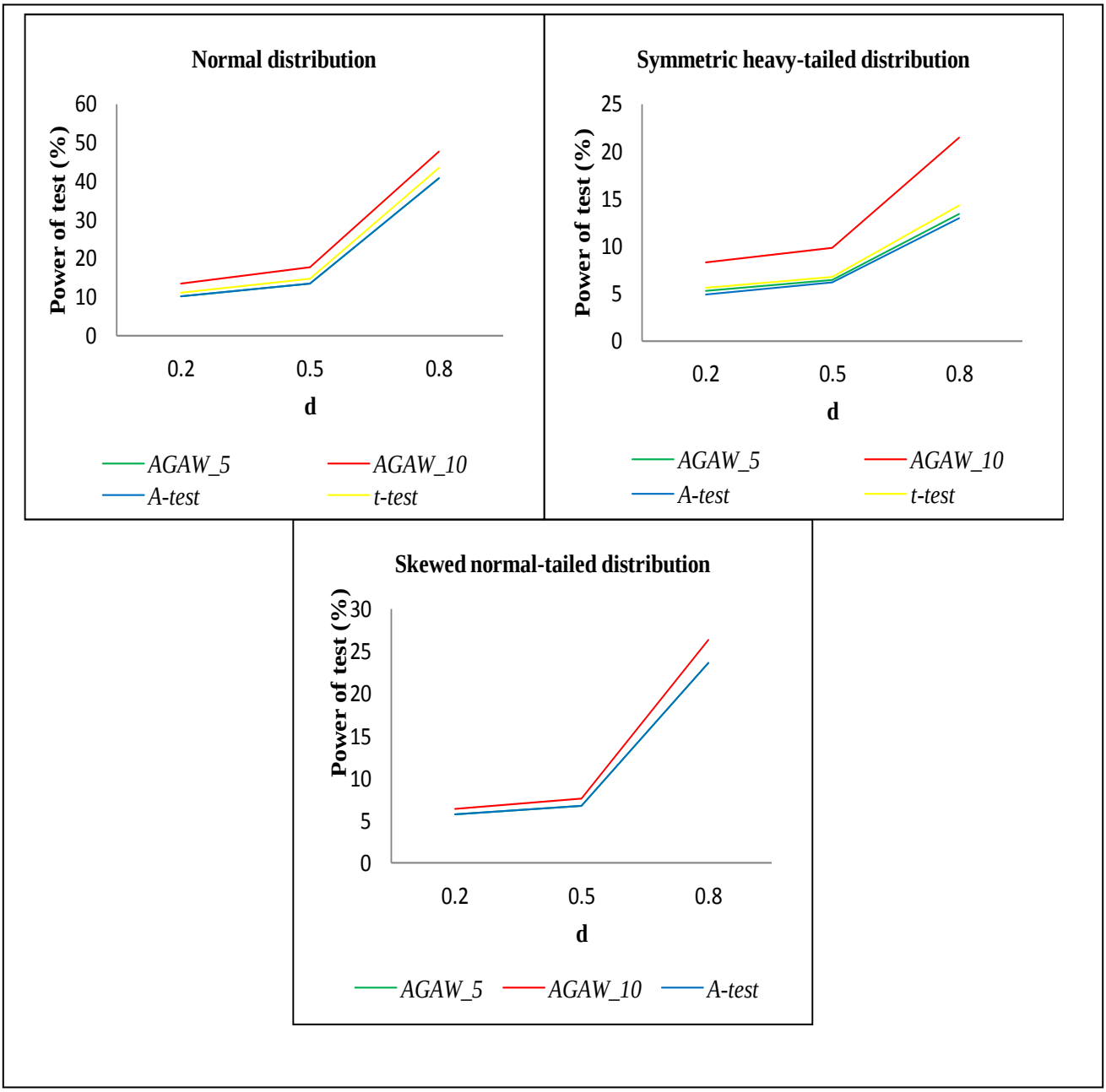

Figure 4.3. Power of Test for $J = 2$ under Balanced Sample Sizes and Heterogeneous Variances

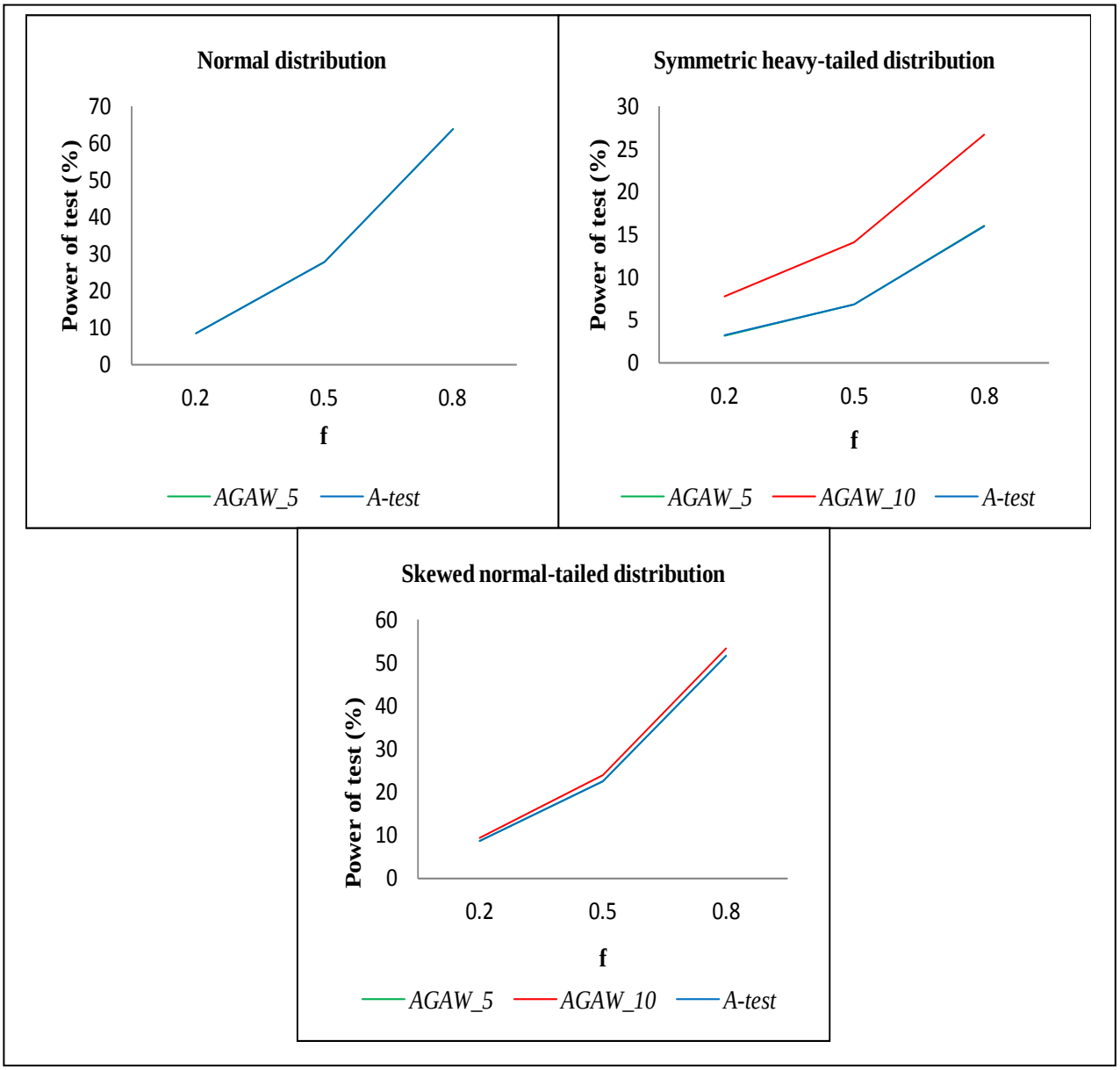

Figure 4.4. Power of Test for $J = 4$ under Balanced Sample Sizes and Heterogeneous Variances

4.3.3 Unbalanced Sample Sizes and Homogeneous Variances

The resulting power of test for $J = 2$ and $J = 4$ are displayed in Figures 4.5 and 4.6, respectively.

4.3.3.1 Two-Group Case

As displayed in Figure 4.5, all the robust tests produce adequate power values for both normal and skewed normal-tailed distributions. The power of the tests under normal distribution is higher than skewed normal-tailed distribution with value of more than 60% when effect size is 0.8. Apparently, the *AGAW_15* test produces the highest power values under three types of distributions: normal, symmetric heavy-tailed and skewed heavy-tailed. Under skewed normal-tailed distribution, the powers produced by the robust tests are quite similar to each other. Meanwhile, under extremely skewed normal-tailed distribution, all tests produce low power values but the *t*-test produces the lowest power.

4.3.3.2 Four-Group Case

The power of the robust tests under this condition is displayed in Figure 4.6. All the robust tests produce adequate power values under normal and skewed normal-tailed distributions. Under both distributions, the tests produce power of more than 50% when the effect size is above 0.5. The *AGAW_10* test and *AGW_5* test produce the highest power values under symmetric heavy-tailed and skewed heavy-tailed distributions, respectively. Under skewed normal-tailed distribution, the power values produced by the tests are comparable to each other except for *ANOVA* test which produces the lowest.

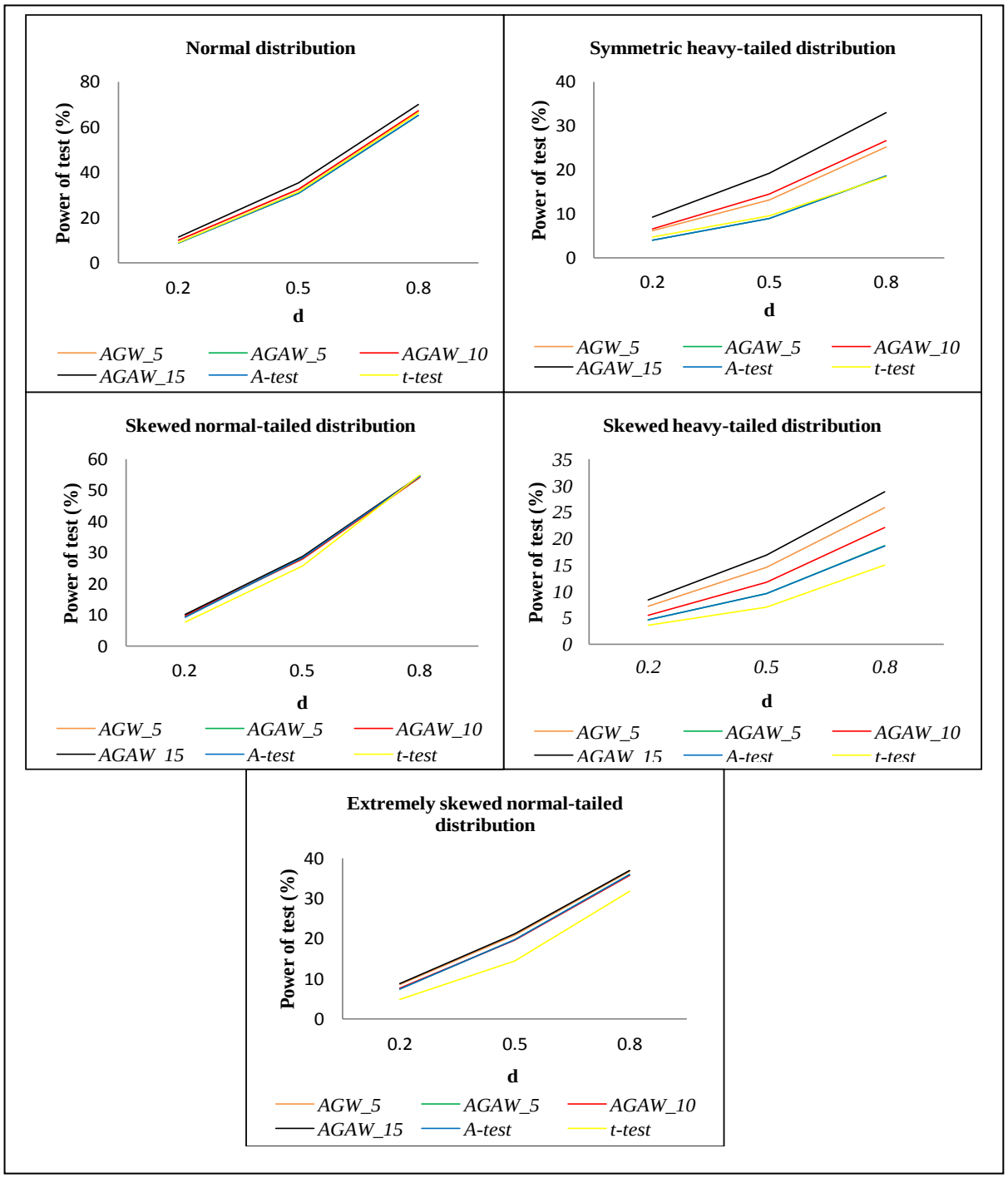

Figure 4.5. Power of Test for $J = 2$ under Unbalanced Sample Sizes and Homogeneous Variances

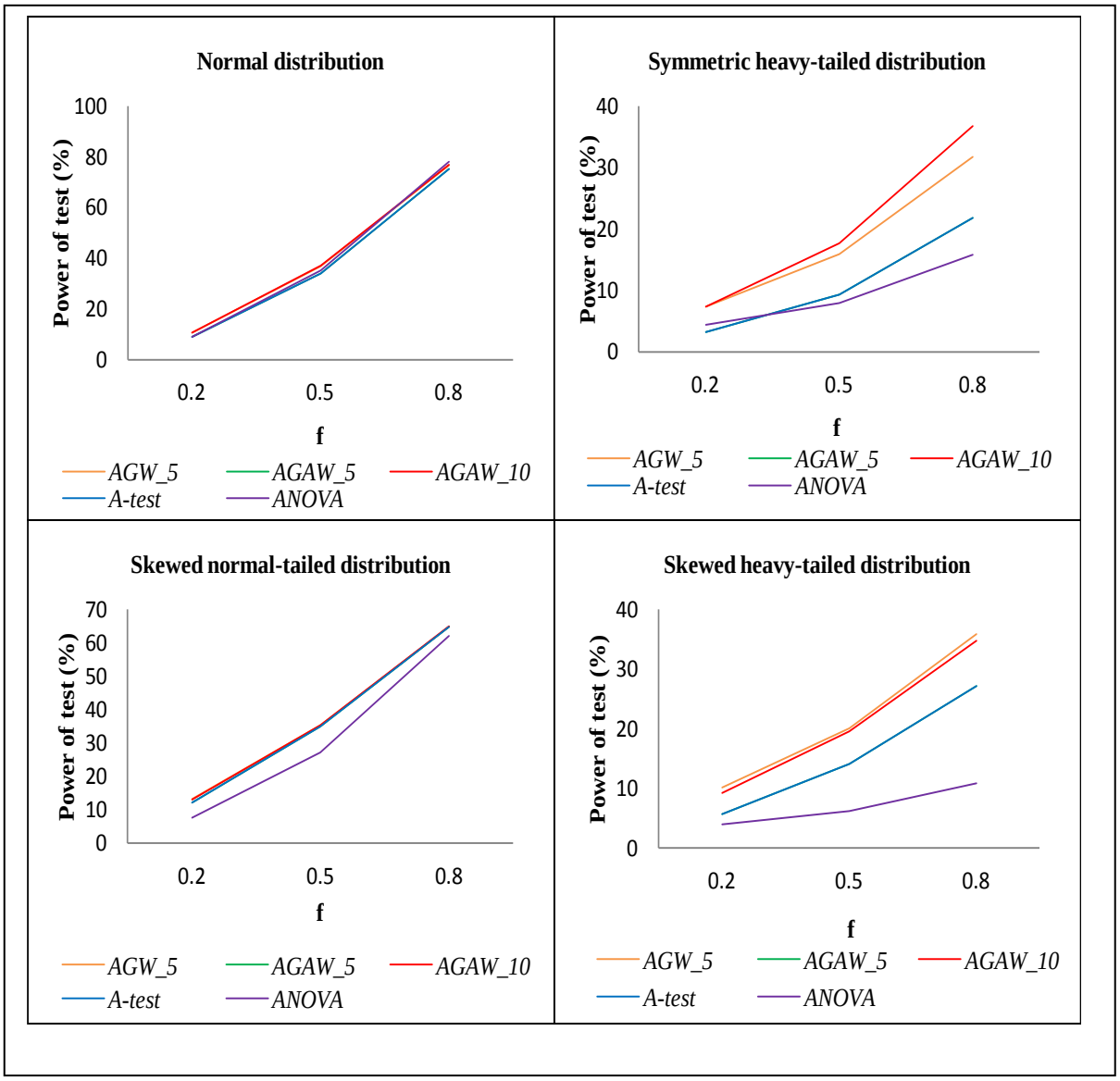

Figure 4.6. Power of Test for $J = 4$ under Unbalanced Sample Sizes and Homogeneous Variances

4.3.4 Unbalanced Sample Sizes and Heterogeneous Variances

The results on the power rates under unbalanced sample sizes and heterogeneous are shown in Figure 4.7 for $J = 2$ and Figure 4.8 for $J = 4$.

4.3.4.1 Two-Group Case

The power produced by the robust tests for both positive and negative pairings are shown in Figure 4.7. Generally, for the two-group case, the performance of the robust tests for positive pairing is better than the negative pairing.

For the case of positive pairing, the robust tests produce high power under normal distribution with the achieved value of 50% when the effect size is greater than 0.5 and above 80% when effect size is 0.8. The power produced by *AGAW_10* test is the highest under both normal and symmetric heavy-tailed distributions. Under skewed normal-tailed distribution, the power of *AGAW_5* test and *A*-test are considered high reaching 80% when the effect size is 0.8.

In contrast, none of the robust tests produce adequate power for the case of negative pairing under all three types of distributions presented. The *AGAW_15* test and *AGAW_10* test produce the highest power values under normal and symmetric heavy-tailed distributions, respectively, whereas for skewed normal-tailed distribution, the power values produced by the robust tests are comparable to each other.

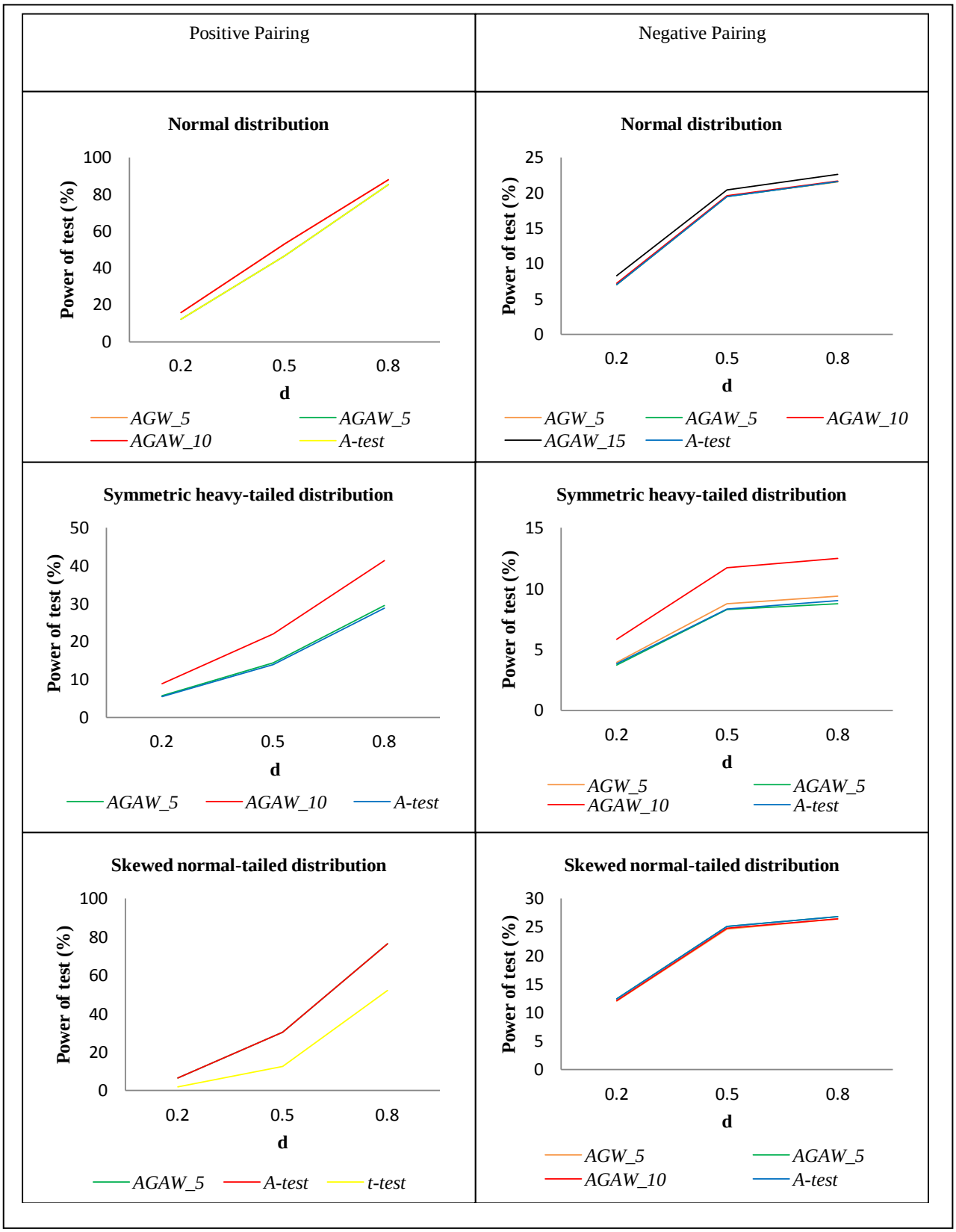

Figure 4.7. Power of Test for $J = 2$ under Unbalanced Sample Sizes and Heterogeneous Variances

4.3.4.2 Four-Group Case

Figure 4.8 displays the power produced by the robust tests for both positive and negative pairing for the four-group case.

For positive pairing, all the robust tests produce adequate power values under normal and skewed normal-tailed distributions except for *ANOVA* test. The tests produce power values of 50% when the effect size is greater than 0.5 for both types of distributions. The power values produced by the *AGW_5* test and *AGAW_10* test are the highest under normal and symmetric heavy-tailed distributions.

On the other hand, for the negative pairing, the robust tests produce adequate power values under normal distribution only. The power values are more than 50% when the effect size is greater than 0.5. The *AGW_5* test and *AGAW_10* test improve the performance of the *A*-test under normal and symmetric heavy-tailed distributions.

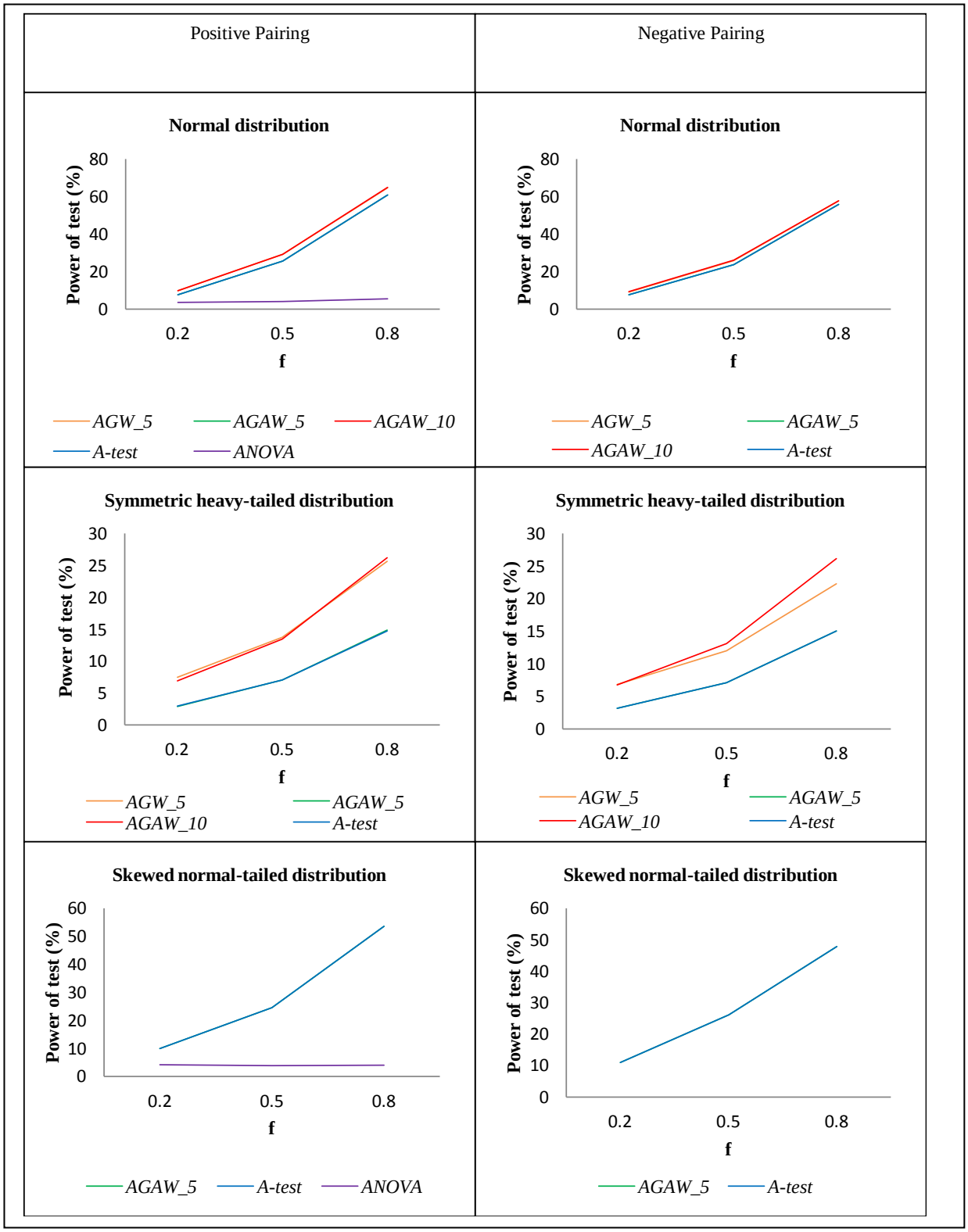

Figure 4.8. Power of Test for $J = 4$ under Unbalanced Sample Sizes and Heterogeneous Variances

4.4 Discussion on Power of Test

The Type I error rates are used to determine robustness of tests. The robust tests are then evaluated in terms of power of test. As observed in Figures 4.1 to 4.8, the performances of the robust tests are considered good under normal and skewed normal-tailed distributions producing adequate power values. However, for heavy-tailed distribution and extremely skewed normal-tailed distribution, the power values drastically drop below the adequate power value.

The presence of homogeneous variances and heterogeneous variances generally influence the performance of the tests. All tests produce adequate power values under normal and skewed normal-tailed distribution when variances are homogeneous. However, when variances are heterogeneous, the power values are sometimes below than the adequate power value.

The sample size does not give much effect to the power value of the tests. Between balanced and unbalanced sample size, the performance of the tests are comparable. In terms of the nature of pairings, the tests produce good power under positive pairing than negative pairing regardless of the distribution.

Overall, the original *A*-test produces lower power compared to the modified *AGAW_10* test which produces the highest power values under most conditions. In contrast, the classical tests (*t*-test and *ANOVA*) produce the lowest power in all the investigated conditions.

4.5 Analysis on Real Data

The performance of the modified Alexander-Govern test namely *AGW* test and *AGAW* test were demonstrated on real data. The following sections discuss the data source, data characteristic and also the performance of all the compared tests.

4.5.1 Data Source

A protoporphyrin dataset is used for performance validation where the dataset consists of the protoporphyrin levels, in milligrams/100 ml RBC, of three groups of subjects. The three groups of subjects are Group I, Group II and Group III where Group I consists of normal healthy laboratory workers, Group II, the group of patients admitted with acute alcoholism with ring sideroblasts in bone marrow and Group III, the group of patients admitted with acute alcoholism without ring sideroblasts in bone marrow. Full details of the protoporphyrin levels in 15 healthy workers and 26 patients are displayed in Table 1 of Appendix D. Testing are done to answer the question of whether it can be concluded that the three groups differ with respect to the protoporphyrin levels.

4.5.2 Data Characteristics

As shown in Figure 4.9 to Figure 4.11 the plotted data values for all three sample groups are not closely following the straight diagonal line which indicates that the distribution are non-normal. From the normal probability plot, we may assume that the distributions are either flatter or peaked since the plotted data are sometimes below and above the diagonal line.

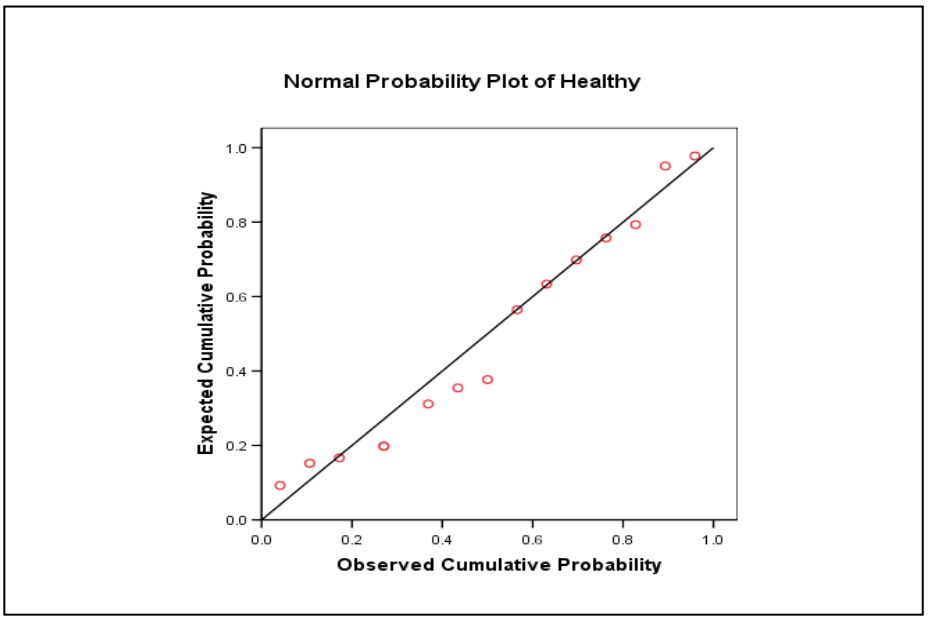

Figure 4.9. Normal Probability Plot for Group I

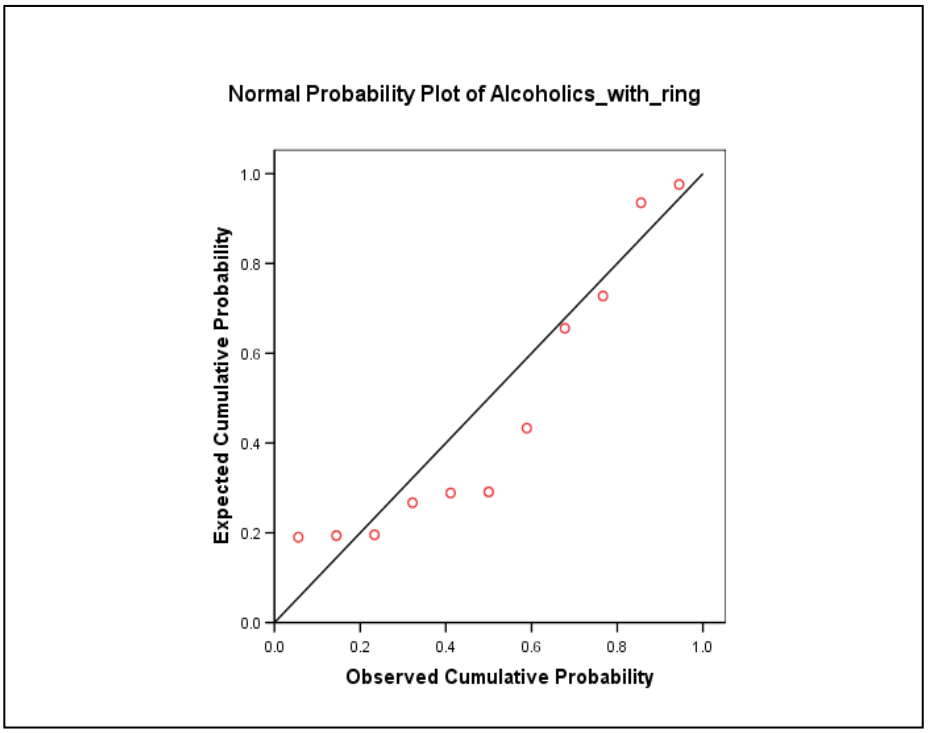

Figure 4.10. Normal Probability Plot for Group II

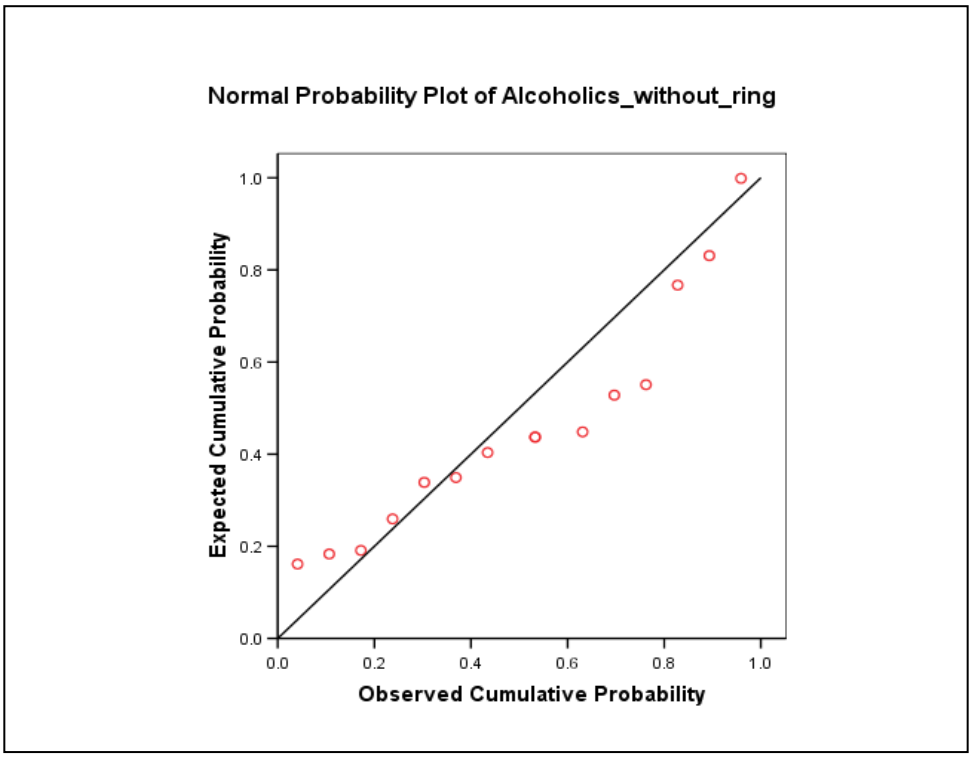

Figure 4.11. Normal Probability Plot for Group III

As shown in Table 4.13, the values of skewness and kurtosis are different than zero. The skewness and kurtosis values for healthy worker are somewhat close to zero, beside the *z* score for both values are within the range of -1.96 to +1.96. This indicates that the distribution of protoporphyrin level for Group I is normal. For Group II and Group III, the values of skewness, kurtosis and *z* score reflect that the distributions are non-normal. The kurtosis value for Group II is positive and very high and possibility of peaked distribution present in this group.

A further testing on the normality assumption of protoporphyrin level is using Shapiro-Wilks test, where the null hypothesis and alternative hypothesis are stated as follow:

H_0: Sample drawn from a population that follows a normal distribution

H_1: Sample drawn from a population that does not follow a normal distribution

The results of the Shapiro-Wilks test for all three groups (Group I, Group II, and Group III) are reported in Table 4.13 (7^{th} column). The Shapiro-Wilks test resulting in rejected the null hypothesis for both groups, Group II and Group III since the *p*-values obtained are less than nominal level, 0.05, which are confirm that both groups are not normally distributed. In contrast, the sample drawn from a population of the Group I follow a normal population since the *p*-value obtained is greater than the nominal level, resulted in fail to reject the null hypothesis.

Table 4.13

Descriptive Statistic for Protoporphyrin Dataset

Groups	Sample size	Skewness	Kurtosis	z-score		Shapiro-Wilks(Sig)
				Skewness	Kurtosis	
Group I	15	0.653	-0.399	1.134	-0.356	0.339
Group II	15	2.141	5.742	3.586	4.976	0.002
Group III	11	1.092	-0.011	1.652	-0.009	0.027

Since the data are not normal, then the nonparametric Levene's test (refer Equation 2.1) is used in testing the equality of variances. The null and alternative hypotheses are as follow:

H_0: The population variances are equal

H_1: The population variances are not equal

A nonparametric Levene's test resulting in rejecting the null hypothesis since the *p*-value is less than 0.05 and verifies the inequality of variances in the samples as can be seen in Table 4.14.

Table 4.14

Nonparametric Levene's Test for Protoporphyrin Dataset

ANOVA

	Sum of Square	*df*	Mean Square	F	Sig.
Between Groups	160.479	2	80.24	4.611	0.016
Within Groups	661.291	38	17.402		
Total	821.771	40			

Therefore, we can conclude that both the normality and homogeneity of variances assumptions are violated for this protoporphyrin dataset.

4.5.3 Testing on Protoporphyrin Dataset

For comparison, the protoporphyrin dataset were tested using all the eight modified *A*-test namely *AGW_5*, *AGW_10*, *AGW_15*, *AGW_20*, *AGAW_5*, *AGAW_10*, *AGAW_15* and *AGAW_20* with *A*-test and *ANOVA*. As can be observed in Table 4.15, all tests produce significant *p*-values implying that there exist differences among the three groups with respect to the protoporphyrin level.

Table 4.15

The p-value of Protoporphyrin Dataset

Tests	*p*-value
AGW_5	0.006624
AGW_10	0.013347
AGW_15	0.014053
AGW_20	0.000576
AGAW_5	0.029533
AGAW_10	0.029533
AGAW_15	0.028761
AGAW_20	0.019035
A-test	0.029533
ANOVA	0.000000

Among these tests, the *ANOVA* shows the strongest significance (*p*-value = 0.000000) followed by *AGW_20* (*p*-value = 0.000576). Although all tests produce significant *p*-values, however the interpretation of the result obtain from *AGW_10*, *AGW_15*, *AGW_20*, *AGAW_15*, *AGAW_20* and *ANOVA* should be done with caution due to these tests fail to control their Type I error rates as shown in the simulation result (refer Table 4.8) for the case of unbalanced sample sizes, heterogeneous variances and non-normal distribution.

CHAPTER FIVE

CONCLUSION

There has been a lot of discussion about the limitation of the Alexander-Govern test, denoted by *A*-test, under non-normal data situations. Thus, previous studies aim to improve the performance of the *A*-test under this limitation (Abdullah, 2011; Lix & Keselman, 1998; Luh & Guo, 2005). They have considered several robust estimators based on trimming approach, namely, trimmed mean, adaptive trimmed mean and modified one-step-*M* estimator (*MOM*) as the central tendency measure in the *A*-test. Their modification enhanced the performance of the *A*-test under skewed distribution. Yet the problem remains under heavy-tailed distribution, producing a robust conservative Type I error rate values with low power.

Hence, this study takes the research into the next step by modifying the *A*-test with a robust central tendency measure based on Winsorization approach. It is believed that this approach is able to improve the performance of the *A*-test generally under non-normal distribution and specifically under heavy-tailed distribution, based on the suggestions in previous research work and the findings on Winsorization approach as reviewed in depth in **Section 1.1** and **Section 2.7**.

This study considers two central tendency measures, namely, Winsorized mean and the adaptive Winsorized mean as replacement to the usual mean in the *A*-test. The Winsorized mean is obtained from symmetrical Winsorization *of* each tail of the

distribution. While, the adaptive Winsorized mean is obtain by symmetrical Winsorization or asymmetrical Winsorization depending on the characteristic of the distribution.

This study considers using 5%, 10%, 15% and 20% as the percentage of Winsorization. Taking into account these four percentages, there are altogether eight modified *A*-tests proposed in this study. Performances of the modified tests are compared to original *A*-test and the two classical tests which are the *t*-test for the two-group case and *ANOVA* for the four-group case. All tests are evaluated under varying conditions to demonstrate the strengths and the weaknesses of the tests. These conditions are created by manipulating five variables that can significantly affect the performance of statistical tests. The five variables are the type of distributions, number of groups, sample size, variance heterogeneity and nature of pairing.

The performance of all the compared tests are measured by the rates of Type I error and power of a test. A test is considered robust if its empirical Type I error rate is between the ranges of 0.025 to 0.075 for 0.05 nominal level (Bradley, 1978). The power of tests is usually judged to be adequate when its value is more than 50% and considered high when it is above 80% (Murphy & Myors, 1998).

5.1 Summary

It is no doubt that the modification in *A*-test capable to enhance the performance of the test when dealing with non-normal data. From using the trimming approach, this study

tries to bring forward the use of Winsorization approach in estimating the central tendency measure used in *A*-test.

We have compared the performance of the two *A*-test modifications: the *AGW* test and *AGAW* test, and found out that the *AGAW* test is superior to *AGW* test with the ability to cater more conditions compared to *AGW* test under various experimental conditions investigated. This finding shows that the utilization of the proposed adaptive Winsorized mean as central tendency measure in *AGAW* test produces a more robust test than the utilization of Winsorized mean in *AGW* test.

The proposed modification of the *A*-test has enhanced the performance of the test where the *AGAW_10* test has succeeded not only in producing a Type I error rate values but also produce the highest power under symmetric heavy-tailed distribution. Regardless of the sample sizes, the number of groups or the variance condition, the *AGAW_10* test performed remarkably well under symmetric heavy-tailed distribution. This test produces Type I error rates nearest to the nominal value as well as the highest power values compared to other tests investigated. This *AGAW_10* test has successfully overcome the limitations of the original *A*-test under this distribution. Besides, for the two-group case under skewed heavy-tailed distribution and extremely skewed distributions with unbalanced sample sizes and homogeneous variances, the *AGW5* is superior to the other tests.

The findings showed that the proposed procedures performed best in terms of controlling the Type I error rates with different Winsorization percentages where the

AGW test performs best with 5% of Winsorizing, while 10% is the best for the *AGAW* test. The amounts of percentages: 5% and 10% are similar to the previous research on trimming approach (Abdullah, 2011; Md Yusof, Abdullah, Syed Yahaya & Othman, 2011). On the other hand, both the *AGW* test and *AGAW* test are not robust for 20% Winsorization. However, this finding is not in line with the finding in the trimming approach where 20% of trimming still suggested as a one of the good trimming percentage (Cribbie, Fisksenbaum, Keselman, & Wilcox, 2012; Keselman et al., 2003; Keselman et al., 2007).

The performance comparison of the *AGW* test and *AGAW* test with the *A*-test and classical test demonstrates that the *AGAW* test and *A*-test are more capable in controlling the Type I error rates than the *AGW* test and classical test. The *AGAW* test and *A*-test are able to cater all conditions considered under normal and skewed normal-tailed distributions. Surprisingly, the *AGAW* test is able to improve the performance of the *A*-test under symmetric heavy-tailed distributions.

Overall the modified tests are capable to achieve the goal of the study in which they produce the tests that are robust under departure from normal and homogeneous variances assumption. They are robust under two types of non-normal distributions namely symmetric heavy-tailed and skewed normal-tailed distribution with the presence of variances heterogeneous. However, none of the modified tests are able to control the Type I error rates under skewed heavy-tailed and extremely skewed normal-tailed distribution when variances are heterogeneous.

The *t*-test and *ANOVA* have performed well under departure from normality as long as the variances remain homogenous. However, when variances are heterogeneous, these classical tests fail to control the Type I error rates even under balanced sample sizes and the results become worse when the sample sizes are unbalanced. This signifies that other than the normality and homoscedasticity assumptions, other factors such as sample size are also necessary to be considered before using the classical tests.

In this study, we also evaluate the performance of the robust tests in producing the power. Observation shows that the power of a test increases as the number of groups increases. In terms of the nature of pairings, positive pairing produce good power values compared to negative pairing regardless of the distribution. In addition, the powers obtained by the tests are adequate under normal and skewed normal tailed distribution. The modified tests produce the highest power in almost all investigated conditions compared to the *A*-test and classical tests. However, the classical tests produce the lowest power in all the investigated conditions. Unfortunately, all tests produce low power when the distribution is heavy-tailed.

This study further validates the performance of the proposed modified test with real data under the violation of both normality and homogeneous variances assumptions. For this purpose, the protoporphyrin dataset consisting of protoporphyrin levels, in milligrams/100 ml RBC, of three groups of subjects are used. The results presented in **Section 4.5.3**, demonstrate that the proposed modified tests are workable on real data application.

5.2 Implication

The proposed modified the *A*-test is robust under non-normal distribution and heterogeneity of variances. Two proposed modified *A*-test, the *AGW* and the *AGAW* tests are proven more robust compared to the *A*-test and classical tests under certain experimental conditions especially under non-normal data distribution.

Therefore, the proposed tests can also serve as an alternative to the classical tests in testing the equality of means of independent groups. Under normal and skewed normal-tailed distribution, the utilization of *A*-test is recommended. However, the *AGAW_10* test is suggested when data comes from symmetric heavy-tailed distribution. With these two alternative tests, the researchers can make use of the original data without having to worry about the shape of distribution or about variance heterogeneity.

5.3 Limitation of the Study

None of the modified tests are robust under all the experimental conditions. Although the modified tests are able to improve the control of Type I error rates especially under symmetric heavy-tailed distribution, they still produce low power. Despite the production of low power, they have improved the performance of the original *A*-test to a certain extent. Another limitation is that they also fail to control Type I error rates under extremely skewed normal-tailed distribution when the variance are heterogeneous.

This study assumes that the type of distribution of each population is the same. For example, in the two-group case under symmetric heavy-tailed distribution, the distributions of both population groups are assumed to be symmetric heavy-tailed

distributions. However, in reality there could be cases where the groups may not have the same type of distribution.

5.4 Suggestion for Future Research

The future research could be undertaken to improve the robustness of the modified tests since none of the modified tests are robust under all the experimental conditions. Researchers may consider using the automatic Winsorization approach in estimating the central tendency measure. This approach is based on the outlier detection method, for example, the modified adjusted box plot as proposed by Hubert and Vandervieren (2008). This box plot is appropriate for skewed and heavy-tailed distribution (Dovoedo, 2011). Therefore, researchers may consider the utilization of the modified adjusted box plot instead of the traditional box plot.

Besides that, Reed and Stark (1996) have defined and introduced seven hinge estimators based on measures of tail-length and skewness for a dataset (i.e. HQ, HQ_1, HH_3, HQ_2, HH_1, HSK_2, and HSK_5). This study only considers the HQ_1 as a hinge estimator in determining the percentage of observation to be winsorized from each tail. The use of other hinge estimators could be put into consideration in future research because there is still a need to improve the performance of the proposed tests under skewed heavy-tailed and extremely skewed distributions.

Finally, since in reality there could be cases where the groups studied may not have the same type of distribution, future research might also consider the condition where different types of distributions exist among the compared groups.

REFERENCES

Abdullah, S. (2011). Kaedah Alexander-Govern Menggunakan Penganggar Teguh Dengan Pendekatan Pangkasan Data: Satu Kajian Simulasi. Unpublished Ph.D thesis, UniversitiUtara Malaysia.

Aberson, C. L. (2010). *Applied power analysis for the behavioral sciences*. New York, NY: Taylor & Francis Group.

Ahad, N. A., Othman, A. R., & Syed Yahaya, S. S. (2011). Comparative Performance of Pseudo-Median Procedure, Welch's Test and Mann-Whitney-Wilcoxon at Specific Pairing. *Modern Applied Statistics, 5*(5), 131-139. doi:10.5539/mas.v5n5p131

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

Ali, M. A. M., & Sweeney, G. (1974). Erythrocyte Coproporpyrin and Protoporphyrin in Ethanol-induced Sideeroblastic Erythropoiesis. *Blood, 43* (2), 291-295.

Amado, C., & Pires, A. M. (2004). Robust Bootstrap with Non Random Weights Based on the Influence Function. *Communications in Statistics-Simulation and Computation, 33* (2), 377-396.

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

Bukat, K., Sitek, J., Kisiel, R., Moser, Z., Gasior, W., Koscielski, M…& Pstrus, J. (2008). Evaluation of the influence of Bi and Sb additions to Sn-Ag-Cu and Sn- Zn alloys on their surface tension and wetting properties using analysis of variance-ANOVA. *Soldering & Surface Mount Technology, 20*(4), 9-19. doi:10.1108/09540910810902660

Chen, E. H., & Dixon, W. J. (1972). Estimates of Parameters of a Censored Regression Sample, *Journal of American Statistical Association, 67*(339), 664-671.

Chen, Y., Wang, Y., & Lin, L. (2014). Independent directors' board networks and controlling shareholders' tunneling behavior. *China Journal of Accounting Research, 7*, 101-118.

Cheng, Z., Cullian, C. P., & Zhang, J. (2014). Free cash flow, growth opportunities, and dividends: Does cross-listing of shares matter? *The Journal of Applied Business Research, 30* (2), 587- 298.

Choi, J., & Zhao, J. (2014). Consumers' behaviors when eating out. Does eating out change consumers' intention to eat healthily? *British Food Journal, 116* (3), 494-509. doi:10.1108/BFJ-06-2012-0136

Choi, J., Yoo, S., Kim, J., & Kim, J. (2014). Capital structure determinants among construction companies in South Korean: A quartile Regression Approach. *Journal of Asian Architecture and Building Engineering, 13* (1), 93-100.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. New York, NY: Academic Press.

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

Dhiren, G., & Andrew, V. (2012). *Survey Research Methods,* 1233-1238.

Dixon, W. J., & Yuen, K. K. (1974). Trimming and Winsorization: A review. *StatistischeHefte, 15*(2), 157-170.

Dovoedo, Y. H. (2011). Contributions to outlier detection methods: Some theory and applications. Published Ph.D thesis, The University of Alabama.

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern Robust Statistical Methods: An Easy Way to Maximize the Accuracy and Power of Your Research. *American Psychological Association, 63*(7), 591–601. doi: 10.1037/0003-066X.63.7.591.

Etzel, C. J., Shete, S., Beasley, T. M., Fernandez, J. R., Allison, D. B., & Amos, C. I. (2003). Effect of Box-Transformation on Power of Haseman-Elston and Maximum- Likelihood Variance Components Tests to Detect Quantitative Trait Loci. *Human Heredity, 55*, 108-116. doi: 10.1159/000072315.

Fan, W., & Hancock, G. H. (2012). Robust Means Modeling: An Alternative for Hypothesis Testing of Independent Means Under Variance Heterogeneity and Nonnormality. *Journal of Educational and Behavioral Statistics, 37* (1), 137-156. doi:10.3102/1076998610396897.

Farcomeni, A., & Ventura, L. (2010). An overview of robust methods in medical research. *Statistical Methods in Medical Research, 21*(2), 111-133. doi:10.1177/0962280210385865.

Farrell-Singleton, P. A. (2010). Critical values for the two independent samples Winsorized T-test. Published Ph.D thesis, Wayne State University.

Ferrara, L., Marsilli, C., & Ortega, J. (2014). Forecasting growth during the Great Recession: Is financial volatility the missing ingredient? *Economic Modeling, 36*, 44-50.

Fried, R. (2004). Robust filtering of time series with trends. *Journal of Nonparametric Statistics, 16* (3-4), 313-328. doi: 10.1080/10485250410001656444.

Fuller, W. A. (1991). Simple estimators for the mean of skewed populations. *Statistica Sinica, 1*, 137-158.

Fung, K. Y., & Rahman, S. M. (1980). The Two-Sample Winsorized T. *Communications in Statistics: Simulation and Computation, 9*(4), 337-347.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance. *American Educational Research Association, 42*(3), 237-288.

Grissom, R. J. (2000). Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology, 68* (1), 155-165.

Hair, J., Black, W., Babin, B., Anderson, R., and Tatham, R. (2006). *Multivariate Data Analysis*, (6th ed.): New Jersey: Pearson Prentice Hall

Hawkins, D. M. (1980). *Identification of outliers*. London: Chapman & Hall.

Hill, M., & Dixon, W. J. (1982). Robustness in real life: A study of clinical laboratory data. *Biometrics, 38*, 377–396.

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

Huber, P. J. (2004). *Robust Statistics*. New Jersey: John Willey & Son, Inc.

Hubert, M., & Vandervieren, E. (2008). An adjusted boxplot for skewed distribution. *Computational Statistics and Data Analysis, 52*, 5186-5201.

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

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

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. (2003). A generally robust approach to hypothesis testing in independent and correlated groups designs. *Psychophysiology, 40*, 586-596.

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 Nonnormality. *Journal of Modern Applied Statistical Methods, 1*(2), 288-309

Keyes, T. M., & Levy, M. S. (1997). Analysis of Levene's test under design imbalance. *Journal of Educational and Behavioral Statistics, 22*, 227-236.

Lance, M. W. (2011). Approximate vs. Monte Carlo Critical Values for the Winsorized T-test. Published Ph.D thesis, Wayne State University.

Lien, D., & Balakrishnan, N. (2005).On regression analysis with data cleaning via trimming, Winsorization and dichotomization. *Communications in Statistics-Simulation and Computation, 34*(4), 839-849.

Lievenbruck, M., & Schmid, T. (2014). Why do firms (not) hedge? Novel evidence on cultural influence. *Journal of Corporate Finance, 25*, 92-106.

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

Lix, L. M., Keselman, H. J., & Hinds, A. M. (2005). Robust tests for the multivariate Behrens-Fisher problem. *Computer Methods and Programs in Biomedicine, 77* (2), 129-139.

Lix, L. M., Keselman, J. C. & Keselman, H. J. (1996). Consequences of Assumption Violations Revisited: A Quantitative Review of Alternatives to the One-Way Analysis of Variance "F" Test. *Review of Educational Research, 66* (4), 579-619.

Lix, L. M., Deering, K. N, Fouladi, R. T. & Manivong, P. (2009). Comparing Treatment and Control Groups on Multiple Outcomes: Robust Procedures for Testing a Directional Alternative Hypothesis. *Education and Psychological Measurement, 69* (2), 198-215. doi:10.1177/0013164408322027.

Locorotondo, R., Dewaelheyns, N., & Van Hulle, C. (2014). Cash holdings and business group membership. *Journal of Business Research, 67*, 316-323.

Luh, W. M., & Guo, J. H. (2005). Heteroscedastic Test Statistics for One-Way Analysis of Variance: The Trimmed Means and Hall's Transformation Conjunction. *The Journal of Experimental Education, 74*(1), 75–100.

Lusk, E. J., Halperin, M. & Heiling, F. (2011).A Note on Power Differentials in Data Preparation between Trimming and Winsorizing. *Business Management Dynamics, 1* (2), 23-31.

MdYusof, Z., Abdullah, S., Syed Yahaya, S. S., & Othman, A. R. (2011). Type I Error Rates of F_t Statistic with Different Trimming Strategies for Two Groups Case. *Modern Applied Sciences, 5* (4), 236-242. doi: 10.5539/mas.v5n4p236.

Mendes, M., & Pala, A. (2003). Type I Error Rate and Power of Three Normality Tests. *Pakistan Journal of Information and Technology, 2* (2), 135-139.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*, 156–166.

Mirtagioglu, H., Yigit, S., Mollaogullari, A., Genc, S., & Mendes, M. (2014). Influence of using alternative means on Type-I error rates in comparison of independent groups. *The Journal of Animal & Plant Sciences, 24* (1), 344-349.

Moir, R. (1998). A Monte Carlo Analysis of the Fisher Randomization Technique: Reviving Randomization for Experimental Economists. *Experimental Economics, 1*, 87-100.

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

Murari, K., & Tater, B. (2014). Employee's attitude towards adaptation of IT-based banking services. A case of Indian private sector banks. *Competitiveness Review, 24* (2), 107-118.doi:10.1108/CR-01-2013-0005.

Murphy, K. R., & Myors, B. (1998). *Statistical power analysis*: *A simple and general model for traditional and modern hypothesis tests*. Mahwah, NJ: Lawrence Erlbaum.

Nilsen, G., LiestAl, K., Loo, P. V., Vollan, H. K. M., Eide, M. B., Rueda, O. M., & LingjArde, O. C. (2012). Copynumber: Efficient algorithms for single- and- multi-track copy number segmentation. *BMC Genomics, 13*, 591-607. doi:10.1186/1471- 2164-13-591.

Nordstokke, D. W., & Zumbo, B. D. (2010). A New Nonparametric Levene Test for Equal Variances. *Psicologica, 31*, 401-430.

Nordstokke, D. W. Zumbo, B. D., Cairns, S. L., & Saklofske, D. H. (2011). The operating characteristics of the nonparametric Levene test for equal variances with assessment and evaluation data. *Practical Assessment, Research & Evaluation, 16* (5), 1-8.

Orr, J. M., Sackett, P. R., & DuBois, C. L. Z. (1991). Outlier detection and treatment in I/O Psychology: A survey of researcher beliefs and an empirical illustration. *Personnel Psychology, 44*, 473-486.

Oshima, T. C., & Algina, J. (1992). Type I error rates for James's second-order test and Wilcoxon's Hm test under heteroscedasticity and non-normality. *British Journal of Mathematical and Statistical Psychology, 45*, 255-263.

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*, 215-234.

Ouyang, B., & Wan, H. (2014). Do analysts understand conservatism? *Accounting and Finance Research, 3* (1), 1-8.

Ozdemir, A. F., Wilcox. R. R., & Yildiztepe, E. (2013). Comparing measures of location: some small-sample result when distributions differ in skewness and kurtosis under heterogeneity of variances. *Communications in Statistics-Simulation and Computation, 42* (2), 407-424. doi: 10.1080/03610918.2011.636163.

Razali. N., & Wah, Y. B. (2011). Power comparison of Shapiro-Wilk, Kolmogrov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics, 2* (1), 21-33.

Reed, J. F. (2005). Contributions to two-sample statistics. *Journal of Applied Statistics, 32*(1), 37- 44. doi: 10.1080/0266476042000305140.

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

Rivest, L. (1994). Statistical properties of Winsorized means for skewed distributions. *Biometrika, 81*(2), 373-383.

SAS Institute Inc. 2009. *SAS/IML 9.2 User's guide*. SAS Institute Inc, Cary, NC.

Schneider, P. J., & Penfield, D. A. (1997). Alexander and Govern's approximation: Providing an alternative to ANOVA under variance heterogeneity. *Journal of Experimental Education, 65*(3), 271-287.

Scholze, M., Boedeker, W., Faust, M., Backhaus, T., Altenburger, R., & Grimme, L. H. (2001). A general best-fit method for concentration-response curves and the estimation of low-effect concentrations. *Environmental Toxicology and Chemistry, 20*(2), 448-457. doi: 10.1002/etc.5620200228.

Shoemaker, L. H. (2003). Fixing the *F* test for equal variances. *American Statistician, 57* (2), 105-114.

Singh, A., Dev, C. S., & Mandelbaum, R. (2014). A flow-through analysis of the U.S. lodging industry during the great recession. *International Journal of Contemporary Hospitality Management, 26* (2), 205-224. doi: 10.1108/IJCHM-12-2012-0260.

Srivastava, D. K., Pan, J., Sarkar, I., & Mudholkar, G. S. (2009). Robust Winsorized Regression Using Bootstrap Approach. *Communications in Statistics-Simulations and Computation, 39* (1), 45-67. doi: 10.1080/03610910903308423.

Syed Yahaya, S. S. (2005). Robust Statistical Procedure for Testing the Equality of Central Tendency Parameters under Skewed Distributions. Unpublished Ph.D thesis, Universiti Sains Malaysia.

Syed Yahaya, S. S., Othman, A. R., & Keselman, H. J. (2006). Comparing the "Typical Score" Across Independent Groups Based on Different Criteria for Trimming. *Metodoskizvezki, 3*(1), 49-62.

Thomas, J. W., & Ward, K. (2006). Economic Profiling of Physician Specialists: Use of Outlier Treatment and Episode Attribution Rules. *Inquiry, 43* (3), 217-282.

Ulusoy, U. (2008). Application of ANOVA to image analysis results of talc particles produced by different milling. *Powder Technology, 188*, 133-138. doi: 10.1016/j.powtec.2008.04.036.

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

Wilcox, R. R. (1990). Comparing the means of two independent groups. *Biometrical Journal, 32*, 771-780.

Wilcox, R. R. (1997). A Bootstrap Modification of the Alexander-Govern ANOVA Method, Plus Comments on Comparing Trimmed Means. *Educational and Psychological Measurement, 57*(4), 655-665.

Wilcox, R. R (2002). Understanding the Practical Advantages of Modern ANOVA Methods. *Journals of Clinical Child and Adolescent Psychology, 31*(3), 399-412.

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing* (2nd ed.): California: Elsevier 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: Simulation and Computation, 15*, 933–943.

Wilcox, R. R., & Keselman, H. (2003). Repeated measures one-way ANOVA based on a modified one-step M-estimator. *British Journal of Mathematical and Statistical Psychology, 56,* 15-25.

Wong, J. D., Mailick, M. R., Greenberg, J. S., Hong, J., & Coe, C. L. (2014). Daily Work Stress and Awakening Cortisol in Mothers of Individuals with Autism Spectrum Disorders or Fragila X Syndrome. *Interdisciplinary Journal of Applied Family Studied, 63,* 135-147. doi: 10.1111/fare.12055.

Yale, C., & Forsythe, A. B. (1976). Winsorized regression. *Technometrics, 18*, 291-300.

Yu, J., & Kabir Hassan, M. (2010). Rational speculative bubbles in MENA stock markets. *Studies in Economics and Finance, 27* (3), 247-264.

Yuen, K. K. (1971). A Note on Winsorized t. *Journal of the Royal Statistical Society Series C (Applied Statistics), 20* (3), 297-304.

Zimmerman, D. W. (2004). Conditional Probabilities of Rejecting *Ho* by Pooled and Separate- Variances *t* Test Given Heterogeneity of Sample Variances. *Communications in Statistics, Simulation and Computation, 33*(1), 69-81.

Zimmerman, D. W., & Zumbo, B. D. (1993). Rank transformations and the Power of the Student t Test and Welch test for Non-Normal Populations with Unequal Variances. *Canadian Journal of Experimental Psychology, 47* (3), 523–539.