MODIFIED WILCOXON PROCEDURE FOR DEPENDENT GROUP

NOR AISHAH AHAD

SUHAIDA ABDULLAH

ZAHAYU MD YUSOF

SHARIPAH SOAAD SYED YAHAYA

LIM YAI FUNG

UNIVERSITI UTARA MALAYSIA

2014

DISCLAIMER

We are responsible for the accuracy of all opinion, technical comment, factual report, data, figure, illustration and photographs in the article. We bear full responsibility for the checking whether material submitted is subjected to copyright or ownership right. UUM does not accept any liability for the accuracy of such comment, report and other technical and factual information and the copyright or ownership right claims.

CHIEF RESEARCHER:

_____
NOR AISHAH AHAD

MEMBER:

_____
SUHAIDA ABDULLAH

_____
ZAHAYU MD YUSOF

_____
SHARIPAH SOAAD SYED YAHAYA

_____
LIM YAI FUNG

ACKNOWLEDGEMENT

First and foremost, we would like to extent our sincere thanks to Universiti Utara Malaysia for the financial support under the LEADS Research Grant Scheme and to RIMC for facilitating the management of the research. Our grateful recognition are due to our family, fellow colleagues and friends who had offered guidance and ideas as well as encouragement that has definitely push us to complete these research.

ABSTRACT

Nonparametric methods require no or very limited assumptions to be made about the format of the data, and they may therefore be preferable when the assumptions required for parametric methods are not valid. The Wilcoxon signed rank test applies to matched pairs studies. For two tail test, it tests the null hypothesis that there is no systematic difference within pairs against alternatives that assert a systematic difference. The test is based on the Wilcoxon signed rank statistic *W,* which is the smaller of the two ranks sums. The step to compute the statistic *W* considered positive and negative differences and omit all the zero differences. In this study, we modify the Wilcoxon signed rank test using the indicator function of positive, zero and negative differences to compute the Wilcoxon statistic, *W*. The empirical Type I error rates of the modified statistical test was measured via Monte Carlo simulation. These rates were obtained under different distributional shapes, sample sizes, and number of replications. The modified Wilcoxon signed rank test was found to be robust under symmetric distributions. The result shows that this test produced liberal Type I error rates under skewed distribution. The use of the indicator positive, zero and negative differences influence the result of the Wilcoxon statistic. These finding was demonstrated using an example data.

.

Keywords: nonparametric, Type I error rate, Wilcoxon signed rank test.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ARE	Asymptotic relative efficiency
MD	Median difference

CHAPTER 1

INTRODUCTION

1.1 Research Background

Drawing conclusions and inferences through statistical hypothesis testing about the differences between two groups is one of the routinely employed processes in educational, behavioral or social research. Group comparisons are at the heart of many research questions addressed by the researchers. Here are some examples of the research questions.

1) Do males and females differ in terms of their exam scores?
2) Is a particular curriculum effective in improving students' achievement?

To answer these questions, researchers typically turn to a time-honored procedure like the independent samples *t*-test. The independent samples *t*-test is based on certain assumptions namely (a) samples are independent and randomly selected, (b) population distributions are normal, and (c) population variances are equal. In order for the test statistic to provide valid results leading to sound and reliable conclusions, this requirement must be satisfied.

In real life, these ideal data are no doubt hard to access. Fortunately, if the assumption of normality is not satisfied, researchers can choose alternative procedures from the nonparametric methods. Wilcoxon rank sum test and Wilcoxon signed rank test can be used for independent and dependent samples respectively. Dependent or paired data are numerical data obtained from two populations that are related, that is, when results of the first group are not independent of the results of the second group. This dependency characteristic of the two groups occurs either because the items or individuals are paired or matched according to some characteristic or because repeated measurements are obtained from the same set of items or individuals. In either case, the variable of interest becomes the difference between the values of the observations rather than the values of the observations themselves.

If the difference scores are assumed to be randomly drawn from a poulation that is normally distributed and sample size is not very small, paired sample *t*-test can be used to determine whether there is a significant population mean difference. The paired *t*-test assumes that the data are measured on interval or a ratio scale. When the paired *t*-test is not suitable due to the violation of the normality assumption, the nonparametric Wilcoxon signed rank test for the median difference can be used. The Wilcoxon signed rank test requires less stringent assumptions, such that the difference score come from a distribution that is approximately symmetric and the data are measured on an ordinal, interval, or ratio scale. When the assumptions for the Wilcoxon signed rank procedure are met, but the assumptions of the paired *t*-test are violated, the Wilcoxon procedure is likely to be the more powerful in detecting the existence of significant differences.

Under ideal condition, where all the assumptions for using *t*-test are fulfilled, the Wilcoxon signed ranks test is almost as powerful.

The Wilcoxon signed ranks test uses the test statistic *W*. The computation of this statistic does not involve raw observations in the two dependent groups, but used the differences between them. For each item in a sample, the absolute difference between the paired values are arranged in increasing order and assign ranks, such that the smallest absolute difference score gets rank 1 and the largest gets the highest rank. Keeping track of which values were originally positive and negative. For tied values, the average of their ranks was computed while for the difference with zero values, it was discarded before ranking. These zero values are not considered in the calculation of Wilcoxon statistic. Lastly, compute the Wilcoxon test statistic, *W*, which is the smaller of the two ranks sums.

In the case of independent sample, there was several studies developed new approach of nonparametric tests. Study by Steland, Padmanabhan and Akram (2011) used pseudo-medians of distribution as a location parameter and applied bootstrap method in testing the differences between groups for the nonparametric Behrens-Fisher problem and the Generalized Behrens-Fisher problem. Ahad, Othman and Syed Yahaya (2013; 2012; 2011) modified the one-sample nonparametric Wilcoxon procedure and employed pseudo-median of differences between group values as the central measure of location in a two independent groups setting. In their study, they considered positive differences, differences equal to zero and negative differences in computing the Wilcoxon statistic. However, none of these approaches take a look into the case of dependent sample.

Therefore, in this study, we focusing on the case of dependent sample (paired data) by modifying the original Wilcoxon signed rank test with the application of the same indicator function where we considered positive, zero and negative differences in calculating the Wilcoxon statistic, *W*. The Type I error rate was measured in order to evaluate the performance of the modified Wilcoxon signed rank test.

1.2 Research Objectives

The primary goal of this study is to evaluate the performance of the modified Wilcoxon signed rank test in controlling the Type I error rates by taking into consideration the positive, zero and negative differences in calculating the Wilcoxon statistic. To achieve this goal, the performance of the modified Wilcoxon signed rank test was measured in terms of Type I error rates.

1.3 Significance of the Study

The contribution of this study is towards the knowledge development in the nonparametric statistics. Original step in Wilcoxon signed rank test will omit any absolute difference score of zero from further analysis and just considered positive and negative differences. However, in this study, we modified the Wilcoxon signed rank test where we considered positive, zero and negative differences in calculating *W*.

1.4 Organization of the Report

In this current chapter, we have provided an introduction of the study which included the scenario of testing the two paired samples, the problem that arise when the assumptions of normality is violated, the alternative test in the nonparametric methods, the objective and the significance of the study. Chapter 2 reviewed the work related to Wilcoxon signed rank test and gives definitions of some important terminologies such as Type I error and robustness. Chapter 3 proposes the research method and testing framework in this study. This chapter also outlines how the empirical investigation was conducted and discuss on the study condition being investigated, followed by the procedure of generating and manipulating selected distributions. The results and discussion of the Monte Carlo simulation study of the modified Wilcoxon signed rank test are presented in Chapter 4. Finally, Chapter 5 summarizes the findings and discusses the strength and implication of the tests. We ended the report with conclusion and recommendations for further studies.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Many parametric statistical methods require assumptions to be made about the format of the data to be analyzed. One of the underlying assumptions of parametric tests used in hypothesis testing is that the populations from which the data are sampled are normal in shape. Very often the variables within data sets from education and psychology are not normally distributed (Cressie & Whitford, 1986; Micceri, 1989). In his study, Micceri (1989) surveyed 440 data sets from psychological and education sources and determined that virtually none of the data sets could be adequately characterized by a normal distribution. Micceri (1989) described the distributions he examined as having varying degrees of multimodality, asymmetry (skew), and excessive tail weight (kurtosis). Although it may be convenient (practically and statistically) for researchers to assume that their samples are obtained from normal populations, this assumption may rarely be accurate (Micceri, 1989; Wilcox, 1990). For example, the paired *t*-test requires that the distribution of the differences be approximately normal. Fortunately, this assumption is

often valid in real data, or the other alternative is to apply suitable transformation. However, there are situations where even transformed data may not satisfy the assumptions. For such case, it may be inappropriate to use traditional (parametric) methods of analysis.

Nonparametric methods provide an alternative series of statistical methods that require very limited assumptions to be made about the data. There is a wide range of methods that can be used in different circumstances. The nonparametric alternative to the paired *t*-test is the Wilcoxon signed rank test. For situations involving either matched items or repeated measurements of the same item, the nonparametric Wilcoxon signed rank test for the median difference can be used when the paired *t*-test for the mean difference is not appropriate due to the violation of the assumptions. The paired *t*-test assumes that the data are measured on an interval or a ratio scale and are normally distributed. The Wilcoxon signed rank test only requires that the differences are approximately symmetric and that the data are at least measured on an ordinal. The Wilcoxon signed ranks test is a more powerful method than *t*-test when the assumptions for *t*-test are violated, and it is as powerful as *t*-test when the assumptions are fulfilled.

Before going in depth into the discussion of the method, the following sections will give a brief explanation on the statistical tests used and the definitions on a few important terminologies that were being used throughout this study.

2.2 Nonparametric Test

A statistical procedure that has certain desirable properties that hold under relatively mild assumptions regarding the underlying population from which the data were obtained is known as a nonparametric procedure. According to Savage (1953), papers related to nonparametric statistics appeared as early as the nineteenth century, however, he designated 1936 as the true beginning of the nonparametric test, because in this year Hotelling and Pabst (1936) published their paper on rank correlation. Hettmansperger, Mckean and Sheather (2000), stated in their paper that the earliest work in nonparametric was done in 1936 by Hotelling and Pabst on rank correlation and followed by Friedman on rank tests in a two-way design a year later.

In testing the equality of two groups, Wilcoxon (1945) introduced the signed rank and rank sum tests while Mann and Whitney (1947) extended his ideas to the case of more than two groups. The development of nonparametric continued after that when Pitman developed efficiency concepts in 1948. During 1950s and 1960s, Lehmann in association with Hodges and his students showed that rank tests are surprisingly efficient and robust. Since then, many articles and books were produced and published by well-known researchers such as Fisher, Scheffe', Wilcoxon, Wolfowitz, Kendall, Siegel, Savage, Hajek, Sidak and many more. Nonparametric statistics continue to flourish today.

According to Hollander and Wolfe (1999), there are many reasons that contribute to the rapid development of nonparametric statistical procedures. Among the reasons are:

a) Nonparametric methods require few assumptions about the underlying populations from which the data are obtained. It forgoes the traditional assumption that the underlying populations are normal.
b) Nonparametric techniques are often (although not always) easier to apply than their normal theory counterparts.
c) Nonparametric procedures are often quite easy to understand.
d) Nonparametric procedures are applicable in situations where the normal theory procedure cannot be utilized. For example, many of the procedures require not the actual magnitudes of the observations, but rather, their ranks.
e) Nonparametric procedures even though are slightly less efficient than their normal theory competitors when the underlying populations are normal, they can be mildly more efficient compared to their competitors when the underlying populations are not normal.

There are many advantages of using nonparametric techniques. Siegel (1956) outlined six main advantages as follows:

a) For most nonparametric statistics, the "accuracy" of the probability statement does not depend on the shape of the population.
b) The size of the sample is not as important, because small sample sizes will not cause the results to be misleading to the extent that small samples unduly affect parametric tests.
c) Nonparametric statistics can be used when observations come from several different distributions.

d) Nonparametric statistics can be used with data that are ordinal, or ranked, as well as with interval and ratio scaled data.
e) Nonparametric statistics can be used with nominal data as well.
f) Nonparametric statistics can be easily learned and applied, at least at the univariate level.

McSeeney and Katz (1978) summarized the reasons for using nonparametric statistics. These include (a) nonparametric statistics have fewer assumptions, (b) nonparametric statistics can be used with rank-ordered data, (c) nonparametric statistics can be used with small samples, (d) data do not need to be normally distributed, and (e) outliers can be present. Conover (1980) provided three reasons for using nonparametric statistics. Specifically, he agreed that nonparametric methods (a) involve less computational work, (b) easier and quicker to apply, and (c) much of the theory behind the nonparametric methods may be developed rigorously, using no mathematics beyond high school algebra.

Furthermore, when approximate normality is met, nonparametric tests are still relatively efficient, the asymptotic relative efficiency (ARE) of nonparametric tests with respect to parametric tests can be as high as 95.5% (Gibbons, 1993; Hollander & Wolfe, 1999). ARE in simple term can be defined as a measure of the large-sample efficiency of one test relative to the other (Higgins, 2004). Consequently, in many cases, researchers have relatively little to lose by using nonparametric tests if the distribution is normal. If the distribution is not normal, tests based on nonparametric tests are likely to be more efficient than their parametric counterparts.

2.3 Wilcoxon Signed Rank Test

Frank Wilcoxon stated in his paper in 1945 that the comparison of two treatments generally falls into one of the following two categories: a) we may have a number of replications for each of the two treatments, which are unpaired, or b) we may have a number of paired comparison leading to a series of differences, some which may be positive and some negative. Wilcoxon (1945) introduced the rank sum tests for unpaired group and signed rank test for paired group which are still named after him.

To perform the Wilcoxon signed rank test for the median difference, below are the steps to obtain the test statistic *W*.

1. For each item in a sample of n items, compute a difference score, D_i, between the two paired values.
2. Neglect the + and - signs and list the set of n absolute differences, $|D_i|$.
3. Omit any absolute difference score of zero from further analysis, thereby yielding a set of n' nonzero absolute difference scores, where $n' \leq n$. After remove values with absolute difference scores of zero, n' becomes the actual sample size.
4. Assign ranks R_i from 1 to n' to each of the $|D_i|$ such that the smallest absolute difference score gets rank 1 and the largest gets rank n'. If two or more are equal, assign each of them the average of the ranks they would have been assigned individually had ties in the data not occurred.
5. Reassign the symbol + or - to each of the n' ranks, R_i, depending on whether D_i was originally positive or negative.

6. Compute the sum of the positive ranks and the sum of the negative ranks. The smaller of the two rank sums is used as the test statistic, *W* as shown in Equation 1.

$$W = \sum_{i=1}^{n'} R_i\,(+) \; or \; W = \sum_{i=1}^{n'} R_i\,(-) \qquad (1)$$

Because the sum of the first n' integers $(1,2,\ldots,n')$ is given by $n'(n'-1)/2$, the Wilcoxon test statistic *W* ranges from a minimum of 0 (where all the observed difference scores are negative) to a maximum of $n'(n'-1)/2$ (where all the observed difference scores are positive). If the null hypothesis is true, the test statistic *W* is expected to take on a value close to its mean $\mu_W = n'(n'-1)/4$. If the null hypothesis is false, the observed value of the test statistic is expected to be close to one of the extremes. The two-tail test of the null hypothesis that the population median difference *MD* is zero can be written as Equation 2.

$$H_0: M_D = 0 \qquad H_1: M_D \neq 0 \qquad (2)$$

2.4 The Error of Statistical Test

Statistical test involves two types of errors which are (1) Type I error (a probability of the true null hypothesis is incorrectly rejected) denoted as α and (2) Type II error (a probability false null hypothesis fail to be rejected) denoted as β. Definition of these two errors can be presented as in Table 2.1. Type I error is the error of rejecting a null hypothesis when it is actually true. This error is also known as significant level, nominal level, “error of the first kind” or “false positive”. In other words, this is the error when we

are observing a difference when in truth there is none, thus indicating a test of poor specificity. Type I error can be viewed as the error of excessive credulity. Conventionally, researchers have chosen either the 0.05 level or the 0.01 level (5% or 1% level) of significance, although the choice of levels is largely subjective. When the significance level is low, more data must be diverging from the null hypothesis to be significant. Therefore, 0.01 level is more conservative compare to the 0.05 level.

Type II error is the error of failing to reject a null hypothesis when it is in fact not true. This error is also known as "error of the second kind", β error, or "false negative". It happens when we fail to observe a difference when in truth there is one, thus indicating a test of poor sensitivity. This error can be viewed as the error of excessive skepticism. Based on their definitions, Type I errors are generally considered more serious than Type II errors.

In many practical applications, Type I errors are more delicate than Type II errors. More attention is given on minimizing the occurrence of this statistical error. Therefore a good statistical test is defined by evaluating how good the test in controlling the occurrence of the probability of Type I error or the Type I error rates. The test with good control of the Type I error rates identified as a robust test.

Table 2.1: Type of errors

Statistical Decision	True State of the Null Hypothesis	
	H_0 True	H_0 False
Reject H_0	Type I error	Correct Decision
Do not Reject H_0	Correct Decision	Type II error

2.5 Robustness

Robustness signifies insensitivity to small deviations from the assumptions (Huber, 2004). According to Sullivan and D`Agostino (1996) and Heeren and D`Agostino (1987), statistical tests are said to be robust if the observed Type I error rates are close to the pre-selected or nominal significance value in the presence of violations of assumptions.

Robustness in the context of hypothesis testing is the ability of a procedure to control the Type I error rate of a test close to the nominal value (significance level) and stable over a range of distributions even with some deviations from its assumptions. Tiku, Tan and Balakrishnan (1986) refer the phenomenon as "robustness of validity" when the Type I error of a test procedure is stable from distribution to distribution; at any rate the Type I error for plausible alternatives to normality is never too large compared to its normal-theory.

In order to provide a quantitative definition of robustness, we have to state for a given α value, the range of the true probability p of a Type I error for which the test would be regarded as robust. Bradley (1978), recommended that a procedure could be considered robust to the violation of an assumption if the Type I error rate is within $\pm 0.5\alpha$. Bradley proposed a 'liberal criterion' by defining robustness as $0.5\alpha \leq p \leq 1.5\alpha$. Thus, when the nominal level is set at $\alpha = 0.05$, the procedure or test is considered robust if its Type I error rate is in between 0.025 and 0.075. If the Type I error was not contained in this interval, then a test procedure was considered non robust for that particular condition. Type I error rates above 0.075 are considered liberal and those below 0.025 are considered conservative. A conservative test is of less concern and at

times considered robust. This situation is different for liberal tests as it is viewed with caution for it has committed a critical Type I error. This research used Bradley's liberal criterion of robustness to measure the performance of the proposed statistical test in controlling its Type I error.

CHAPTER 3

METHODOLOGY

3.1 Introduction

The Wilcoxon signed rank test is a nonparametric procedure which is suitable to be applied to test two dependent groups (matched pairs). It tests the null hypothesis that there is no systematic difference within pairs against alternatives that assert a systematic difference. The test is based on the Wilcoxon signed rank statistic *W,* which is the smaller of the two rank sums. The step to compute the statistic *W* considered positive and negative differences and omit all the zero differences. Recent works by Ahad et al. (2013; 2012; 2011) and Steland et al. (2011) considered indicator function of positive, zero and negative differences to compute the Wilcoxon statistic, *W* for two independent groups. In this study, the same idea of using the indicator function of positive, zero and negative differences to compute the Wilcoxon statistic, *W* for dependent group will be employed.

In order to accomplish the objective of the study which is to investigate the performance of the modified Wilcoxon signed rank test in controlling the Type I error rate, a few variables have been manipulated to create conditions which may be able to

highlight the strengths and weaknesses of the method. The variables include total sample sizes, types of distributions, and number of simulation.

A simulation study was conducted to evaluate and compare the performance of the method for each of the conditions investigated. The robustness of the method or test was determined by performing the test under constant conditions a large number of times. The frequencies of rejecting the null hypotheses were recorded. The proportion rejected (frequencies of rejecting the null hypotheses divide by the number of simulations) was used as an estimate of the probability of committing the Type I error.

3.2 Procedure Employed

The procedure employed in this study was the modification of the Wilcoxon signed rank test with the inclusion of the indicator function zero differences to obtain the Wilcoxon statistic, *W*. The two-tail test of the population median difference, *MD* is given by Equation 3.

$$H_0: M_D = 0 \qquad H_1: M_D \neq 0 \tag{3}$$

Generate two sequences of uncorrelated X_1 and X_2 using specified distribution with equal sizes. Define a new sequence as Equation 4 to get a correlated or paired data (Thijs van den Berg, 2013). This new Y_1 sequence will have a correlation of ρ with the X_1 sequence. The value of ρ is set as 0.8.

$$Y_1 = \rho X_1 + X_2\sqrt{1-\rho^2} \tag{4}$$

Find sequence difference between X_1 and Y_1 where

$$D_i = X_{1i} - Y_{1i} \tag{5}$$

where $i = 1,2, \ldots, n$. Let $|D_i|$ denotes the absolute value of D_i, and R_i denotes the rank of $|D_i|$. Define the indicator function as

$$e_i = \begin{cases} 1 & if\ D_i > 0 \\ 0.5 & if\ D_i = 0 \\ 0 & if\ D_i < 0 \end{cases} \tag{6}$$

Based on Equation 6, determine e_i with regards to the differences, D_i. Then the Wilcoxon statistic is defined as

$$W = \sum_{i=1}^{n} R_i e_i \tag{7}$$

For a two-tail test and for a particular level of significance, if the observed value of W equals or is greater than the upper critical value or is equal to or less than the lower critical value in the Wilcoxon table, the null hypothesis is rejected.

3.3 Variables Manipulated

A few conditions that were identified to have effect on the robustness of test for paired group were considered. These conditions were created by manipulating a few variables namely sample sizes, distributional shapes and simulation number. The purpose was to highlight the strength and weaknesses of the method in the aspect of robustness. Discussion on the variables manipulated is given in the next subsection.

3.3.1 Sample Sizes

This study focused on paired groups with small sample sizes. To examine the effect of sample size on the test, we manipulated the sample size (n) to be 10, 15, 20, 25 and 30. We focus on small sample sizes since we want to accommodate the critical values provided by the Wilcoxon table when making decision whether to reject or not to reject the null hypothesis. For a large sample, the test statistic W is approximately normally distributed.

3.3.2 Types of Distributions

The next variable of interest was the population distributional shape. We chose to employ various distribution from both types of symmetrical and nonsymmetrical distributions to study the effects of distributional shape on Type I error for the procedures investigated.

The shape of a distribution is usually depicted by skewness and kurtosis. Skewness is a departure from symmetry (Hoaglin, 1985a). In other words, skewness is a measure of symmetry, or more precisely, the lack of symmetry. The coefficient γ_1 is used to measure the skewness. γ_1 provides an indication of departure from symmetry in a distribution and its value can be positive or negative. Generally, a distribution is symmetric if the median divides the left side and the right side into two identical areas. A symmetric distribution has a skewness value of zero $(\gamma_1 = 0)$. Left skewed occurs when the left tail is longer with regard to the right tail, i.e. $\gamma_1 < 0$ and right skewed occurs

when the right tail is longer with regard to the left tail, i.e. $\gamma_1 > 0$. Refer to Figure 3.1 for illustration.

Figure 3.1. Type of skewness

Kurtosis, on the other hand, is a measure of whether the data are peaked or flat relative to a normal distribution. In a simple definition, kurtosis also refers to heavier tails (Hoaglin, 1985a). The coefficient γ_2 is used to measure the kurtosis. High kurtosis is usually represented by a distinct peak near the mean, which subside rather rapidly, and have heavy tails. While low kurtosis tends to have a flat top near the mean. Zero kurtosis $(\gamma_2 = 0)$ indicates normal tail or mesokurtic distribution. Short-tailed distributions have $\gamma_2 < 0$, and long-tailed distributions have $\gamma_2 > 0$ (Algina, Keselman & Penfield, 2005). Leptokurtic distributions have a positive kurtosis while platykurtic distributions have a negative kurtosis. According to Miles and Shevlin (2001), the term 'leptokurtic' is originally from the Greek word 'leptos', meaning small or slender. On the other hand, the word 'platykurtic' comes from the French word 'plat', meaning flat. In other word, positive kurtosis indicates a "peaked" distribution while negative kurtosis indicates a "flat" distribution. A general form of kurtosis is illustrated in Figure 3.2.

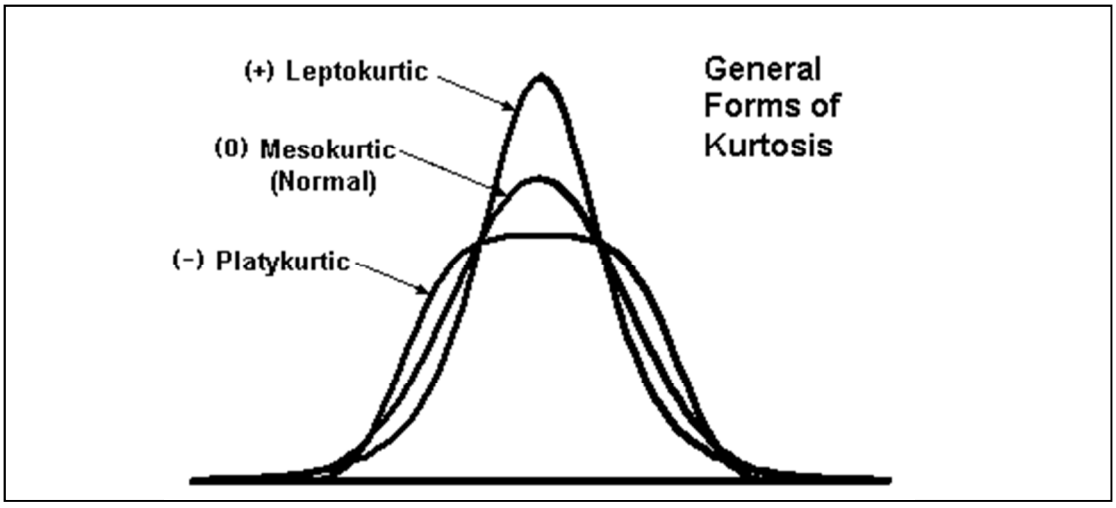

Figure 3.2. General forms of kurtosis

To observe the effect of distributional shapes on Type I error of the modified Wilcoxon signed rank test, this study focused on four distributions representing different degrees of skewness and kurtosis from both spectrum of symmetric and asymmetric distribution. For symmetric distributions, the distributions used in this study were standard normal, Beta (0.5, 0.5) and the *g*-and-*h* distribution from Hoaglin (1985b) with $g = 0$ and $h = 0.225$, representing symmetric mesokurtic, platykurtic and leptokurtic, respectively.

The normal distribution is probably the most important distribution in statistics and it represents distribution with zero skewness. Normal distribution was used as the basis of comparison. Beta in general is an asymmetric distribution. But one advantage of the Beta distribution is that it can take on many different shapes. We can set the Beta's parameters in order to generate a desired distribution. Beta (0.5, 0.5) is a platykurtic

distribution and has a *u*-shape. The *g*-and-*h* distribution was obtained from the transformation of the normal distribution to skewed or longer tailed by controlling the *g* and *h* parameters. The parameter *g* controlled the amount of skewness, while parameter *h* controlled the kurtosis. The tails of the distribution were further skewed as *g* increased and became heavier as *h* increased. Meanwhile, for asymmetric distributions, the chi-square distribution with three degrees of freedom $\left(\chi_3^2\right)$ was chosen to represent skewed leptokurtic distribution. Table 3.1 shows the types of symmetrical and nonsymmetrical distributions used in this study together with their levels of skewness and kurtosis.

Table 3.1: Distributions used in the study

Distributional Shape		Distribution Identified	Skewness	Kurtosis
Symmetrical	Platykurtic	Beta(0.5,0.5)	0	-1.5
	Normal tail	Normal(0,1)	0	0
	Leptokurtic	$g = 0$, $h = 0.225$	0	154.84
Asymmetrical	Leptokurtic	Chi-square(3)	1.63	4.00

3.3.3 Number of Simulation

The last variable manipulated was number of simulation. In this study, we used 1000, 5000 and 10,000 replications for each distribution with each study condition. These three different replication sizes have been used by Kang and Harring (2012) in their simulation

study. According to the literature, there are various numbers of simulations being used by previous researchers. For example, Othman, Padmanabhan and Keselman (2003) used one thousand replications of each condition when they extended the Mann-Whitney procedure to *J*-samples. The same number of simulations also been used in Wilcox, Keselman and Kowalchuk (1998) and Keselman, Wilcox, Taylor and Kowalchuk (2000). Greater numbers of simulations like ten thousands were used when sampling distribution was really intractable to derive analytically as done in Keselman, Wilcox, Lix, Algina and Fradette (2007) and Guo and Luh (2000).

However, the number of simulations frequently used is five thousand. Syed Yahaya, Othman and Keselman (2006) generated five thousand data sets in examining the Type I error rates of their modified robust statistical procedure and when comparing the typical score across independent groups based on different criteria for trimming. Othman, Keselman, Padmanabhan, Wilcox and Fradette (2004) also used five thousands replications of each study condition when comparing a number of adaptive robust methods with respect to their ability to control Type I error. Other researchers that used this number of simulations in their study were Keselman, Othman and Wilcox (2013), Keselman, Wilcox, Othman and Fradette (2002), Keselman, Othman, Wilcox and Fradette (2004) and Hess, Olejnik and Huberty (2001).

3.4 Design Specification

Table 3.2 shows the design specifications from the combination of sample sizes, types of distribution and number of simulation. The test conditions simulated for this study

involved the association of the five sample sizes paired with the three simulation number which produced fifteen conditions. These conditions were then applied on the four suggested distributions. Overall, these specifications produced 60 conditions in total. Type I error rates were computed within each condition examined. The results of the analysis and discussion in Chapter 4 were based on these design specifications.

Table 3.2: Design specifications and test conditions of the study

Distribution	Sample Sizes	Number of Simulation
Normal Beta (0.5,0.5) $g = 0, h = 0.225$ χ_2^3	10 15 20 25 30	1000 5000 10000

3.5 Data Generation

This study was based on simulated data. The simulation of data according to types of distribution was the key step in the analytical and empirical computational studies of the test procedure. The simulation was carried out using random-number-generating function in SAS and the simulation program was written in SAS/IML (SAS, 2006). The data were generated by using the following steps for each condition:

1. Generate two groups of n observations from the target population where n is the sample size (n = 10, 15, 20, 25, 30).

2. Standardize the distributions using the population expected value and standard deviation.

3. Generate a paired group that use both groups in Step 1.

In terms of the data generation procedure, pseudo-random variates for each particular distributional shape were obtained in the following manner:

a) Standard normal distribution

Pseudo-random normal variates were generated by employing the SAS generator RANDGEN (SAS, 2006). This involved the straight forward usage of the (RANDGEN(Y, 'NORMAL')) to generate normal variates with means equals to zero and standard deviation equals to one.

b) Beta (0.5,0.5) distribution

Data for Beta (0.5,0.5) distribution was generated using the RANDGEN subroutine with the beta distribution option, (RANDGEN(Y, 'BETA',0.5,0.5)). Beta (0.5,0.5) is a symmetric *u*-shaped distribution, hence the negative kurtosis.

c) *g*-and-*h* distribution with $g = 0$ and $h = 0.225$

To generate data from a *g*- and *h*- distribution, standard normal variates Z_{ij} were generated using (a). Transform the standard normal variates to *g*- and *h*- variates via Equation 8 to obtain the symmetric leptokurtic distribution.

$$Y = Ze^{\frac{hZ^2}{2}} \qquad (8)$$

d) Chi-square distribution with three degrees of freedom

To generate the chi-square variates with three degrees of freedom, we used the straight forward SAS/IML function i.e. (RANDGEN(Y, 'CHISQUARE', 3)).

Normal distribution already have variances equal to one, therefore standardization was not required. Observations generated from the Beta (0.5,0.5), *g*-and-*h* and χ_3^2 distributions, where the variances were not equal to one, were standardized so that they were one. The standardization for each of these distributions was done using the equation below: .

$$Y_i = (X_i - Theoretical\ Mean)/\sqrt{Theoretical\ Variance} \qquad (9)$$

Table 3.3 gives the theoretical mean and variance for Beta (0.5,0.5), *g*-and-*h* and χ_3^2 distributions.

Table 3.3: Theoretical mean and variance

Distribution	Theoretical Mean	Theoretical variance
Beta (0.5,0.5)	$\frac{A}{(A+B)} = 0.5$	$\frac{AB}{\left((A+B)^2(A+B+1)\right)} = 0.125$
$g = 0, h = 0.225$	0	$\frac{1}{(1-2h)^{\frac{3}{2}}} = 2.4516$
χ_3^2	$v = 3$	$2v = 6$

3.6 Monte Carlo Simulation

Monte Carlo simulations are computer experiments involving random sampling from known probability distributions to study properties of statistical methods (Mooney, 1997). By generating data under a variety of model-specific and distributional misspecification, one can monitor statistics of interest in order to understand their behavior across varying conditions (e.g., severity of non-normal distribution conditions, sample sizes). Monte Carlo simulation is a class of computational algorithms that rely on repeated random sampling to compute their results. Monte Carlo simulation is usually used to estimate the Type I error in situations where the assumptions of the test are violated or when analytical approach are not available. The next section will discuss on the Monte Carlo simulation used on the statistical test investigated to assess the Type I error rates.

3.7 Monte Carlo Assessment of Type I Error

The simulation of data according to the systematic manipulation of the group sizes and the underlying shape of the distribution was used to compute the Type I error for the modified Wilcoxon signed rank test. The number of simulations or replications used in this study were 1000, 5000 and 10,000. The algorithm of the modified Wilcoxon signed rank test for estimating the Type I error is as follows:

1. Initialize a variable, count = 0.
2. Generate samples data based on design specification.

3. Perform the hypothesis test based on the generated data at the predetermined significance level ($\alpha = 0.05$).
4. Reject H_0 if $W \geq$ upper critical value or $W \leq$ lower critical value.
5. If the decision is reject H_0 , then increase count by one (count = count + 1).
6. Repeat step 2 to step 4 for 1000 times.
7. Calculate the average Type I error rates by dividing count by 1000.
8. Repeat this simulation for 20 different conditions (4 distributions x 5 sample sizes).

Repeat all the steps (step 1 to step 8) for five thousand and ten thousand replications.

CHAPTER 4

RESULT AND ANALYSIS

4.1 Introduction

The performance of the modified Wilcoxon signed rank test was measured in terms of robustness. The test was conducted on data with various combinations of test conditions to highlight the strengths and weaknesses of the statistical test. The test conditions involved the various combinations of types of distribution, sample sizes and number of replication.

The performance of the test in terms of robustness was assessed at $\alpha = 0.05$ level of significance. To evaluate a particular condition under which a test is sensitive to assumption violations, Bradley's (1978) liberal criterion of robustness was employed. In order for a test to be considered robust, its empirical Type I error rate must be contained in the interval of [0.5α, 1.5α] or [0.025, 0.075]. A test is considered to be non robust if, for a particular condition, its Type I error rate is not within this interval. We adopted this standard because it was widely used by most researchers studying on robustness (e.g. Keselman et al., 2002; Wilcox et al., 1998; Othman et al., 2004; Syed

Yahaya, Othman & Keselman, 2004; Syed Yahaya et al., 2006). A procedure is considered highly robust if its estimated Type I error falls within the Bradley's liberal criterion and close to the nominal (significance) level. Estimated Type I error rates outside these intervals are considered either conservative or liberal for $p < 0.025$ and $p > 0.075$, respectively.

4.2 Type I Error Rates

The outcome measures for this study which is the Type I error rates were shown in Table 4.1. As mentioned in the previous chapter, the Wilcoxon signed rank test requires that the difference score come from a distribution that is approximately symmetric. The result from Table 4.1 shows that the modified Wilcoxon signed rank test is able in controlling the Type I error rates for all symmetric distributions. The cell with bold values represent the conditions with Type I error rates outside the Bradleys interval. All the Type I error rates obtained under normal, Beta and *g*-and-*h* distributions are lower than the nominal level of 0.05. However, there are a few conditions where the test produced conservative Type I error rates (below 0.025). Some researchers would consider that the procedures with conservative Type I error rates fail to perform. However, according to Mehta and Srinivasan (1970) and Hayes (2005), conservative procedures can still be considered as robust.

The finding displays that the test do not perform well under skewed distribution (chi-square with 3 d.f.) where it produced liberal Type I error rates (above 0.075) under all replications. The bold values under chi-square distributions indicate the Type I error

rates are liberal as it exceed Bradley liberal criterion of 0.075. With respect to the number of replications, different number of replications used do not influence the outcome because there are not much different in the Type I error rates produced by the test among the three values.

Table 4.1: Type I error rates

Distribution	Sample sizes	Type I error		
		1000 Replications	5000 Replications	10,000 Replications
Normal	10	**0.023**	**0.021**	**0.020**
	15	**0.024**	**0.023**	0.025
	20	0.027	0.025	0.026
	25	0.027	0.027	0.027
	30	**0.024**	**0.022**	**0.022**
Beta(0.5,0.5)	10	0.029	0.030	0.026
	15	**0.021**	0.025	**0.024**
	20	0.035	0.027	0.025
	25	**0.021**	**0.024**	**0.024**
	30	0.025	**0.024**	**0.023**
g=0 h=0.225	10	**0.023**	**0.021**	**0.020**
	15	0.025	**0.023**	0.026
	20	0.026	0.026	0.026
	25	0.027	0.025	0.025
	30	0.025	**0.022**	**0.023**
Chi-Square(3)	10	0.067	0.067	0.062
	15	0.054	0.064	0.067
	20	0.067	**0.084**	**0.082**
	25	**0.010**	**0.092**	**0.091**
	30	**0.091**	**0.087**	**0.092**

4.3 Calculation of Wilcoxon Statistic Based on Example Data

We believe that the inclusion of indicator zero differences might have some effect in the calculation of Wilcoxon statistic, *W*. However, based on simulation data, we cannot show the result for the original Wilcoxon signed rank test due to the time constrain in developing the programming. For that reason, both Wilcoxon tests were demonstrated using an example data as shown in Table 4.2 to show that the difference indicator influence the result of the Wilcoxon statistics. This section will be separated into two parts which are the case without zero difference and the case with zero difference. The Wilcoxon statistic, *W* is compared with the critical value from the Wilcoxon Table. For two-tail test, H_0 is rejected if *W* is less than or equal to the lower critical value or *W* is greater than or equal to the upper critical value. We used the example of two-tail test from Lind, Marchal and Wathen (2005), page 560. The situation is;

> *Suppose Toyota Motor Corporation is studying the effect of regular versus high-octane gasoline on fuel economy of its new high-performance, 3.5-liter, V6 engine. Ten executives are selected and asked to maintain records on the number of miles traveled per gallon of gas. Is there a difference in the number of miles traveled per gallon between regular and high-octane gasoline?*

Table 4.2: Number of miles traveled per gallon of gas

Executive	Miles per Gallon	
	Regular	High-Octane
Bowers	25	28
Demars	33	31
Grasser	31	35
DeToto	45	44
Kleg	42	47
Rau	38	40
Greolke	29	29
Burns	42	37
Snow	41	44
Lawless	30	44

4.3.1 Calculate the Wilcoxon signed rank test with zero difference

Both calculation are based on $\alpha = 0.05$. Based on the Wilcoxon signed rank test, the statistic W is 11. Since there is zero difference, so remove one executive and the sample become 9. From the Wilcoxon Table, lower and upper critical value is 5 and 40, respectively. Since $W = 11$ is between 5 and 40, we fail to reject the null hypothesis.

Table 4.3: Wilcoxon signed rank test with zero difference

Executive	Miles per Gallon		D_i	$\lvert D_i \rvert$	R_i	Sign of D_i
	Regular	High-Octane				
Bowers	25	28	-3	3	4.5	-
Demars	33	31	2	2	2.5	+
Grasser	31	35	-4	4	6	-
DeToto	45	44	1	1	1	+
Kleg	42	47	-5	5	7.5	-
Rau	38	40	-2	2	2.5	-
Greolke	29	29	0	Discard	Discard	Discard
Burns	42	37	5	5	7.5	+
Snow	41	44	-3	3	4.5	-
Lawless	30	44	-14	14	9	-

Next, Table 4.4 shows the calculation based on the modified Wilcoxon signed rank test. The Wilcoxon statistic W is equal to 14.5. Sample size is still ten executives. Based on the Wilcoxon Table, lower and upper critical value is 8 and 47, respectively. Since 14.5 falls between 8 and 47, we fail to reject the null hypothesis. Even though both test give different value of Wilcoxon statistic, but both test have the same conclusion which is fails to reject the null hypothesis.

Table 4.4: Modified Wilcoxon signed rank test with zero difference

Executive	Miles per Gallon		D_i	$\|D_i\|$	R_i	e_i	$R_i e_i$
	Regular	High-Octane					
Bowers	25	28	-3	3	5.5	0	0
Demars	33	31	2	2	3.5	1	3.5
Grasser	31	35	-4	4	7	0	0
DeToto	45	44	1	1	2	1	2
Kleg	42	47	-5	5	8.5	0	0
Rau	38	40	-2	2	3.5	0	0
Greolke	29	29	0	0	1	0.5	0.5
Burns	42	37	5	5	8.5	1	8.5
Snow	41	44	-3	3	5.5	0	0
Lawless	30	44	-14	14	10	0	0

4.3.2 Calculate the Wilcoxon signed rank test without zero difference

To make the data without zero difference, we changed the value of miles per gallon for regular gasoline for the Greolke executive from 29 to 31. Then we proceed to calculate the two Wilcoxon statistic. Both Wilcoxon statistic W based on the Wilcoxon signed rank test and the modified Wilcoxon signed rank test give the same value which is 15.5. From the Wilcoxon Table with sample size of ten, lower and upper critical value are 8 and 47, respectively. Since 15.5 is upper than 8 but lower than 47, we fail to reject the null hypothesis.

Table 4.5: Wilcoxon signed rank test without zero difference

Executive	Miles per Gallon		D_i	$\lvert D_i \rvert$	R_i	Sign of D_i
	Regular	High-Octane				
Bowers	25	28	-3	3	5.5	-
Demars	33	31	2	2	3	+
Grasser	31	35	-4	4	7	-
DeToto	45	44	1	1	1	+
Kleg	42	47	-5	5	8.5	-
Rau	38	40	-2	2	3	-
Greolke	31	29	2	2	3	+
Burns	42	37	5	5	8.5	+
Snow	41	44	-3	3	5.5	-
Lawless	30	44	-14	14	10	-

Table 4.6: Modified Wilcoxon signed rank test without zero difference

Executive	Miles per Gallon		D_i	$\|D_i\|$	R_i	e_i	$R_i e_i$
	Regular	High-Octane					
Bowers	25	28	-3	3	5.5	0	0
Demars	33	31	2	2	3	1	3
Grasser	31	35	-4	4	7	0	0
DeToto	45	44	1	1	1	1	1
Kleg	42	47	-5	5	8.5	0	0
Rau	38	40	-2	2	3	0	0
Greolke	31	29	2	2	3	1	3
Burns	42	37	5	5	8.5	1	8.5
Snow	41	44	-3	3	5.5	0	0
Lawless	30	44	-14	14	10	0	0

From these two examples, it is clearly shown that for situation without zero difference, there is no different in the Wilcoxon statistic obtained. Therefore, the used of indicator differences have no effect on the Wilcoxon statistic under this situation. However, when there is zero difference, the use of indicator differences do have an effect on the Wilcoxon statistic, W.

CHAPTER 5

CONCLUSION

5.1 Introduction

The goal of this research is to evaluate the performance of the modified Wilcoxon signed rank test in controlling the Type I error rates when we considered positive, zero and negative differences in calculating the Wilcoxon statistic. The performance of the modified Wilcoxon signed rank test was measured in terms of the Type I error rates. The robustness of the statistical test was evaluated based on the Bradley liberal criterion of robustness. The test was considered robust if the Type I error rates fall between 0.025 and 0.075.

5.2 Performance of the Modified Wilcoxon Signed Rank Test

The modified Wilcoxon signed rank test requires that the difference score come from a distribution that is approximately symmetric. The test is able in controlling the Type I

error rates for all symmetric distribution even though the values obtained is quite conservative. From this finding, the modified Wilcoxon signed rank test is robust for non skewed distribution. However, under skewed distribution, some of the conditions tested showed non robust with liberal Type I error rates. Different numbers of replication used do not influence the outcome of the test.

5.3 Limitation and Suggestion for Future Research

From this study, we notice that the use of the indicator differences (positive, zero and negative) could work for Wilcoxon Signed rank test. The performance of the modified Wilcoxon signed rank test should be compared with the original Wilcoxon signed rank test. However, due to the time constraint, the comparison between this two Wilcoxon tests could not be done because longer period is needed to write the programming for the original Wilcoxon signed rank test. Therefore, for future research we are interested to compare the performance of the original and the modified Wilcoxon singed rank test in controlling the Type I error rate and also to conduct the power analysis for these two Wilcoxon tests.

REFERENCES

Ahad, N. A., Othman, A. R., & Syed Yahaya, S. S. (2013). New procedure in testing differences between two groups. *Applied Mathematics & Information Sciences*, 7, No. 2L, 397-401.

Ahad, N. A., Othman, A. R., & Syed Yahaya, S. S. (2012). Performance of two-samples pseudo-median procedure. *Sains Malaysiana*, 41(9), 1149-1154.

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. *Journal of Modern Applied Science*, *5*(5), 131-139.

Algina, J., Keselman, H. J., & Penfield, R. D. (2005). An alternative to Cohen's standardized mean difference effect size: A robust parameter and confidence interval in the two independent groups case. *Psychological Methods*. *10*(3), 317-328.

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

Conover, W. J. (1980). *Practical nonparametric statistics* (2nd ed.). New York: Wiley.

Cressie, N. A. C., & Whitford, H. J. (1986). How to use the two sample *t*-test. *Biometrical Journal*, *28*, 131-148.

Gibbons, J. D. (1993). *Nonparametric statistics: An introduction*. Newbury Park, CA: Sage.

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

Hayes, A. F. (2005). *Statistical methods for communication science*. Mahwah, NJ: Erlbaum.

Heeren, T., & D`Agostino, R. B. (1987). Robustness of the two independent simples *t*-test when applied to ordinal scaled data. *Statistics in Medicine*, *6*, 79-90.

Hess, B., Olejnik, S., & Huberty, C. J. (2001). The efficacy of two improvement-over-chance effect sizes for two-group univariate comparisons under variance heterogeneity and nonnormality. *Educational and Psychological Measurement*, *61*, 909-936.

Hettmansperger, T. P., Mckean, J. W., & Sheather, S. J. (2000). Robust nonparametric methods. *Journal of the American Statistical Association*, *95*, 1308-1312.

Higgins, J. J. (2004). *An introduction to modern nonparametric statistics*. Belmont, CA: Duxbury Press.

Hoaglin, D. C. (1985a). Using quantiles to study shape. In D. Hoaglin, F. Mosteller, and J. Tukey (Eds.), *Exploring Data Tables, Trends, and Shapes* (pp. 417-458) New York: Wiley.

Hoaglin, D. C. (1985b). Summarizing shape numerically: The *g*-and-*h* distributions. In D. Hoaglin, F.Mosteller, and J. Tukey (Eds.), *Exploring Data Tables, Trends, and Shapes* (pp. 461-508). New York: Wiley.

Hollander, M., & Wolfe, D. A. (1999). *Nonparametric statistical methods* (2^{nd} ed.*)*. New York: Wiley.

Hotelling, H., & Pabst, M. R. (1936). Rank correlation and tests of significance involving no assumption of normality. *The Annals of Mathematical Statistics*, *7*(1), 29-43.

Huber, P. J. (2004). *Robust statistics*. New York: Wiley.

Kang, Y. & Harring, J. R. (2012). Investigating the Impact of Non-Normality, Effect Size, and Sample Size on Two-Group Comparison Procedures: An Empirical Study. Retrieved from http://education.umd.edu/EDMS/fac/Harring/Misc/Kang&H-2012.pdf.

Keselman, H. J., Othman, A. R. & Wilcox, R. R. (2013). Preliminary Testing for Normality: Is This a Good Practice? *Journal of Modern Applied Statistical Methods*, 12(2), 2-19.

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

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.

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.

Keselman, H. J., Wilcox, R. R., Taylor, J., & Kowalchuk, R. K. (2000). Test for mean equality that do not require homogeneity of variances: Do they really work? *Communication in Statistics: Simulation and Computation*, *29*, 875-895.

Lind, D. A., Marchal, W. G. & Wathen, S. A. (2005). *Statistical techniques in business & economics*. New York: McGraw-Hill.

Mann, P. S., & Whitney, D. R. (1947). On a test of whether on of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, *18*(1), 50-60.

McSeeney, M., & Katz, B. M. (1978). Nonparametric statistics; Use and nonuse. *Perceptual and Motor Skills*, *4*, 1023-1032.

Mehta, J. S., & Srinivasan, R. (1970). On the Behren-Fisher problem. *Biometrika*, *57*, 649-655.

Miles, J., & Shevlin, M. (2001). *Applying regression and correlation*. London, United Kingdom: Sage.

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

Mooney, C. Z. (1997). *Monte Carlo simulation*. Thousand Oaks, CA: Sage.

Othman, A. R., Padmanabhan, A. R., & Keselman. H. J. (2003). Extending the Mann-Whithney procedure to *J*-samples. In A. Ahmed, Z. Jubok, C. M. Ho, R. Roslan, and A. F. Pang (Eds.), *Prosiding Simposium Kebangsaan Sains Matematik ke-XI: Penyelidikan dan Pendidikan Sains Matematik Teras Kecemerlangan Ilmu [Proceedings of the Eleventh National Mathematical Sciences Symposium: Mathematical Science Research and Education, Pillars of Academic Excellence*] (pp. 554-562). Kota Kinabalu, Malaysia: Universiti Malaysia Sabah.

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.

SAS Institute Inc. (2006). *SAS online doc*. Cary, NC: SAS Institute Inc.

Savage, I. R. (1953). Bibliography of a nonparametric statistics and related topics. *Journal of American Statistical Association*, *48*, 844-906. Correction 53 (1958), 1031.

Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

Steland, A., Padmanabhan, A. R., & Akram, M. (2011). Resampling methods for the nonparametric and generalized Behrens-Fisher problems. *Sankhya: The Indian Journal of Statistics Series A*, *73*(2), 267-302.

Sullivan, I. M., & D`Agostino, R. B. (1996). Robustness and power of analysis of covariance applied to data distorted from normality by floor effects. *Statistics in Medicine*, *15*, 477-496.

Syed Yahaya, S. S., Othman, A. R., & 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*. (pp. 319 – 328). Basel, Switzerland: Birkhausher.

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

Thijs van den Berg (2013). Generating correlated random numbers. Retrieved from http://www.sitmo.com/article/generating-correlated-random-numbers/

Tiku, M. L., Tan, W. Y., & Balakrishnan, N. (1986). *Robust inference*. New York: Marcel Dekker.

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

Wilcox, R. R., Keselman, H. J., & Kowalchuk, R. K. (1998). Can tests for treatment group equality be improved?: The bootstrap and trimmed means conjecture. *British Journal of Mathematical and Statistical Psychology*, *51***,** 123-134.

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