ROBUST STATISTICAL PROCEDURES FOR TESTING THE EQUALITY OF CENTRAL TENDENCY PARAMETERS UNDER SKEWED DISTRIBUTIONS

by

SHARIPAH SOAAD SYED YAHAYA

Thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy

May 2005

ACKNOWLEDGEMENTS

Firstly, my sincere appreciations to my supervisor Associate Professor Dr. Abdul Rahman Othman without whose guidance, support, patience, and encouragement, this study could not have materialized. I am indeed deeply indebted to him. My sincere thanks also to my co-supervisor, Associate Professor Dr. Sulaiman for his encouragement and support throughout this study. I would also like to thank Universiti Utara Malaysia (UUM) for sponsoring my study at Universiti Sains Malaysia.

Thanks are due to Professor H.J. Keselman for according me access to his UNIX server to run the relevant programs; Professor R.R. Wilcox and Professor R.G. Staudte for promptly responding to my queries via emails; Pn. Azizah Ahmad and Pn. Ramlah Din for patiently helping me with some of the computer programs and En. Sofwan for proof reading the chapters of this thesis.

I am deeply grateful to my parents and my wonderful daughter, Nur Azzalia Kamaruzaman for their inspiration, patience, enthusiasm and effort. I would also like to thank my brothers for their constant support. Finally, my grateful recognition are due to (in alphabetical order) Azilah and Nung, Bidin and Kak Syera, Izan, Linda, Min, Shikin, Yen, Zurni and to those who had directly or indirectly lend me their friendship, moral support and endless encouragement during my study.

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

LIST OF ABBREVIATIONS

ANOVA	Analysis of variance
H	A statistical method for testing the equality of central tendency measures
MAD_n	Median absolute deviation about the median
MOM	Modified one-step *M*-estimator
Med	Median
MOM-H	A statistical method for testing the equality of central tendency measures
Q_n	A scale estimator
S_1	A statistical method for testing the equality of central tendency measures
S_n	A scale estimator
T_n	A scale estimator

LIST OF PUBLICATIONS

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

Syed Yahaya, S.S., Othman, A.R., and Keselman,H.J. (2004). An Alternative Approach for Testing Location Measures in the One way Independent Group Design. *In Proceedings of the International Conference on Statistics and Mathematics and Its Applications in the Development of Science and Technology*, 4 – 6 October, Bandung, Indonesia.

Syed Yahaya, S.S., Othman, A.R., and Keselman, H.J. (2004). Type I error rates of robust statistical procedures based on robust scale estimators. In *Proceedings of the Simposium Kebangsaan Sains Matematik ke 12 "The Role of Mathematical Sciences in the Development of Biotechnology and k-economy"* 23 -24 Disember, Kuala Lumpur.

Syed Yahaya, S.S., Othman, A.R., and Keselman, H.J. (2005). Comparing the "typical score" across independent groups based on different criteria for trimming. Manuscript submitted for publication in *Advances in Methodology and Statistics.*

Syed Yahaya, S.S., Othman, A.R., and Keselman, H.J. (2005). Testing the equality of central tendency measures for skewed distributions using *MOM-H.* Paper will be presented at the *International Conference Applied Statistics 2005*, 18 – 21 September, Ribno, Slovenia,

PROSEDUR STATISTIK TEGUH BAGI PENGUJIAN KESAMAAN PARAMETER KECENDERUNGAN MEMUSAT UNTUK TABURAN TERPENCONG

ABSTRAK

Kajian ini menyelidik kesan ralat Jenis I dan kuasa keatas dua jenis kaedah teguh. Kaedah pertama dikenali sebagai statistik S_1 yang julung kalinya diselidik oleh Babu et al. (1999). Kaedah ini menggunakan median sebagai sukatan kecenderungan memusat. Satu ciri menarik statistik S_1 ialah data terpencong yang dikaji tidak memerlukan pemangkasan (trimming). Kaedah kedua yang dicadangkan oleh Othman et al. (2004) dikenali sebagai statistik *MOM-H*. Berbeza dengan statistik S_1, *MOM-H* akan memangkas (trim) sebarang nilai yang ekstrim dan tidak seperti min trim, statistik ini secara empirik akan menentukan jumlah pangkasan yang diperlukan dan dengan itu mengelakkan pangkasan yang tidak perlu. Sukatan kecenderungan memusat untuk statistik ini ialah penganggar-*M* satu-langkah terubahsuai seperti yang dicadangkan oleh Wilcox dan Keselman (2003). Dalam kajian ini, kami telah mengubahsuai kedua-dua kaedah statistik tersebut dengan menggabungkan beberapa penganggar skala teguh ke dalam statistik itu. Kami mengenalpasti 4 penganggar skala teguh yang memiliki titik kegagalan yang tertinggi dan fungsi pengaruh yang terbatas seperti yang ditetapkan oleh Rouesseuw dan Croux (1993) iaitu MAD_n, Q_n, S_n, dan T_n. Keempat-empat penganggar skala ini berfungsi secara berbeza dalam setiap kaedah yang dikaji. Untuk statistik S_1, penganggar tersebut menggantikan penganggar skala yang asal bagi membentuk prosedur S_1 terubahsuai dan untuk statistik *MOM-H*, penganggar skala ini telah diguna sebagai kriteria pemangkasan (trimming criterion) yang berfungsi menentukan nilai-nilai sampel

untuk *MOM*. Untuk mengenalpasti keteguhan setiap prosedur, beberapa pembolehubah telah dimanipulasi untuk mewujudkan keadaan yang mampu mengutarakan kekuatan dan kelemahan semua ujian yang dibentuk bagi menentukan kesamaan pengukur kecenderungan memusat. Kesemua pembolehubah ini merangkumi jenis taburan, bilangan kumpulan, sampel bersaiz sama atau sebaliknya, kehomogenan varians dan sifat pasangan saiz sampel dan varians kumpulan. Kajian ini adalah berdasarkan data simulasi dan disebabkan taburan pensampelan statistik S_1 dan *MOM-H* sukar dirungkai, kami telah menggunakan kaedah bootstrap bagi menguji hipotesis kesamaan parameter kecenderungan memusat. Prosedur terubahsuai ini secara lazim menjana kadar Ralat jenis I yang baik dan kadar kuasa yang sederhana. Gabungan skala penganggar T_n atau MAD_n dengan kedua-dua kaedah statistik menghasilkan prosedur teguh yang berpotensi dan mampu menangani masalah pengujian kesamaan sukatan kecenderungan memusat terutamanya bagi taburan terpencong

Katakunci : Statistik teguh, ralat Jenis I, kuasa, bootstrap, taburan terpencong, titik kegagalan

ROBUST STATISTICAL PROCEDURES FOR TESTING THE EQUALITY OF CENTRAL TENDENCY PARAMETERS UNDER SKEWED DISTRIBUTIONS

ABSTRACT

This study examined the effect of Type I error and power on two types of robust methods. The first method is known as the S_1 statistic, which was first studied by Babu et al. (1999). This statistic uses median as the central tendency measure. An interesting characteristic of the S_1 statistic is that the data needs no trimming when skewed. The second method, proposed by Othman et al. (2004) is known as the *MOM-H* statistic. In contrast to the S_1 method, the *MOM-H* statistic will trim any extreme values, and unlike trimmed means, this statistic empirically determines the amount of trimming needed thus avoiding unnecessary trimming. The central tendency measure for this statistic is the modified one-step *M*-estimator (*MOM*) proposed by Wilcox and Keselman (2003). In this study, we modified the two statistical methods by incorporating some of the more robust scale estimators to these statistics. We identified four robust scale estimators with highest breakdown point and bounded influence functions as ascertained by Rouesseuw and Croux (1993) i.e. MAD_n, Q_n, S_n, and T_n. These scale estimators functioned differently in each of the two statistical methods. For the S_1 statistic, the estimators replaced the default scale estimator to form modified S_1 procedures, and for the *MOM-H* statistic, these scale estimators were used as the trimming criterion used to determine the sample values for modified one-step *M*-estimator (*MOM*). To identify the sturdiness or robustness of each procedure, some variables were manipulated to create conditions which are known to highlight the strengths and weaknesses of tests designed to determine the central tendency measures equality. These variables encompassed the types of

distributions, number of groups, equal or unequal sample sizes, variance homogeneity, and nature of pairings of sample sizes and group variances. This study was based on simulated data and since the sampling distributions of the S_1 and *MOM-H* statistics were intractable, we used the bootstrap method for testing the hypotheses of the equality of central tendency measures. The modified procedures, generally, generated good Type I error control and moderate power rates. The combinations of either T_n or MAD_n scale estimators with the two statistical methods produced promising robust procedures that are capable of addressing the problem of testing the equality of central tendency measures especially for skewed distributions.

Keywords: Robust statistics, Type I error, power, bootstrap, skewed distributions, breakdown point

CHAPTER 1
INTRODUCTION

1.1 Introduction

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

The parametric approach in testing the equality of the central tendency parameters continued to play a prominent role because of its capacity to comprehensively describe information contained in a data. However, the good performance and valid application of the procedures require strict adherence to certain assumptions, which do not always operate as predicatively as assumed in the real world. Some of the most common statistical procedures are extremely sensitive to these minor deviations from assumptions such as in the case of normality of distributions and homogeneity of variances. As an example, when computing confidence intervals and testing hypothesis about means, the

methods are based on the assumption that observations are randomly sampled from normal distributions. Another instance is when comparing independent groups; where the methods are also assume that groups have a common variance. Currently, these methods form the backbone of most applied research that involves statistical methodology. It is therefore desirable to construct methods of inference that do not depend on distributional and homoscedasticity (equal variances) assumptions for their validity.

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

1.2 Robust Statistics

In view of all the aforementioned violations, an estimator that is stable and insensitive to all these violations is needed. In other words, the estimator has to be robust. In 1960's, Huber (1964) and Hampel (1968) developed the theory of robustness that paved the way for finding practical solutions in statistics. The theory of robustness developed was basically centered on parametric models. That is whilst their methods recognized that the parametric model might not be the "true" model, but nevertheless made inferences about its parameters with robust and efficient methods. Robustness signifies insensitivity to small deviations from the assumptions (Huber, 1981).

Robust statistics combine the virtues of both, the parametric and the nonparametric approach. In nonparametric inference, few assumptions are made regarding the distribution from which the observations are drawn. In contrast, the approach in robust inference is different wherein there is a working assumption about the form of the distribution, but we are not entirely convinced that the assumption is true. Robustness theories can be viewed as stability theories of statistical inference. What is desired is an inference procedure, which in some sense does almost as well as possible if the assumption is true, but does not perform much worse within a range of alternatives to the assumption. The theories of robustness consider neighborhoods of parametric models and thus belong to parametric statistics. A robust procedure usually adopts what might be called an "applied parametric viewpoint", which according to Huber (1981) uses a parametric model. This model is hopefully a good approximation to the true underlying

situation, but we cannot and do not assume that it is exactly correct. Frequently in discussions of robustness, the assumed distribution (probability density function) is normal, therefore, the type of robustness of interest is "robustness to non-normality".

As mentioned by Wilcox and Keselman (2003) small departures from normality can substantially lower the power when comparing the means of two or more groups. Let us look at the example of analysis of variance (ANOVA), and the drawbacks of this method when assumptions are not met. ANOVA is one of the most commonly used statistical methods for locating treatment effects in one-way independent group design. Generally, violating the assumptions associated with standard ANOVA method can seriously hamper the ability to detect true differences. Non-normality and heteroscedasticity are the two usual assumption violations detected in ANOVA. In particular, when these problems occur at the same time, rates of Type I error are usually inflated, thus resulting in spurious rejections of the null hypotheses. They can also substantially reduce the power of a test, resulting in treatment effects going undetected. Reduction in the power to detect differences between groups occurs because the usual population standard deviation (σ) is very sensitive to outliers and will be greatly influenced by their presence. Consequently, the standard error of the mean (σ^2/n) can become seriously inflated when the underlying distribution has heavy tails (Wilcox and Keselman, 2002). Therefore, the standard error of the F statistics is larger than it should be and power accordingly will be depressed. In order to achieve a good test, one needs to be able to control Type I error and power of test. In other words, neither should power be lost nor Type I error be inflated.

In their efforts to control the Type I error and power rate, investigators looked into numerous robust methods since these methods generally are insensitive to assumptions about the overall nature of the data (e.g. Babu et al., 1999; Keselman et al., 2004b; Kulinskaya, 2003; Luh and Guo, 1999; Othman et al., 2004). Any small deviations from the model assumptions should only slightly impair the performance, for example, the level of a test should be close to the nominal value calculated at the model, and larger deviations from the model should not cause catastrophe. Robust measures of central tendency such as trimmed means, medians or *M*-estimators (refer to Huber, 1981; Staudte and Sheather, 1990; Wilcox, 1997) have been considered as alternatives for the usual least squares estimator, i.e., the usual least squares means, in most research recently (e.g. Keselman et al., 2004b; Luh and Guo, 1999; Wilcox et al., 1998; Wilcox and Keselman, 2002). These measures of central tendency had been shown to have better control over Type I error and power to detect treatment effects (see e.g. Lix and Keselman, 1998; Othman et al., 2004; Wilcox, 1997; Yuen, 1974). Yuen (1974) found these benefits in the two-group case of trimmed means and Lix and Keselman (1998) demonstrated similar results in the more than two-group problem. Other investigators, e.g. Babu, et al. (1999) used median as the central tendency measure when dealing with skewed distribution and Wilcox and Keselman (2003) introduced a modified one-step *M*-estimator (*MOM*) as the central tendency measure when testing for treatment effects.

One might ask whether robust procedures are needed at all. What about using the two step approach; 1) clean the data by using some outlier rejection rule, and 2) use

classical method to test the remainder of the data. Will this suffice? Unfortunately, the process is not as simple as that. Huber (1981, p4) explained the reasons:

> "*...a. It is rarely possible to separate two steps cleanly; for instance, in multi parameter regression problems outliers are difficult to recognize unless we have reliable, robust estimates for the parameters.*
>
> *b. Even if the original batch of observations consists of normal observations interspersed with some gross errors, the cleaned data will not be normal (there will be statistical errors of both kinds, false rejections and false retentions), and the situation is even worse when the original batch was derived from a genuine non-normal distribution, instead of from a gross-error framework. Therefore the classical normal theory is not applicable to cleaned samples, and the actual performance of such a two step procedure may be more difficult to work out than that of a straight robust procedure.*
>
> *c. It is an empirical fact that the best rejection procedures do not quite reach the performance of the best robust procedures. The latter apparently are superior because*

they can make a smooth transition between full acceptance and full rejection of an observation..."

To illustrate the usefulness of a robust method, consider the following example. In many glass samples, the concentration of a given compound SiO_2 needed to be estimated. For this purpose, 4 to 5 measurements were taken from each sample. The usual concentration estimate is then the average of these 4 to 5 measurements. However, this estimate can be way off due to the presence of extreme values. For example, in one sample, the measured concentrations of SiO_2 were 68.52, 68.23, 67.42, 68.94, and 68.34 (in units of weight percent). The usual average of these five values is 68.29, whereas their median is 68.34. If the first measurement were wrongly recorded as 18.52 the average would become 58.29 (compared to 68.29), i.e. the average is strongly attracted by the extreme value. On the other hand, the median becomes 68.23 (compared to 68.34), indicating that the value is not much affected. This usefulness of a robust method will be much more appreciated if the measurement process is computerized (i.e. the measurement instrument sends the data directly to the computer, which process it without human intervention).

Among the latest procedures in robust statistics are S_1 (Babu et al., 1999) and *MOM-H* (Othman et al., 2004). These two procedures will be the main focus of this study, which will look at the problem of comparing central tendency measures for J groups with

$$H_0 : \theta_1 = \theta_2 = ... = \theta_J,$$

where θ_j is the central tendency parameter corresponding to distribution $F_j : j = 1, 2, \ldots, J$.

1.3 S_1 Statistic

Babu et al. (1999) proposed a more flexible statistical method dealing with asymmetric distributions and heteroscedastic settings. Known as the S_1 statistic, this method is one of the latest procedures in assessing the effects of a treatment variable across groups. S_1 uses median as the central tendency measure and is still the most widely known robust estimator of central tendency especially for skewed distribution. Characterized by the highest breakdown point (0.5), this estimator can withstand large proportions of very bad observations without breaking down completely. In a finite sample, the breakdown point of an estimator is the smallest proportion of the observations that must be replaced by arbitrary values in order to force the estimator to produce values arbitrarily far from the parameter values that generated the original data (Donoho and Huber, 1983). The finite sample breakdown point of an estimator is a measure of its resistance to contamination. The classical estimator and test of central tendency measures based on an ideal parametric model are often distorted by a few extreme values in the data set because they have lower breakdown points (see Huber, 1981). For example, if the *i*th observation among the observations $X_1, \ldots, X_n$ approach infinity, the sample mean gravitates to infinity as well. This means that the sample breakdown point of the sample mean is only $1/n$. The breakdown point of the population mean, μ, is 0, which is $1/n$ as n approaches infinity. The median has a breakdown point

of 50% (which is the highest possible), because the estimate remains bounded when fewer than 50% of the data points are replaced by arbitrary numbers.

When using the S_1 statistic, one can work with the original data without having to transform or trim the data to achieve symmetry. Simple transformations may fail to deal effectively with extreme values and heavy tailed distributions. Moreover, an ill-considered transformation can do more harm than good in terms of results interpretability. Even the popular strategy of taking logarithms of all the observations does not necessarily reduce problems due to the existence of extreme values (Wilcox, 1997).

1.4 *MOM-H* Statistic

One more strategy when dealing with extreme values is trimming. There are two approaches of trimming: (1) trimming a predetermined amount and then computing $\hat{\theta}$, (2) empirically determine the amount of trimming, trim, and then computing $\hat{\theta}$. Trimming needs to be done carefully to avoid the loss of information during the process. For example when sampling from a light tailed distribution, it might be desirable to trim very few observations, or if sampling is from a normal distribution, trimming might not be needed at all. For a right skewed distribution, a natural reaction is to trim more observations from the right versus the left tail of the empirical distribution. In essence, this is what the modified one-step *M*-estimator (*MOM*) does.

When using trimmed means, the strategy for reducing the effects of the tails of a distribution is by simply removing them based on the predetermined amount. By using this method of trimming, even observations from a normal distribution will be trimmed according to the predetermined amount such as 10% or 20% on both tails, whereas observations from a normal distribution need no trimming at all. Furthermore, the trimmed mean has a breakdown point just as much as the percentage of trimmings and this indicates that trimmed mean is not so robust, and cannot withstand large proportions of extreme values.

Wilcox and Keselman (2003) introduced the latest version of the *M*-estimator known as modified one-step *M*-estimator or *MOM*. The *MOM* estimator is calculated using data left from empirically determined trimming, and this estimator competes well with methods based on trimmed means in terms of both power and control over the probability of a Type I error. Like the sample median, the sample *MOM* estimator is a robust central tendency estimator that possesses highest breakdown point. Othman et al. (2004) used *MOM* as the central tendency measure in their work on the *H* statistic. Denoted as *MOM-H*, this statistic was shown to have the ability to control the Type I error at a nominal level in testing the equality of central tendency measures.

1.5 Scale Estimators

Based on Hall and Sheather's (1988) work on sample medians, Othman et al. (2004) noted that the standard errors of the sample medians in the S_1 statistic can be replaced by asymptotic variances. However their findings showed that the method does

not result in better Type I error control. Syed Yahaya et al. (2004a, 2004b) identified four scale estimators, namely Q_n, S_n, T_n and the popular robust scale estimator, MAD_n, with highest breakdown point and bounded influence function that were capable of maintaining the robustness of the S_1 statistic. Using these scale estimators to substitute the standard errors of the sample medians, $\hat{\omega}$, in the S_1 statistics, they observed that the combination of S_1 with these scale estimators produced good Type I error. The first three aforementioned scale estimators were introduced by Rousseeuw and Croux (1993) and each of them possess simple and explicit formula that guarantees uniqueness. In fact, the combination of the S_1 method with scale estimator T_n as examined by Syed Yahaya et al.(2004a, 2004b) showed it to be a very promising procedure in robust statistics.

Considering the good performance in controlling Type I error rates, these four scale estimators were selected for use on the S_1 statistic under a more skewed distribution in this study. These highly robust scale estimators; Q_n, S_n, T_n and MAD_n were also incorporated in the *MOM-H* procedure for the same purpose, i.e. to test the equality of central tendency measures. However, instead of being the scale estimators for the statistic, these scale estimators are used as the criterion for choosing sample values (trimming criterion) for *MOM*.

The combinations of the two robust test statistics with the highly robust scale estimators generated certain new procedures known as the modified S_1 and *MOM-H*. These modified procedures were then examined for their Type I error and power rates. With these modifications, new robust statistical procedures that can hopefully address the

issues of non-normality and variance heterogeneity when dealing with comparisons of central tendency measures are expected to be generated. Similar to results obtained in previous research on these two statistics, we have no doubt that this study will produce results that are as good if not better in terms of Type I error and power rates.

There is no perfect method in statistics and the same applies to robust methods. However, they offer a substantial improvement over standard techniques which have to comply with rigid assumptions. Besides, these methods (robust) can improve the performance of a test in terms of Type I error and power. For instance when comparing two or more groups, the significant differences reported in applied journals in all actuality may reflect a true difference, but many non-significant differences would have been significant if only the investigator had used a robust method.

1.6 Objective of the Study

The goal of this study is to search for alternative methods in testing for the equality of central tendency measures by simultaneously controlling Type I error and improving power rates in the one-way independent group design under skewed distributions. In achieving this goal, the following objectives need to be accomplished,

(a) To modify two of the latest methods in robust statistics i.e. S_1 and *MOM-H* by integrating some of the most robust scale estimators to these methods.

(b) To evaluate the modified methods using simulation data.

(c) To compare the modified with the original methods in terms of Type I error and power rates.

(d) To determine the best methods.

1.7 Significance of the Study

This study will contribute towards knowledge development in experimental design methodology especially in the experimental sciences. Statisticians are aware that experimental design methodology depends on assumptions of normality and treatment groups having equal variances. However, in the real world, data are not always normally distributed. The benefit of this research is that with these new alternative methods, researchers (in various fields, especially the experimental sciences) will not be constrained with all the assumptions such as normality and homogeneity of variances. They can instead work with the original data without having to worry about the shape of the distributions.

1.8 Organization of the Thesis

In this current chapter, we have reviewed the importance of robust methods and briefly introduced the latest methods in robust estimation, namely the S_1 and *MOM-H* statistics. The detail of these methods and the suggested scale estimators, and their underlying theories will be reviewed in Chapter 2. The definitions of some important terminologies such as Type I error, power of a test, breakdown point, influence function and percentile bootstrap will also be explained in Chapter 2 which would also review some of the past and contemporary research on robust statistics. Chapter 3 describes how the empirical investigation was conducted. The discussion in this chapter encompasses the selection of the conditions being investigated, followed by data

generation, with focus on the transformation of skewed distributions from the standard normal. There is a section on power analysis that focuses on the pattern of separation of the central tendency measures in the setting of the alternative hypotheses. The results from the analyses of Type I error and power are presented in Chapter 4. Finally, we end the thesis with conclusion and suggestions for further studies in Chapter 5.

CHAPTER 2
LITERATURE SURVEY

2.1 Introduction

In testing central tendency (location) measures for two or more groups, the classical methods such as the Student's two-sample *t*-test and the ANOVA are among the most commonly used statistical methods in the one-way independent group design. However, these methods are adversely affected by non-normality, particularly when variances are heterogenous and group sizes are unequal (Lix and Keselman, 1998). Violating the assumptions associated with these methods will cause the Type I error and power rates to be disrupted. The Type I error rates will be inflated from the nominal value and power rates can be substantially reduced from the theoretical value. These liberal values of Type I error rates will subsequently result in spurious rejections of the null hypotheses while low power rates will result in differences going undetected. Even though it is well established that the conventional ANOVA for comparing means is not robust if the homogeneity assumption does not hold (Wilcox et al., 1986), the *F*-test in ANOVA, for example, is often employed in statistical practice even when the data suggest that population variances are unequal (Kulinskaya et al., 2003). It is also well known that a slight departure from normality has a great effect on power for the methods mentioned (Sawilowsky and Blair, 1992; Wilcox, 1995). The sensitivity of *t*-test towards the violation of normality was illustrated by Wilcox (1995) on contaminated normal distribution where 10 percent of the observations came from a normal population having variance 10, rather than 1. When switching from standard normal to contaminated normal, the power is reduced from 0.975 to 0.28! However, consequences of the effects

of these violations for test statistics are hard to gauge, and are thus important issues that need further investigation.

In the effort to overcome the sensitivity of these procedures to the violations of the assumptions, researchers in this area have sought to find alternative methods. Cochran (1937) suggested weighting the terms in the sum of squares explained by the respective inverses of the sample variances, and he provided a chi-square test for equal means based on a transformation of the *F*-test for ANOVA. However in this case, the design has to be balanced. For unbalanced design, James (1951) and Welch (1951) had suggested weighting the terms in the sum of squares explained by estimates of the inverses of the variances of the respective sample means. This weighted sum of squares has an approximate chi-square distribution under the null hypotheses of equal population means for large sample sizes. Even if the problem of unequal variances could be overcome, the assumption of normality will always be associated with ANOVA. Even though ANOVA is known to be robust to small deviations from normality, to what extent can this method hold is unknown as there is no exact measurement of the violation or deviation from normality that we can base on, unless the sample size is big enough to guarantee the normality of sample means. This problem is common in the experimental sciences where measurements are typically repeated only a few times thus yielding small sample sizes, which of course will violate the normality assumption. Any violations, be it the non-normality or heterocedasticity, will always have some impact on the result of the ANOVA as well as the *t*-test.

In order to achieve a good test of a statistical hypothesis, Type I error rates need to be control at the nominal level while power rates need to be simultaneously inflated. As alternatives to the ANOVA or *t*-test, one can seek methods from the less powerful nonparametric statistics but these less powerful methods need large sample sizes to increase power. Centering on parametric models, but not entirely convinced that the assumption is true, robust statistical methods will give ways of finding practical solutions in statistics. With high speed computers, it is now possible to apply robust statistical methods that were heretofore impractical to use.

Robust statistical methods offer useful and viable alternatives to traditional analytic methods, often yielding greater statistical power and increased sensitivity. These methods were also proven to be able to control the Type I error rates at the nominal level (Keselman et al., 2002a; 2004b; Othman et al., 2004; Syed Yahaya et al., 2004a; 2004b; Wilcox et al., 2001; Wilcox et al., 1988).

Luh and Gou (1999) agreed that approximate tests are well known alternatives for dealing with the problem of heteroscedasticity. Nevertheless they noted that although these tests are known to be the most valid tests under various conditions of heteroscedasticity investigated (Wilcox, 1989), they could not simultaneously handle the problem of non-normality. They cited the case of the Welch test (Welch, 1951) and the James second-order test (James, 1951) which were though the most valid tests under various conditions of heterogeneity investigated were nevertheless affected by non-normality conditions (Keselman et al., 1995).

Babu et al. (1999) proposed the S_1 method that can handle the problems of non-normality and heteroscedasticity simultaneously in an adaptive test setting. This method needs no trimming or transforming and will be selected when the data are skewed.

Another way of dealing with skewed data is by trimming the tail of the distribution. Working with actual data, Wilcox et al. (2000) found that power can be greatly increased by comparing trimmed means versus means and control over the probability of a Type I error can be better. However, there are practical concerns regarding trimmed means. The utmost concern is that by assumption, the amount of trimming is fixed prior to analyzing the data. The next concern is that trimming is typically assumed to be symmetric. Given these concerns, the question is how can we determine the best percentage of trimmings especially when the distribution is skewed?

In dealing with the problem of predetermined amount of trimming, Wilcox et al. (2002) suggested modified one-step *M*-estimator (*MOM*) which addresses the problems with trimmed means. For example, if sampling is from a light-tailed distribution or normal distribution, it might be desirable to trim very few observations or perform no trimming at all. If the distribution is skewed, a natural reaction is to trim more observations from the skewed side of the empirical distribution. This central tendency estimator, like trimmed mean, can be applied to test statistics to investigate the equality of central tendency measures across treatment groups (Keselman et al., 2002; Othman et al., 2004). By using a statistic mentioned by Schrader and Hettsmansperger (1980),

as the probability of rejecting a true null hypothesis. Since it refers to the rate of rejecting a "true" null hypothesis, therefore, it should be of a relatively small value. The null hypothesis for testing the equality of central tendency measures is given as

$$H_0 : \theta_1 = \theta_2 = ... = \theta_j ,$$

where θ_i is the central tendency parameter for F_i: $i = 1, 2, ..., j$, and F_i is the distribution for group i.

Type I error rate is easy to access, because it involves calculating the proportion or percentage of significant statistical test (e.g. t's and F's) when the underlying population means are the same. When assumptions are met, the proportion of significance should come close to the set significance level.

In the context of hypothesis testing, the aspect of robustness is the ability of a procedure to control the Type I error rate of a test close to the nominal value (significance level), i.e. α, and stable over a range of distributions even with some deviations from its assumptions and Tiku et al. (1986) referred to this as "robustness of validity". Robust statisticians are looking for test procedures which are able to control the Type I error rates at the nominal value. By convention, a procedure can be considered robust if its Type I error is in between $0.5\alpha \leq \hat{\alpha} \leq 1.5\alpha$. For the nominal level $\alpha = 0.05$ the Type I error rate should be in between 0.025 and 0.075. Empirical Type I errors rates above 0.075 are considered liberal and those below 0.025 are considered conservative. However, Guo and Luh (2000) considered a test to be robust if

its empirical Type I error rate does not exceed 0.075 for the 5% level of significance used.

Type I error control is affected by the extreme conditions of non-normality and variance heterogeneity. In their investigation on the robustness of Student's *t*-test, Sawilowsky and Blair (1992) found that distributions with the extreme degree of skewness (e.g. $\gamma_1 = 1.64$) affected the Type I error control of the independent sample *t* statistic. Apart from these two problems, rates of Type I error can also be subjected to the unbalanced design, and even the pairings of unbalanced sample sizes with the unbalanced group variances. It is well known that the combination of larger variance with smaller sample size will disrupt the Type I error (Spector, 1993).

2.3 Power of a Statistical Test

The power of a statistical test is the probability of correctly rejecting a false null hypothesis, that is the probability that the test will conclude that the phenomenon exists (Cohen, 1988). Power is defined as $1-\beta$, where β is the Type II error probability. Type II error is the probability of failing to reject the null hypothesis when it needs to be rejected in favor of the alternate hypothesis. If the power of an experiment is low, then there is a good chance that the experiment will be inconclusive. Hence it is important to consider power in the design of experiments. However, in most of the work on robustness of tests, the level of the test (robustness of validity) was accorded prominence while power analysis, which is also known as "robustness of efficiency" in the robustness aspect, continued to be ignored until recently. As recent as 1997, a methodological study

has found that the power of statistical tests were not taken into account by researchers and that they continued to run a high risk of Type II error (Clark-Carter, 1997). Cohen (1988) has suggested that the neglect of power analysis exemplifies the slow movement of methodological advance. Neglect of power not only decreases the recognition of interesting effects (Type II error), but it also has a negative effect on the ability of researchers to establish statistical consensus through replication. Ottenbacher (1996, p274) points out that,

> "*...The apparently paradoxical conclusion is that the more often we are well guided by theory and prior observation, but conduct a low power study, the more we decrease the probability of replication... The responsible investigator must be concerned with statistical power. A concern with power, however, cannot end with its calculation. Because the ability to detect treatments must be optimized, the responsible scientist must also be concerned with factors that determine effect size...*"

Most studies on power deal with the calculation of power for parametric statistics where normal theory assumptions are required, for example, the *t*-test and *F*-tests. The calculations of power for robust statistics or nonstandard nonparametric statistics are not addressed at a practical level. For example, the most sought after tome on power by Cohen (1988) concentrates mainly on ANOVA and regression models and some standard nonparametric tests such as the chi-square test. What is not addressed is how violations of normality assumptions affect power estimates. Our study on power focused on the violations of normality assumptions by assuming that the factors affecting power are kept

constant except for the effect size. The power of a statistical test depends upon three parameters: i) the significance criterion, ii) the sample size, and iii) the effect size (Cohen, 1988).

The significance criterion

The significance criterion represents the standard of proof that the phenomenon exists, or the risk of mistakenly rejecting the null hypothesis. Denoted by α, it is known as the Type I error rate. The more conservative the significance level, the lower the power. Thus, using the .01 level will result in lower power than using the .05 level.

The directionality of the significance criterion also gives some impact to the power of a statistical test (Cohen, 1988). When no direction is specified, the resulting test will have less power than the test with the same α value which is directional, as long as the effect is in the expected direction.

The sample size

The reliability of sample result is always dependent upon the size of the sample. All things being equal, the larger the sample size, the greater the reliability or precision of the results, thus the greater the probability of detecting a non-null state of affairs, that is, the phenomenon under test can manifest itself more clearly against the background of variability. By increasing the sample size, the statistical power will increase.

The effect size

Cohen (1988) defined "effect size" as the degree to which the phenomenon is present in the population, or the degree to which the null hypothesis is false in relation to the alternate hypothesis. The null hypothesis always means that the effect size is zero. Specifically, in using the *t*-test for two independent groups, effect size is simply the difference between the two averages divided by the standard deviation i.e. the standardized mean difference. In the *F*-test for two or more population means, the "effect size" is the standard deviation of standardized means. Effect size measures provide a standardized index of how much impact treatments actually have on the dependent variable. Conventionally, the measures of effect size can be categorized into small, medium, and large effects depending on the on the area of research. The values are arbitrary, but the conventional definitions of effect size by Cohen (1988) are given in Table 2.1:

Table 2.1: Conventional effect size values by Cohen (1988)

# of Groups / Effect size	2 Groups	≥ 2 Groups
Small	0.20	0.10
Medium	0.50	0.25
Large	0.80	0.40

Several other factors such as variance and population distribution can affect power. Increasing the variance will lower the power of a statistical test. A homogenous population reduces the variance thus increasing power. With regard to population distribution, deviations from the assumption of normality usually lower the power. The