Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.12785/amis/072L03

New Procedure in Testing Differences between Two Groups

Nor Aishah Ahad[1], *Abdul Rahman Othman*[2] *and Sharipah Soaad Syed Yahaya*[3]

[1] UUM College of Arts and Sciences, Universiti Utara Malaysia, 06010 Sintok, Kedah Malaysia
[2] School of Distance Education, Universiti Sains Malaysia, 11800 Pulau Pinang, Malaysia
[3] UUM College of Arts and Sciences, Universiti Utara Malaysia, 06010 Sintok, Kedah Malaysia

Received: 3 Oct. 2012, Revised: 25 Dec. 2012, Accepted: 28 Dec. 2012
Published online: 1 Jun. 2013

Abstract: Despite the theoretical correctness of the *t*-test in testing differences between two groups and the existence of the nonparametric backup, i.e. Mann-Whitney-Wilcoxon test, these test fail to simultaneously control Type I error and maintain adequate power under certain condition. This study intends to alleviate this problem by applying the pseudo-median as the location measure of interest into the one-sample nonparametric Wilcoxon procedure in a two group setting. Pseudo-median is the median of all possible differences of observations from the two groups. Since the sampling distribution of this procedure is intractable, the bootstrap method was used to achieve the significance level. The finding shows that the new procedure has the ability to control Type I error rates and maintaining high power rates regardless of distributional shape whether symmetrical or asymmetrical. The performance of the new procedure is compatible to *t*-test and Mann-Whitney-Wilcoxon test.

Keywords: Pseudo-median, robust, Type I error, power

1 Introduction

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. If the underlying distributions are normally distributed with equal population variances, it is well known that the most suitable test statistic to use is the Student's *t*-test. Unfortunately, this test statistic is sensitive to non-normality of data and heterogeneity of variances. For this situation, Welch's approximate test [1] offers the best practical solution. However, until today this statistic still has problems in controlling Type I error probabilities under non-normal distributions.

A popular alternative for analyzing data from non-normal populations is to select a nonparametric method such as the Mann-Whitney-Wilcoxon (MWW) test. Even though nonparametric methods are distribution free, they are not assumptions free. Usually the underlying distribution has to be symmetric.

The issue whether any methods for comparing two independent groups can provide reasonable control over Type I error and simultaneously improve power rates under the violations of normality and variance homogeneity has received considerable attention. The development of new methods in testing the equality of location measures in the one-way independent groups design by controlling Type I error and power rates does raise a serious attention and remains a very active area of study. Even though many methods have been proposed, researchers realized that no single statistical method is ideal in all situations encountered in applied work because different methods are sensitive to different features of the data.

In this study, a method to work in both normal and non-normal distributions was suggested. The proposed method known as a pseudo-median (PM) procedure, used pseudo-median of differences between group values as the statistic of interest in the modification of the one sample nonparametric Wilcoxon procedure in a two groups setting. The pseudo-median of a distribution F is defined to be the median of distribution $\left(\frac{u+v}{2}\right)$, where u and v are independent, each with the same distribution F [5-6]. Hodges-Lehmann estimator was used to estimate the pseudo-median values.

* Corresponding author e-mail: aishah@uum.edu.my

This paper is organized as follows. In Section 2, we provide the method used in this study, followed by empirical investigation. The results of Monte Carlo study and discussion are displayed in Section 3 and finally conclusion in Section 4.

2 Methodology

This study covers both symmetric and asymmetric distributions. The methods applied to the two types of distributions are quite different. Let $X_1 = (X_{11}, X_{12}, ..., X_{1n_1})$ be samples from distribution F_1 and $X_2 = (X_{21}, X_{22}, ..., X_{2n_2})$ be samples from distribution F_2, respectively. The pseudo-median is defined as

$$\hat{d} = median\left(\frac{(X_{1i} - X_{2j}) + \left(X_{1i'} - X_{2j'}\right)}{2}\right) \quad (1)$$

where $i \neq i'$and $j \neq j'$. When F_1 and F_2 are symmetric, d can be defined as the difference between the centres of symmetry. Hence, the null hypothesis is $H_0 : d = 0$. Let$D_{ij} = X_{1i} - X_{2j}$, $i = 1,2,...,n_1$ and $j = 1,2,...,n_2$ and $N = n_1 n_2$. The statistic is a one-sample Wilcoxon statistic based on the $ND'_{ij}s$. Let R_{ij} denotes the rank of $|D_{ij}|$ and let the indicator function be

$$e_{ij} = \begin{cases} 0, & D_{ij} < 0 \\ 0.5, & D_{ij} = 0 \\ 1, & D_{ij} > 0 \end{cases} \quad (2)$$

Then the statistic is defined as

$$W = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} R_{ij} e_{ij} \quad (3)$$

The modification of the Wilcoxon procedure is performed by adding the pseudo-median value to the second sample to form a new sample, $X_2 + \hat{d} = \left(X_{21} + \hat{d}, X_{22} + \hat{d}, ..., X_{2n_2} + \hat{d}\right)$ where $\hat{d}$ is the estimate of d. Then the aligned difference based on the location-aligned samples becomes, $\hat{D}_{ij} = X_{1i} - \left(X_{2j} + \hat{d}\right)$. Define the aligned statistic as Equation 4 where $\hat{W}$represents the (approximate) value of the statistic, when H_0 is true.

$$\hat{W} = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \hat{R}_{ij}\hat{e}_{ij} \quad (4)$$

Since we have realigned the second sample with the estimated d, we need to find the pseudo sampling distribution for the estimated W. We proposed to use bootstrap procedure to construct the hypothesis test. The reason of using bootstrapping method was due to the fact that the sampling distribution for the statistic used was intractable. Bootstrapping was conducted by separately bootstrap n_1 observations from X_1 group and n_2 observations from $\left(X_2 + \hat{d}\right)$ group to obtain bootstrap samples, $X_1^* = \left(X_{11}^*, X_{12}^*, ..., X_{1n_1}^*\right)$ and $X_2^* = \left(X_{21}^*, X_{22}^*, ..., X_{2n_2}^*\right)$. Then the bootstrap differences become $D_{ij}^* = X_{1i}^* - X_{2j}^*$. Therefore, the bootstrap statistic is defined as

$$W^* = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} R_{ij}^* e_{ij}^* \quad (5)$$

In the case of symmetric distribution, d coincides with the difference between the center of symmetry between two groups. Therefore, without loss of generality, we may assume that $d = 0$. For asymmetric distributions, we cannot assume the difference between the center of symmetry between two groups as zero. Therefore, to ensure the setting for the null condition, we have to determine a constant a and add it to the members of the second sample. Determination of a algebraically or analytically seems intractable, so we use simulation to obtain its value. To calculate the constant a, two samples of equal size $n_1 = n_2 = 2$ are generated from the same distribution as X_1 and X_2. For example, let F_1 and F_2 be two skewed distributions where the standard deviations need not be the same. Let $Y_1 = (Y_{11}, Y_{12})$ and $Y_2 = (Y_{21}, Y_{22})$ be any samples of size two from F_1 and F_2, respectively. Compute a as given in Equation (6).

$$a = \left[\frac{(Y_{11} - Y_{21}) + (Y_{12} - Y_{22})}{2}\right] \quad (6)$$

Repeat the computation 10,000 times to get $a_1, a_2, ..., a_{10,000}$. The rule of thumb in simulation studies requires computation of 1000 times if the sampling distribution is known. In this case, the sampling distribution is unknown, and therefore requires a larger number of trials. So we choose 10,000 for this purpose. The median of these 10,000 values is the value of a.

In this study, the effect size or the shift parameter used to obtain the statistical power is computed based on the common language (CL) statistic proposed by McGraw and Wong [2] and A from Vargha and Delaney's [3]. The effect size or the shift parameter used in this study is not a single point but its value varies from 0.2 to 2.0 with increment of 0.2 units.

In studying the robustness of this procedure, two variables were manipulated to create conditions that are known to highlight the strengths and weaknesses of the test for the equality of location parameters. The variables are sample sizes and types of distributions. This study was conducted under homogeneous variances 1:1. Empirical Type 1 error rates and statistical power were collected and later compared under various study conditions.

The number of groups and sample sizes were fixed. This study only covered the two groups case and the total sample sizes was set at N = 40. This value was later divided into two groups forming the balanced and unbalanced designs. For the balanced design, the value is

equally divided into $(n_1, n_2) = (20, 20)$ while for the unbalanced design, the groups were divided into (15,25).

To observe the effect of distributional shapes on Type I error and power of the procedure, this study focused on six distributions representing different degrees of skewness and kurtosis from both spectrum of symmetric and asymmetric distributions. 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 [4] with $g = 0$ and h =0.225. These distributions represent symmetric mesokurtic, platykurtic and leptokurtic, respectively. The normal distribution was used as the basis of comparison. Meanwhile, for asymmetric distributions, two distributions based upon Fleishman [5] transformation of the standard normal distribution with different skewness and kurtosis and the chi-square distribution with three degrees of freedom (χ^2_3) were chosen to represent skewed mesokurtic, platykurtic and leptokurtic, respectively. Table 1 shows the types of symmetrical and nonsymmetrical distributions used in this study together with their levels of skewness and kurtosis.

Table 1 Distributions used in the study

Distribution		Skewness	Kurtosis
Symmetric	Beta (0.5,0.5)	0	-1.5
	Normal (0,1)	0	0
	g=0, h=0.225	0	154.84
Asymmetric	Fleishman 1	0.5	-0.5
	Fleishman 2	0.75	0
	Chi-square (3)	1.63	4.00

Data from all of the distributions were generated using RANDGEN function [6]. To generate data from the *g*-and-*h* distribution, standard unit normal variates (Z_{ij}) were converted to *g*-and-*h* random variates via $Y_{ij} = Z_{ij} \exp\left(\frac{hZ_{ij}^2}{2}\right)$. The Z_{ij} scores were generated using the RANDGEN generator with normal distribution option. For each design, 599 bootstrap samples were generated, and 5,000 data sets were simulated.

3 Results and Discussions

To evaluate each particular condition under which a test was insensitive to assumption violations, Bradley's criterion of robustness [7] was employed. According to this criterion, for the five percent nominal level used in this study, a test is considered robust if its empirical rate of Type I error fell within [0.025, 0.075]. Correspondingly, a test is considered to be non-robust if, for a particular condition, its Type I error rate is not contained within the interval. We choose this criterion as it provides a reasonable standard for judging robustness.

The empirical Type I error rates for the investigated procedures are displayed in Table 2. The results showed that all the procedures produced robust Type I error rates under Bradley's liberal criterion of robustness. The disparity between Type I error rates from balanced and unbalanced design is minute and the rates are consistent across the investigated conditions. The nature of the sample sizes be it balanced or unbalanced, did not show much difference in the procedure's ability to control Type I error rates.

Table 2 Empirical Type I error rates

Distribution	Sample Sizes	PM	*t*-test	MWW
Normal	(20,20)	0.0552	0.054	0.0516
	(15,25)	0.0526	0.054	0.0456
Beta	(20,20)	0.046	0.0536	0.0546
	(15,25)	0.0528	0.0562	0.0492
g =0, h = 0.225	(20,20)	0.0588	0.0522	0.0516
	(15,25)	0.0566	0.0504	0.0456
Fleishman1 (F1)	(20,20)	0.0456	0.052	0.0508
	(15,25)	0.0528	0.0532	0.0458
Fleishman2 (F2)	(20,20)	0.0482	0.0524	0.0506
	(15,25)	0.0532	0.0522	0.0452
χ^2_3	(20,20)	0.0454	0.052	0.052
	(15,25)	0.0526	0.0482	0.0514

There are no formal standards for power. In determining the desired power levels, most researchers assess the power of their tests using 0.80 as the standard for adequacy. There are no hard and fast rules about how much power is enough, but according to Murphy and Myors [8], there seems to be a consensus about two things. First, power should be above 0.50. When power drops below 0.50, the study is more likely to fail. Second, power of 0.80 or above is usually judged to be adequate. Most power analyses specify 0.80 as the desired level of power to be achieved, and this convention seems to be widely accepted. In this study, 0.80 was used as the standard for adequacy in power analysis.

Power rates for all procedures are displayed in Table 3 to 5. All power rates that reached the standard level are in bold. We observe that all the investigated procedures achieved the adequate power rate of 0.80 at the shift parameter between 0.8 and 1.0. In Table 3, at shift parameter of 1.0, the psudo-median procedure achieved greater power rate under balanced group sample sizes as compared to unbalanced group sample sizes except for Fleishman 2 and chi-square distributions. These two distributions show that the procedure produced greater power rates under unbalanced group sample sizes.

The results in Table 4 and Table 5 show that *t*-test and Man-Whitney-Wilcoxon test achieved greater power rate

under balanced group sample sizes as compared to unbalanced group sample sizes at shift parameter of 1.0.

Table 3 Statistical power of pseudo-median procedure

	Normal	Beta	g=0, h=.225	F1	F2	χ^2_3
Shift Parameter	Group sizes (20,20)					
0.2	0.095	0.085	0.128	0.091	0.095	0.105
0.4	0.231	0.195	0.333	0.219	0.227	0.296
0.6	0.450	0.406	0.597	0.436	0.453	0.541
0.8	0.681	0.649	**0.803**	0.673	0.684	0.758
1.0	**0.860**	**0.835**	**0.929**	**0.854**	**0.859**	**0.894**
1.2	**0.956**	**0.959**	**0.978**	**0.954**	**0.953**	**0.961**
1.4	**0.990**	**0.991**	**0.992**	**0.989**	**0.988**	**0.987**
1.6	**0.998**	**0.998**	**0.996**	**0.999**	**0.999**	**0.998**
1.8	**1.000**	**1.000**	**0.999**	**0.999**	**0.999**	**0.999**
2.0	**1.000**	**1.000**	**0.999**	**1.000**	**0.999**	**0.999**
	Group Sizes (15,25)					
0.2	0.089	0.090	0.119	0.079	0.077	0.092
0.4	0.216	0.176	0.309	0.202	0.206	0.251
0.6	0.429	0.389	0.565	0.422	0.436	0.522
0.8	0.670	0.610	0.791	0.653	0.671	0.766
1.0	**0.829**	**0.806**	**0.898**	**0.848**	**0.862**	**0.917**
1.2	**0.947**	**0.941**	**0.967**	**0.956**	**0.960**	**0.975**
1.4	**0.984**	**0.988**	**0.984**	**0.988**	**0.989**	**0.994**
1.6	**0.997**	**0.998**	**0.996**	**0.999**	**0.999**	**0.999**
1.8	**0.999**	**0.999**	**0.998**	**0.999**	**0.999**	**0.999**
2.0	**1.000**	**1.000**	**0.999**	**1.000**	**1.000**	**1.000**

Table 4 Statistical power of *t*-test

	Normal	Beta	g=0, h=.225	F1	F2	χ^2_3
Shift Parameter	Group sizes (20,20)					
0.2	0.099	0.097	0.110	0.093	0.093	0.096
0.4	0.236	0.214	0.281	0.236	0.242	0.249
0.6	0.460	0.445	0.528	0.462	0.468	0.485
0.8	0.696	0.699	0.744	0.698	0.700	0.712
1.0	**0.866**	**0.869**	**0.875**	**0.869**	**0.866**	**0.863**
1.2	**0.962**	**0.968**	**0.947**	**0.963**	**0.961**	**0.951**
1.4	**0.991**	**0.993**	**0.972**	**0.992**	**0.991**	**0.982**
1.6	**0.998**	**0.998**	**0.987**	**0.999**	**0.998**	**0.994**
1.8	**1.000**	**1.000**	**0.993**	**1.000**	**1.000**	**0.999**
2.0	**1.000**	**1.000**	**0.996**	**1.000**	**1.000**	**0.999**
	Group Sizes (15,25)					
0.2	0.088	0.098	0.099	0.090	0.091	0.097
0.4	0.224	0.191	0.265	0.223	0.226	0.235
0.6	0.444	0.432	0.505	0.432	0.437	0.457
0.8	0.676	0.663	0.734	0.675	0.681	0.678
1.0	**0.841**	**0.845**	**0.856**	**0.842**	**0.843**	**0.843**
1.2	**0.953**	**0.954**	**0.938**	**0.954**	**0.954**	**0.939**
1.4	**0.987**	**0.990**	**0.968**	**0.988**	**0.987**	**0.980**
1.6	**0.997**	**0.999**	**0.985**	**0.998**	**0.998**	**0.994**
1.8	**0.999**	**0.999**	**0.991**	**1.000**	**1.000**	**0.999**
2.0	**1.000**	**1.000**	**0.996**	**1.000**	**1.000**	**0.999**

The results of the statistical power for all procedures as tabulated in Table 3 to Table 5 show the range of shift parameters when the procedures achieved the desired power of 0.80. The approximated shift parameters when the three procedures achieved power of 0.80 are shown in Table 6. Linear approximation is chosen here because of the two reasons. First, the equation of the power curve is unknown. Second, the approximation is carried out over a small range of effect size. Hence the portion of the power curve used in the approximation resembles a straight line.

Table 5 Statistical power of Mann-Whitney-Wilcoxon

	Normal	Beta	g=0, h=.225	F1	F2	χ^2_3
Shift Parameter	Group sizes (20,20)					
0.2	0.091	0.124	0.128	0.093	0.101	0.131
0.4	0.223	0.263	0.366	0.228	0.251	0.352
0.6	0.442	0.467	0.667	0.446	0.487	0.632
0.8	0.672	0.666	**0.871**	0.669	0.707	**0.842**
1.0	**0.845**	**0.804**	**0.965**	**0.838**	**0.863**	**0.948**
1.2	**0.954**	**0.913**	**0.993**	**0.949**	**0.955**	**0.984**
1.4	**0.987**	**0.959**	**0.999**	**0.984**	**0.986**	**0.996**
1.6	**0.997**	**0.986**	**0.999**	**0.995**	**0.996**	**0.999**
1.8	**1.000**	**0.998**	**1.000**	**1.000**	**0.999**	**0.999**
2.0	**1.000**	**0.999**	**1.000**	**1.000**	**0.999**	**1.000**
	Group Sizes (15,25)					
0.2	0.082	0.121	0.114	0.083	0.089	0.114
0.4	0.207	0.223	0.339	0.205	0.227	0.305
0.6	0.411	0.438	0.627	0.408	0.445	0.604
0.8	0.647	0.617	**0.852**	0.645	0.691	**0.827**
1.0	**0.812**	0.771	**0.948**	**0.816**	**0.848**	**0.941**
1.2	**0.937**	**0.892**	**0.989**	**0.939**	**0.953**	**0.986**
1.4	**0.980**	**0.946**	**0.996**	**0.982**	**0.987**	**0.997**
1.6	**0.996**	**0.980**	**0.999**	**0.995**	**0.997**	**0.999**
1.8	**0.999**	**0.998**	**1.000**	**0.999**	**0.999**	**1.000**
2.0	**1.000**	**0.999**	**1.000**	**0.999**	**0.999**	**1.000**

Table 6 Approximated shift parameter when power achieve 0.80.

Distribution	Sample Sizes	PM	*t*-test	MWW
Normal	(20,20)	0.93	0.92	0.95
	(15,25)	0.96	0.95	0.99
Beta	(20,20)	0.96	0.92	0.99
	(15,25)	0.99	0.95	1.05
g =0, h = 0.225	(20,20)	0.80	0.89	0.73
	(15,25)	0.82	0.91	0.75
Fleishman1 (F1)	(20,20)	0.94	0.92	0.96
	(15,25)	0.95	0.95	0.98
Fleishman2 (F2)	(20,20)	0.93	0.92	0.92
	(15,25)	0.93	0.95	0.94
χ^2_3	(20,20)	0.86	0.92	0.76
	(15,25)	0.85	0.95	0.78

Generally, the performance of the new procedure in achieving the desired power level is comparable to *t*-test and Mann-Whitney-Wilcoxon test. Under Normal and Beta distributions, the results from Table 6 show that the *t*-test achieved the desired power of 0.80 at lower shift parameter values, followed very closely by the pseudo-median procedure. The Mann-Whitney-Wilcoxon test however, required larger shift parameters to achieve the same power value. For the *g*-and-*h* distribution, the Mann-Whitney-Wilcoxon test achieved the desired power of 0.80 at the lowest shift parameters followed by the pseudo-median procedure and the *t*-test.

Under the Fleishman 1 distribution, the findings showed that the rate at which the *t*-test reached the power of 0.80 is on par with the pseudo-median procedure. However, the results indicate that the Mann-Whitney-Wilcoxon test reached 0.80 at larger shift parameters compared to the *t*-test and the new procedure. Under the Fleishman 2 distribution, all procedures are on par with each other in achieving the 0.80 level. Lastly, under the chi-square distribution, the Mann-Whitney-Wilcoxon test achieved the desired power of 0.80 faster than the pseudo-median procedure and the *t*-test for all conditions. However, the pseudo-median procedure reached 0.80 at lower shift parameters than the *t*-test.

4 Conclusions

In this paper, we investigated the performance of the new procedure, known as the pseudo-median procedure to the violations of normality. This procedure recorded empirical Type I error rates within the robustness criterion. The findings suggest that under all conditions used in this study, the pseudo-median procedure has the ability to control Type I error rates and maintaining high power rates regardless of distributional shapes. The performance of the pseudo-median procedure is compatible to the *t*-test and Mann-Whitney-Wilcoxon test. Thus, this new procedure can be considered as an alternative procedure for comparing two groups especially when the violations of assumptions of normality exist.

Acknowledgement

This work is partially funded by Universiti Sains Malaysia (USM-RU-PRGS) and supported by Universiti Utara Malaysia.

References

[1] B. L. Welch, *The significance of the difference between two means when the population variances are unequal*, Biometrika. **29**, (1938) 350–362.

[2] K. O. McGraw and S. P. Wong, *A common language effect size statistic*, Psychological Bulletin. **111**, (1992) 361-365.

[3] A. Vargha and H. D. Delaney, *A critique and improvement of the CL common language effect size statistics of McGraw and Wong*, Journal of Educational and Bahavioral Statistics. **25** (2000) 101-132.

[4] D. C. Hoaglin, *Summarizing shape numerically: The g-and h-distributions*. In D. C. Goaglin, F. Mosteller and J. W. Tukey (Eds.), Exploring data tables, trends, and shapes. (1985) 461 – 513.

[5] A. I. Fleishman, *A method for simulating non-normal distributions*, Psychometrika. **43** (1978) 521-532.

[6] SAS Institute Inc. *SAS/IML user's guide* version 8. Cary, NC: SAS Institute Inc. (1999).

[7] J. V. Bradley, *Robustness?*. British Journal of Mathematical and Statistical Psychology. **31** (1978) 321-339.

[8] K. R. Murphy and B. Myors, *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests (2nd ed.)*. Mahwah, NJ: Erlbaum.

Nor Aishah Ahad is a senior lecturer at School of Quantitative Sciences, Universiti Utara Malaysia. She received her Ph.D from Universiti Sains Malaysia in 2012. Her research interests include Applied Statistics, Robust Statistics, and Nonparametric.

Abdul Rahman Othman is a professor at School of Distance Education, Universiti Sains Malaysia. He received his Ph.D (Education) from University of California, Santa Barbara in 1995. His research interests are in the areas of Applied Statistics, Robust Statistics, and Psychometrics.

Sharipah Soaad Syed Yahaya received her Ph.D (Statistics) from Universiti Sains Malaysia in 2005. She is currently an associate professor at School of Quantitative Sciences, Universiti Utara Malaysia. She specializes in Robust Statistical Methods and Statistical Quality Control.