CHAPTER 19
Type I Error of the Modified Wilcoxon Signed Rank Test under Leptokurtic Distribution

Nor Aishah Ahad, Sharipah Soaad Syed Yahaya, Suhaida Abdullah, Lim Yai Fung and Zahayu Md Yusof

Abstract. Group comparisons are at the heart of many research questions addressed by researchers. Making inferences and drawing conclusions through statistical hypothesis testing on the differences between groups is actively adopted by researchers in many disciplines. When the groups are dependent, and violation of normality assumption occurred, the most commonly used method like paired t-test, usually produced doubtful result which will lead to misleading conclusions. As alternative, researchers tend to choose nonparametric Wilcoxon signed rank test for the purpose. The computation of this statistic involves ranking the absolute difference of each pair of observations and any pair with 0 differences will be discarded. In this study, the statistic was modified by including the 0 differences in the ranking. The empirical Type I error rates of the modified statistical test was measured via Monte Carlo simulation. These rates were obtained under the combination of leptokurtic distributional shapes with various sample sizes and number of replications. The modified Wilcoxon signed rank test was found to be more robust under symmetric leptokurtic with conservative values as compared to the skewed leptokurtic distribution. The finding also indicated that different number of replications had no effect on Type I error.

Keywords: Wilcoxon signed rank test; zero difference; type I error rate; Leptokurtic distribution

Nor Aishah Ahad (✉) • Sharipah Soaad Syed Yahaya • Suhaida Abdullah • Lim Yai Fung • Zahayu Md Yusof
School of Quantitative Sciences, Universiti Utara Malaysia
e-mail: aishah@uum.edu.my, sharipah@uum.edu.my, suhaida@uum.edu.my, yaifung@uum.edu.my, zahayu@uum.edu.my

1 Introduction

Statistical tests for comparing groups have been developed to permit comparison regarding the degree to which qualities of one group differ from the other groups. Each test is based on certain assumptions about the population(s) from which the data are drawn. If a particular statistical test is used to analyze data collected from a sample that does not meet the expected assumptions, then the conclusion drawn from the results of the test will be flawed.

The two major classes of statistical tests are parametric and nonparametric. Before a parametric test can be undertaken, it must be ascertained that the data are normally distributed. Very often the variables within data sets from education and psychology are not normally distributed [1][2]. In his study, Micceri [2] surveyed 440 data sets from psychology and education sources and determined that virtually none of the data sets could be adequately characterized by a normal distribution. Micceri [2] described the distributions he examined as having varying degrees of multimodality, asymmetric, and excessive tail weight. 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 [2][3]. 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. In some cases, transformation can be applied to rectify the problem. 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 do not need such rigid assumptions. The nonparametric alternative to the paired *t*-test is the Wilcoxon signed rank test. The Wilcoxon signed rank test requires less stringent assumptions, such that the difference in scores come from a distribution that is approximately symmetric and the data are measured on either ordinal, interval, or ratio scale. Nonparametric tests use rank or frequency information to draw conclusions about differences between populations.

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. The dependency 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 is the difference between the values of the observations rather than the values of the observations themselves.

One assumption needed in Wilcoxon signed rank tets is that the differences represent observations on a continuous random variable and zero differences do not exist in the calculation of the statistic. In practice, however, zero differences do occur. The usual procedure in such cases is to discard observations that lead to zero differences and thus will reduce the sample size accordingly [4].

The Wilcoxon signed ranks test uses the test statistic W. The computation of this statistic uses the differences between paired items. First, find absolute differences, then, arrange the differences 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 the sign of the differences (positive or negative) and for the tied values, get the average of their ranks. 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 rank sums.

Studies by [5][6][7][8] 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. In this study, we employed the same indicator function where we considered positive, zero and negative differences in calculating the Wilcoxon statistic, W, for paired (dependent) sample. The performance of the statistic in terms of controlling Type I error rate was then measured via Monte Carlo simulation.

This paper is organized as follows. The next section will be reviewing on the procedure employed in the Wilcoxon signed rank test. Description of the design specification is given in the third section. The fourth section describes the results and discussion. The conclusion is elaborated in the final section.

2 Procedure Employed in the Wilcoxon Signed Rank Test

Wilcoxon [9] introduced the rank sum tests for unpaired groups and signed rank test for paired groups which are named after him. He stated 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.

In this study, the Wilcoxon signed rank test was modified with the inclusion of the indicator function zero difference to obtain the Wilcoxon statistic, W. The two-tail test of the population median difference, M_D is given by Eq. (1).

$$H_0: M_D = 0 \qquad H_1: M_D \neq 0 \tag{1}$$

Given two sets of paired data (X_1 , Y_1). Find the sequence difference between X_1 and Y_1 where

$$D_i = X_{1i} - Y_{1i} \tag{2}$$

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 Eq. (3).

$$e_i = \begin{cases} 1 & if\ D_i > 0 \\ 0.5 & if\ D_i = 0 \\ 0 & if\ D_i < 0 \end{cases} \tag{3}$$

Based on Equation 3, determine e_i with regards to the differences, D_i. Then the Wilcoxon statistic is defined as Eq. (4).

$$W = \sum_{i=1}^{n} R_i e_i \tag{4}$$

For a two-tailed test and for a particular level of significance, if the observed value of W is equal or 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 Design Specification

A few conditions that have effect on the performance of the 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 is to scrutinize the strength and weakness of the method as well as its robustness.

The sample size was manipulated to be 10, 15, 20, 25 and 30. We focus on small sample sizes because of the availability of the Wilcoxon table for the critical values. For large sample size, the test statistic W is approximately normally distributed.

The next variable considered was the distributional shape. The shape of a distribution is usually depicted by skewness and kurtosis. Skewness is a departure from symmetry [10]. Kurtosis, on the other hand, is a measure of whether the data are peaked or flat relative to a normal distribution. In a simple description, larger kurtosis refers to heavier tails [10]. This study focused on leptokurtic distribution where the distributions have a positive kurtosis. According to Miles and Shevlin [11], the term 'leptokurtic' is originally from the Greek word 'leptos', meaning small or slender. In other words, positive kurtosis indicates a "peaked" distribution. The distributions used in this study were the *g*-and-*h* distribution from Hoaglin, [12] with $g = 0$ and $h = 0.225$, and chi-square with three degree of freedom $\left(\chi_3^2\right)$ representing symmetric and asymmetric leptokurtic, respectively. 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 became more skewed as *g* increased and heavier as *h* increased. Table 1 shows the distributions used in this study together with their levels of skewness and kurtosis.

Table 1. Distributions used in the study.

Distributional Shape	Distribution Identified	Skewness	Kurtosis
Symmetric Leptokurtic	$g = 0, h = 0.225$	0	154.84
Asymmetric Leptokurtic	Chi-square (3)	1.63	4.00

In this study, we used 1000, 5000 and 10,000 replications for each distribution for each study condition. These different replication sizes have been used by Kang and Harring [13] in their simulation study. The same number of replications were also used by many researchers [14][15][16][17][18][19].

Table 2 shows the design specification of this study. The combination of five sample sizes paired with types of distribution and number of simulation produced a total of 30 different conditions for testing. Type I error rate for each condition was examined. The following table shows the design specifications and the test conditions.

Table 2. Design specifications and test conditions of the study.

Distribution	Sample Sizes	Number of Simulation
	10	
$g = 0, h = 0.225$	15	1000
	20	5000
χ_3^2	25	10000
	30	

This study was based on simulated data. The simulation was carried out using random-number-generating function in SAS and the simulation program was written in SAS/IML[20]. In terms of the data generation procedure, pseudo-random variates for each particular distributional shape was obtained in the following manner:

a) Standard normal distribution.
Pseudo-random normal variates were generated by employing the SAS generator RANDGEN [20]. This involved the straight forward usage of the (RANDGEN(Y, 'NORMAL')) to generate normal variates with mean equals to zero and standard deviation equals to one.

b) 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 5 to obtain the symmetric leptokurtic distribution.

$$Y = Ze^{\frac{hZ^2}{2}} \tag{5}$$

c) 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)).

4 Results and Discussion

To evaluate the robustness of the test to a particular condition, Bradley's [21] liberal criterion of robustness was employed. According to Bradley's liberal criterion of robustness, a test can be considered robust if its empirical rate of

Type I error is within the interval [0.5α, 1.5α] or [0.025, 0.075] when α=0.05. A test is considered liberal if its Type I error rate is greater than the nominal level. Whereas, it is considered conservative if its Type I error rate is less than the nominal level. The outcome measures for this study are shown in Table 3.

Table 3. Type I error rates.

Replications	Sample sizes	Type I error	
		g=0 h=0.225	Chi-square (3)
1000	10	**0.023**	0.067
	15	0.025	0.054
	20	0.026	0.067
	25	0.027	**0.010**
	30	0.025	**0.091**
5000	10	**0.021**	0.067
	15	**0.023**	0.064
	20	0.026	**0.084**
	25	0.025	**0.092**
	30	**0.022**	**0.087**
10000	10	**0.020**	0.062
	15	0.026	0.067
	20	0.026	**0.082**
	25	0.025	**0.091**
	30	**0.023**	**0.092**

Note: Highlighted values show that the Type I error is not within the interval [0.025, 0.075]

As mentioned in the previous section, the Wilcoxon signed rank test requires that the difference score come from a distribution that is approximately symmetric. The result shows that the Type I error rates for Wilcoxon signed rank test are all conservative for symmetric leptokurtic and liberal for asymmetric leptokurtic. For some researchers, the tests with conservative Type I error rates are considered as non robust or fail in controlling Type I error. However, according to Mehta and Srinivasan [22] and Hayes [23], conservative procedures can still be considered as robust. Under *g*-and-*h* distribution, the empirical Type I error rates are not within the Bradley's liberal criterion when sample size equal to 10 for all number of simulations. However, under chi-square (3 df), large sample size (n = 25 and 30) produced empirical Type I error rates beyond the interval of [0.025, 0.075]. In terms of number of simulation, the finding indicates that different number of simulations have no effect on Type I error because the empirical values are barely different with each other within the distribution.

5 Conclusion

Group comparisons are the common statistical methods employed by researchers. When the groups are dependent, and violation of normality assumption occurred, the most commonly used method is the nonparametric Wilcoxon signed rank test. The computation of this statistic involves ranking the absolute difference of each pair of observations and any pair with 0 differences will be discarded. In this study, the statistic was modified by including the 0 differences in the ranking. The empirical Type I error rates of the modified statistical test was measured via Monte Carlo simulation. The finding shows that this test is able to control the Type I error rate within the Bradley's liberal criterion. However, under certain conditions, the Type I error rate is too conservative and liberal.

Acknowledgements.
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 (Code S/O 12406) and to RIMC for facilitating the management of the research.

References

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

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

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

[4] Daniel, W.W.: Applied nonparametric statistics (2nd ed.). Pacific Grove, CA: Duxbury. (1990)

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

[6] Ahad, N. A., Othman, A. R., Syed Yahaya, S. S.: 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 (2011)

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

[8] Ahad, N.A., Othman, A.R., Syed Yahaya, S.S.: New procedure in testing differences between two groups. Applied Mathematics & Information Sciences. 7, 397-401 (2013)

[9] Wilcoxon, F.: Individual comparison by ranking methods. Biometrics. 1, 80-83 (1945)

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

[11] Miles, J., Shevlin, M.: Applying regression and correlation. Sage. United Kingdom (2001)

[12] Hoaglin, D.C.: 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). Wiley, New York (1985b)

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

[14] Othman, A.R., Padmanabhan, A.R., Keselman, H.J.: 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. Universiti Malaysia Sabah, Sabah (2003)

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

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

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

[18] Keselman. H.J., Wilcox. R.R., Othman, A.R., Fradette. K.: Trimming, transforming statistics, and bootstrapping: Circumventing the biasing

effects of heteroscedasticity and nonnormality. Journal of Modern Applied Statistical Methods. 1(2), 288-309 (2002)

[19] Hess, B., Olejnik, S., Huberty, C.J.: 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 (2001)

[20] SAS Institute Inc. SAS online doc. Cary, NC: SAS Institute Inc. (2006)

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

[22] Mehta, J.S., Srinivasan, R.: On the Behren-Fisher problem. Biometrika. 57, 649-655 (1970)

[23] Hayes, A.F.: Statistical methods for communication science. Frlbaum, Mahwah, NJ (2005)