Comparative Performance of Pseudo-Median Procedure, Welch's Test and Mann-Whitney-Wilcoxon at Specific Pairing

Nor Aishah Ahad (Corresponding author)
UUM College of Arts and Sciences, Universiti Utara Malaysia
06010 Sintok, Kedah Malaysia
Tel: 60-19-419-3066 E-mail: aishah@uum.edu.my

Abdul Rahman Othman
Institute of Postgraduate Studies, Universiti Sains Malaysia
11800 Minden, Penang Malaysia
Tel: 60-4-653-4882 E-mail: arothman60@yahoo.com

Sharipah Soaad Syed Yahaya
UUM College of Arts and Sciences, Universiti Utara Malaysia
06010 Sintok, Kedah Malaysia
Tel: 60-4-928-6932 E-mail: sharipah@uum.edu.my

Received: July 2, 2011 Accepted: July 21, 2011 doi:10.5539/mas.v5n5p131

The research is funded by Universiti Sains Malaysia (USM-RU-PRGS) and supported by Universiti Utara Malaysia.

Abstract

The objective of this study is to investigate the performance of two-sample pseudo-median based procedure in testing differences between groups. The procedure is the modification of one-sample Wilcoxon procedure using pseudo-median of differences between group values as the central measure of location. The test was conducted on two groups setting with moderate sample sizes of symmetric and asymmetric distributions. The performance of the procedure was measured and evaluated in terms of Type I error and power rates obtained via Monte Carlo methods. Type I error and power rates of the procedure were then compared with the alternative parametric and nonparametric procedures namely the Welch's test and Mann-Whitney-Wilcoxon test. The findings revealed that the pseudo-median procedure is capable in controlling its Type I error close to the nominal level when heterogeneity of variances exists. In terms of robustness, the pseudo-median procedure outperforms the Welch's and Mann Whitney Wilcoxon tests when distributions are skewed. The pseudo-median procedure is also capable in maintaining high power rates especially for negative pairing.

Keywords: Mann-whitney-wilcoxon, Power, Pseudo-median, Type I error, Welch's test

1. Introduction

Testing the equality of central tendency (location) parameters or differences between two groups is a common statistical problem. Under traditional parametric test statistics, it is well known that Student's two-independent sample *t*-test (Student, 1908) can be highly unsatisfactory when the distribution of the data is non-normal and variances are unequal (Teh & Othman, 2009; Zimmerman, 2004; Zimmerman & Zumbo, 1993). This test also produces low power under arbitrarily small departures from normality (Keselman, Othman, Wilcox & Fradette, 2004). In cases where distributions are normal but population variances are unequal, Welch (1938) gave the solution to this problem. His solution is an approximate degrees of freedom *t* test. However, Welch's test still has problems in controlling Type I error under non-normal distributions (Algina, Oshima & Lin, 1994; Zimmerman & Zumbo, 1993).

A popular alternative for analyzing data from non-normal populations is to use nonparametric test statistics such

as Mann-Whitney-Wilcoxon test. Nonparametric statistics are insensitive to the deviation of normality. Even though nonparametric methods are distribution free, but they are not assumptions free. Usually the underlying distribution has to be symmetric (Gibbons & Chakraborti, 2003). Nonparametric procedures are more appropriate for data based on weak measurement scales and appropriate for symmetric shape (Syed Yahaya, Othman & Keselman, 2004). In addition, procedures in nonparametric statistics are less powerful than the parametric ones and therefore, require larger sample sizes to reject false hypotheses. Thus, choosing non parametric tests as alternative to the classical tests might not guarantee a reliable method due to the weakness of the tests.

To circumvent the effects of assumptions violations on the classical procedures, researchers have been advised to adopt heteroscedastic test statistics, replace the conventional methods with permutation, or transform their data to achieve normality and/or homogeneity. Some studies suggested substituting robust estimators (e.g. trimmed means and Winsorized variances) for the least square estimators (i.e. the usual mean and variance). Robustness to non-normality and variance heterogeneity in unbalanced independent group designs can be achieved by using robust estimators with heteroscedastic test statistics as demonstrated by a number of papers (Keselman, Algina, Wilcox & Kowalchuk, 2000; Keselman, Kowalchuk & Lix, 1998; Wilcox, Keselman, Muska & Cribbie, 2000). These literatures also indicated that by applying robust estimators with heteroscedastic test statistics, distortion in rates of Type I error could generally be eliminated. However, the use of trimmed means for example, required a percentage of observations to be discarded from the whole data which might cause some useful information from the data to be lost.

Over the years, many procedures were developed to handle the violation of the assumptions. However, each of the aforementioned procedures can only handle certain violations and so far, no single statistical method can be considered ideal. In this study, we proposed a statistical procedure which is based on the pseudo-median to deal with the problem of multiple violations such as non-normality, variance heterogeneity and unbalanced group sizes occurring simultaneously. This study also investigates the performance of the pseudo-medians procedure in terms of controlling Type I error and maintaining high power rates under these multiple violations. The performance of this procedure was then compared with the parametric and nonparametric tests namely the Welch's test and the Mann-Whitney-Wilcoxon test, respectively. This method optimistically, will help researchers in conducting their research in a more flexible situation without having to worry about the rigid assumptions.

The rest of the paper is organized as follows. The second section briefly explains the criteria for evaluating the performance of a statistical test. The third section elaborates on the methods used in this study. The design specifications of the data are described in the fourth section while the fifth section discusses the results. The final section concludes our study.

2. Performance Evaluation of the Statistical Test

The evaluation of any statistical test involves two attributes namely Type I error and power rates. Type I error and power rates are measured in the form of *p*-values. Type I error happens when a true null hypothesis (H_0) is incorrectly rejected. The power of a statistical test of H_0 is the probability that the H_0 will be rejected when it is false, that is, the probability of obtaining a statistically significant result or the test resulting in the conclusion that the phenomenon exists (Cohen, 1988; 1992). A procedure having its Type I error close to nominal value is considered as robust. If a procedure is able to control its Type I error rates close to the nominal value and generates good statistical power simultaneously, then the procedure is deemed to be the procedure of choice. These properties are usually used as the criteria for evaluating the performance of a statistical test.

3. Methods

The pseudo-median procedure is generated from the modification of one-sample nonparametric Wilcoxon procedure with the incorporation of pseudo-median of differences between group values as the statistic of interest in a two groups setting. As stated in Hoyland (1965), the pseudo-medians of a distribution *F* is defined as the median of the distribution of $(X_1 + X_2)/2$ where X_1 and X_2 are independently and identically distributed according to *F*. Hollander and Wolfe (1999) noted that the pseudo-median of a distribution *F* is the median of $(Z_1 + Z_2)/2$, where Z_1 and Z_2 are independent, each with the same distribution *F*.

In this procedure, suppose $X_1 = (X_{11}, X_{12}, ..., X_{1n})$ and $X_2 = (X_{21}, X_{22}, ..., X_{2m})$ be samples from distributions F_1 and F_2, respectively. Let the differences between the observations from both samples be $D_{ij} = X_{1i} - X_{2j}$, $i = 1, 2, ..., n$ and $j = 1, 2, ..., m$. The absolute value of the differences is given by $|D_{ij}|$ and R_{ij}

denote the rank of $|D_{ij}|$. An indicator function, e_{ij}, is defined as in Equation 1.1.

$$e_{ij} = \begin{cases} 0, & D_{ij} < 0 \\ 0.5, & D_{ij} = 0 \\ 1, & D_{ij} > 0 \end{cases} \tag{1.1}$$

Then the Wilcoxon statistic is defined as Equation 1.2.

$$W = \sum_{i=1}^{n} \sum_{j=1}^{m} R_{ij} e_{ij} \tag{1.2}$$

The pseudo-median is a location parameter and its value has to be estimated. The estimation is done using the Hodges-Lehmann estimator (Hollander & Wolfe, 1999). The Hodges-Lehmann estimator $(\hat{\theta})$ of pseudo-median is given in Equation 1.3 where Z_i are the differences between the observations from both samples.

$$\hat{\theta} = median\left\{ \frac{Z_i + Z_j}{2}, i \le j = 1, ..., n \right\} \tag{1.3}$$

The modification of the Wilcoxon procedure is performed by adding the pseudo-median value to all observations in the second sample. A bootstrap procedure was employed to test the hypothesis as given in Equation 1.4 where d is the pseudo-median.

$$H_0 : d = 0 \quad \text{versus} \quad H_1 : d \neq 0 \tag{1.4}$$

The algorithm of the bootstrap procedure is enumerated below.

1. Based on the two samples, find W and estimate the pseudo-median, $(\hat{d})$.
2. Shift the second sample by adding $\hat{d}$ to all members.
3. Calculate $\hat{W}$ from X_1 and new sample, $(X_2 + \hat{d})$.
4. Generate bootstrap samples from X_1 and $(X_2 + \hat{d})$ yielding X_1^* and X_2^*.
5. Calculate W^* from the bootstrap samples.
6. Calculate $(W^* - \hat{W})$.
7. Repeat step 4 to step 6 for *B* times.
8. Compare the value of $(W^* - \hat{W})$ with $(W - E(W \mid H_0))$, where $E(W \mid H_0) = [N(N+1)/4]$. Let $U = (W^* - \hat{W}) > (W - E(W \mid H_0))$ and $L = (W^* - \hat{W}) < (W - E(W \mid H_0))$ where $U = 1$ or 0 and $L = 1$ or 0.
9. Calculate the *p*-value as 2 × minimum (number of *L*, number of *U*)/*B*.

4. Design Specifications

This study focused on completely randomized design containing two groups with moderate sample size. The total sample size was set to be 40 and then split to form unbalanced design with sample sizes (15, 25), respectively. The test was conducted under heterogeneous group variances as variance heterogeneity can affect both Type I error and power of the analysis (Wilcox, Charlin & Thompson, 1986). Luh and Olejnik (1990) stated that when the population variances differ, the actual statistical power could be less than that desired. To examine the effect of variance heterogeneity on the procedure, in this study, the group variances were set to be 1:36. This ratio was chosen as it reflects extreme variance heterogeneity. This variance ratio was used by a number of researchers in their study for two groups case (Keselman, Wilcox, Lix, Algina & Fradette, 2007; Othman, Keselman, Padmanabhan, Wilcox & Fradette, 2004; Luh & Guo, 1999).

Unequal group sizes, when paired with unequal group variances, will produce either positive or negative pairings.

A positive pairing occurs when the largest group size is associated with the largest group variance, while the smallest group size is associated with the smallest group variance. On the other hand, a negative pairing referred to the case in which the largest group size is paired with the smallest group variance and the smallest group size is paired with the largest group variance. These conditions were chosen since the test for equality of central tendency parameters typically produces conservative results for the positive pairings and liberal results for the negative pairings (Syed Yahaya et al., 2004; Othman et al., 2004; Keselman et al., 2004). According to Cribbie and Keselman (2003), when variance and sample size are directly paired, Type I error estimates can be conservative and power correspondingly will be deflated. On the other hand, when variance and sample size are inversely paired, Type I error estimates can be liberal and power correspondingly will be inflated. Therefore, all the tests were examined under these two types of pairings to appraise their ability in controlling the Type I error and maintaining good power value.

In terms of distributions, we chose a g = 0, h = 0.225 (Hoaglin, 1985) distribution to represent symmetric leptokurtic and the chi-square distribution with three degrees of freedom $\left(\chi_3^2\right)$ to represent skewed leptokurtic. The former distribution has zero skewness and kurtosis equal to 154.84 while the later distribution has skewness and kurtosis equal to 1.63 and 4.0, respectively. Both distributions have positive kurtosis which indicates a peaked distribution with heavy tails. Normal distribution was used as a basis of comparison.

This study was based on simulated data. The simulation was carried out using the random-number-generating function in SAS and the simulation program was written in SAS/IML (SAS, 2006). In terms of data generation, pseudo-random standard normal variates were generated by employing the SAS generator RANDGEN and this involved the straight forward usage of the (RANDGEN(Y, 'NORMAL')). To generate the chi-square variates with three degrees of freedom, we used RANDGEN(Y, 'CHISQUARE', 3). To generate data from a g- and h-distribution, standard normal variates were converted to g- and h- variates via

$$Y = Ze^{\frac{hZ^2}{2}} \tag{1.5}$$

where Z values were generated using the generator RANDGEN with the normal distribution option.

The effect size or the shift parameter used in this study is not a single point but its values ranging from 0.2 to 2.0 with increment of 0.2 units. Therefore, for each condition, ten power values were obtained. This effect size is computed based on the common language (CL) statistics proposed by McGraw and Wong (1992) and Vargha and Delaney (2000). In this study, 0.80 was used as the standard for adequacy in power analysis. There are no hard and fast rules about how much power is enough, but according to Murphy and Myors (2004), power of 0.80 or above is usually judged to be adequate. Most power analyses specify 0.80 as the desired level, and this convention seems to be widely accepted. For each condition examined, 599 bootstrap samples were generated and 5000 data sets were simulated. The nominal level of significance was set at $\alpha = 0.05$.

5. Results and Discussion

The simulation results of Type I error for pseudo-median (PM), Welch's-test (W) and Mann- Whitney-Wilcoxon (MWW) procedures are presented in Table 1. This study uses the Bradley's (1978) liberal criterion of robustness to quantify the performance of a statistical test to control its probability of Type I error. 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 $\left[0.5\alpha, 1.5\alpha\right]$. 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. Type I error rates greater than 0.075 are considered liberal and those less than 0.025 are considered conservative.

Under normal distribution, all procedures are able to control their Type I error rates close to the nominal level of 0.05 for positive pairing. The error rates are 0.0486, 0.0492 and 0.0458 for the pseudo-median procedure, the Welch's test and the Mann-Whitney-Wilcoxon, respectively. Under negative pairing, the pseudo-median procedure shows outstanding performance in controlling the Type I error rates. The recorded rate is 0.0492, very close to the nominal value. Regardless of pairing, the pseudo-median procedure produces consistent Type I error rates under normal distribution. On the other hand, Welch's test produced Type I error rate with value of 0.0514 that is slightly greater than 0.05 but still very close to the nominal level. Under the same condition, Mann-Whitney-Wilcoxon has Type I error rate that is beyond Bradley's liberal criterion with value equal to 0.1142.

Under the g-and-h distribution, both Welch's test and pseudo-median procedure produced Type I error within the Bradley's liberal criterion. For both pairings, the pseudo-median procedure and the Welch's test produced good

and consistent Type I errors. For positive pairing, the values for the pseudo-median procedure and the Welch's test are 0.0518 and 0.0448, respectively. As for negative pairing, the value for pseudo-median is slightly inflated to 0.0532 while the value for Welch's test is consistently around 0.044. Meanwhile, the Mann-Whitney-Wilcoxon has good Type I error (0.0436) for positive pairing but very liberal Type I error (0.108) for negative pairing.

Under skewed distribution, pseudo-median procedure produced Type I error rates within the Bradley's liberal criterion. The result seems to follow the norm where positive and negative pairings typically produce smaller and larger rates, respectively. The rates of Type I error for both pairings are 0.0476 and 0.055. However, Welch's test produced Type I error considerably greater than 0.05 for both pairings with the rates equal to 0.0654 and 0.0736, however, the rates are still within the robustness criterion. Unfortunately, Mann-Whitney-Wilcoxon produced very liberal Type I error for both pairings with values equal to 0.1812 (positive pairing) and 0.2398 (negative pairing).

The last row of Table I displays the "Average" values obtain by averaging both *p*-values corresponding to each procedure and distribution. Underlined average values denote that the "Average" is within the Bradley's liberal criterion. As we can observe, regardless of distributions the "Average" values for pseudo-median procedure and Welch's test are within the robustness criterion. However, Mann-Whitney-Wilcoxon depicts liberal "Average" values for all distributions.

In statistical power analysis, we only considered procedures which were identified to be in control of Type I error rates. The comparisons of statistical power will only be meaningful if the procedures being compared are capable of controlling their rates of Type I error. The results of power analysis are tabulated in Table 2 and also illustrated in Figure 1. Table 2 is divided into two parts (above and below) based on the pairings. The first column of Table 2 represents the shift parameter used in the study. The rest of the columns record the power rates corresponding to each of the procedures tested under each type of distribution.

As we can observe from Table 2, the power rates for all the tests fail to achieve the desired level for both pairings. Between the pairings (table above and below), the comparison shows that all the procedures under positive pairings produce greater power rates than the negative pairing. When scrutinizing the results under positive pairing, the analysis reveals that under normal distribution, the power of pseudo-median procedure is just slightly below the Welch's procedure, but performs much better than Mann-Whitney-Wilcoxon procedure. However, the power of pseudo-median procedure improves under *g*-and-*h* distribution but decline again when the skewness of the distribution gets larger as shown in the second last column. Under negative pairing, even though the power values for the pseudo-median procedure slightly dropped from the positive pairing, but the procedure performs better than the Welch's test under *g*-and-*h* and chi-square distributions. Under this pairing, we did not include the Mann-Whitney-Wilcoxon because of its inability to control Type I error.

6. Conclusion

The objective of this study is to investigate the performance of the pseudo-median procedure in terms of controlling its Type I error rates and maintaining high power value. With respect to robust performance, the pseudo-median procedure is capable in controlling its Type I error close to nominal level when heterogeneity of variances exists. The pseudo-median procedure also outperforms the Welch's test and Mann-Whitney-Wilcoxon under skewed distributions. The popular Mann-Whitney-Wilcoxon is capable in controlling its Type I error only for positive pairing under symmetric distribution but fails in controlling its Type I error under asymmetric distribution. The study also reveals that pseudo-median procedure perform better than the other procedures especially under the influence of negative pairing.

References

Algina, J., Oshima, T. C. & Lin, W. Y. (1994). Type I error rates for Welch's test and James's second order test under nonnormality and inequality of variances when there are two groups. *Journal of Educational and Behavioral Statistics*. 19, 275-292. doi:10.3102/10769986019003275, http://dx.doi.org/10.3102/10769986019003275

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*. 31, 144-151.

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

Cohen, J. (1992). Statistical Power Analysis. *Current Directions in Psychological Science*. 1, 98-101. doi:10.1111/1467-8721.ep10768783, http://dx.doi.org/10.1111/1467-8721.ep10768783

Cribbie, R. A. & Keselman, H. J. (2003). Pairwise multiple comparisons: A model comparison approach versus

stepwise procedures. *British Journal of Mathematical and Statistical Psychology*. 56, 167-182. doi:10.1348/000711003321645412, http://dx.doi.org/10.1348/000711003321645412

Gibbons, J. D. & Chakraborti, S. (2003). *Nonparametric Statistical Inference* (4th ed.) Marcel Dekker, Inc. New York.

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

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

Hoyland, A. (1965). Robustness of the Hodges-Lehmann estimates for shift. *The Annals of Mathematical Statistics*. 36, 174-197. doi:10.1214/aoms/1177700281, http://dx.doi.org/10.1214/aoms/1177700281

Keselman, H. J., Algina, J., Wilcox, R. R. & Kowalchuk, R. K. (2000). Testing repeated measures hypotheses when covariance matrices are heterogeneous: Revisiting the robustness of the Welch-James test again. *Educational and Psychological Measurement*, 60, 925-938. doi:10.1177/00131640021970998, http://dx.doi.org/10.1177/00131640021970998

Keselman, H. J., Kowalchuk, R. K. & Lix, L. M. (1998). Robust nonorthogonal analyses revisited: An update based on trimmed means. *Psychometrika*, 63, 145-163. doi: 10.1007/BF02294772, http://dx.doi.org/10.1007/BF02294772

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. doi:10.1348/000711005X63755, http://dx.doi.org/10.1348/000711005X63755

Luh, W. M. & Guo, J. H. (1999). A powerful transformation trimmed mean method for one-way fixed effects ANOVA model under non-normality and inequality of variances. *British Journal of Mathematical and Statistical Psychology*. 52, 303-320. doi:10.1348/000711099159125, http://dx.doi.org/10.1348/000711099159125

Luh, W. M. & Olejnik, S. (1990). Two-stage sampling procedures for comparing means when population distributions are non-normal. ERIC Document. ED319-773.

McGraw, K. O. & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*. 111, 361-365. doi:10.1037/0033-2909.111.2.361, http://dx.doi.org/10.1037/0033-2909.111.2.361

Murphy, K. R. & Myors, B. (2004). *Statistical power analysis: A simple and general model for traditional and modern hypothsis tests* (2nd ed.) Lawrence Erlbaum Associates, Mahwah, New Jersey.

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. doi:10.1348/0007110042307159, http://dx.doi.org/10.1348/0007110042307159

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

Student (1908). The probable error of a mean. *Biometrika*. 6, 1-25.

Syed Yahaya, S. S., Othman, A. R. & Keselman, H. J. (2004). Type 1 error rates of a modified robust statistical procedure. In N. Ganikhodjaev and C. H. Pah (Eds.), *Prosiding Simposium Kebangsaan Sains Matematik ke-XII: Peranan Sains Matematik dalam Pembangunan Bioteknologi dan K-Ekonomi [Proceedings of the Twelveth National Mathematical Sciences Symposium: The Role of Mathematical Sciences in the Development of Biotechnology and K-Economy*] [CD-ROM]. Kuala Lumpur, Malaysia: International Islamic University Malaysia.

Teh, S. Y. & Othman, A. R. (2009). When does the pooled variance t-test fail? *African Journal of Mathematics and Computer Science Research*. 2, 056-062.

Vargha, A. & Delaney, H. D. (2000). A critique and improvement of the CL common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics*. 25, 101-132. doi:10.3102/10769986025002101, http://dx.doi.org/10.3102/10769986025002101

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*. 29, 350-362. doi:10.2307/2332010, http://dx.doi.org/10.2307/2332010

Wilcox, R. R., Charlin, V. L. & Thompson, K. L. (1986). New Monte Carlo result on the robustness of the

ANOVA F, W and F* statistics. *Communication in Statistics: Simulation and Computation*. 15**,** 933-943. doi:10.1080/03610918608812553, http://dx.doi.org/10.1080/03610918608812553

Wilcox, R. R., Keselman, H. J., Muska, J. & Cribbie, R. (2000). Repeated measures ANOVA: Some new results on comparing trimmed means and means. *British Journal of Mathematical and Statistical Psychology*, 53**,** 69-82. doi:10.1348/000711000159187, http://dx.doi.org/10.1348/000711000159187

Zimmerman, D. W. & Zumbo, B. D. (1993). Rank transformations and the power of the Student t test and Welch t' test for non-normal populations with unequal variances. *Canadian Journal of Experimental Psychology,* 47*,* 523 -539. doi: 10.1037/h0078850, http://dx.doi.org/10.1037/h0078850

Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology,* **57**, 173 – 181. doi: 10.1348/000711004849222, http://dx.doi.org/10.1348/000711004849222

Table 1. Type I error rates for all procedures under specific pairing

Pairings	Normal			*g*-and-*h*			χ_3^2		
	PM	W	MWW	PM	W	MWW	PM	W	MWW
Positive	**0.0486**	**0.0492**	**0.0458**	**0.0518**	**0.0448**	**0.0436**	**0.0476**	**0.0654**	0.1812
Negative	**0.0492**	**0.0514**	0.1142	**0.0532**	**0.044**	0.108	**0.055**	**0.0736**	0.2398
Average	0.0489	0.0503	0.08	0.0525	0.0444	0.0758	0.0513	0.0695	0.2105

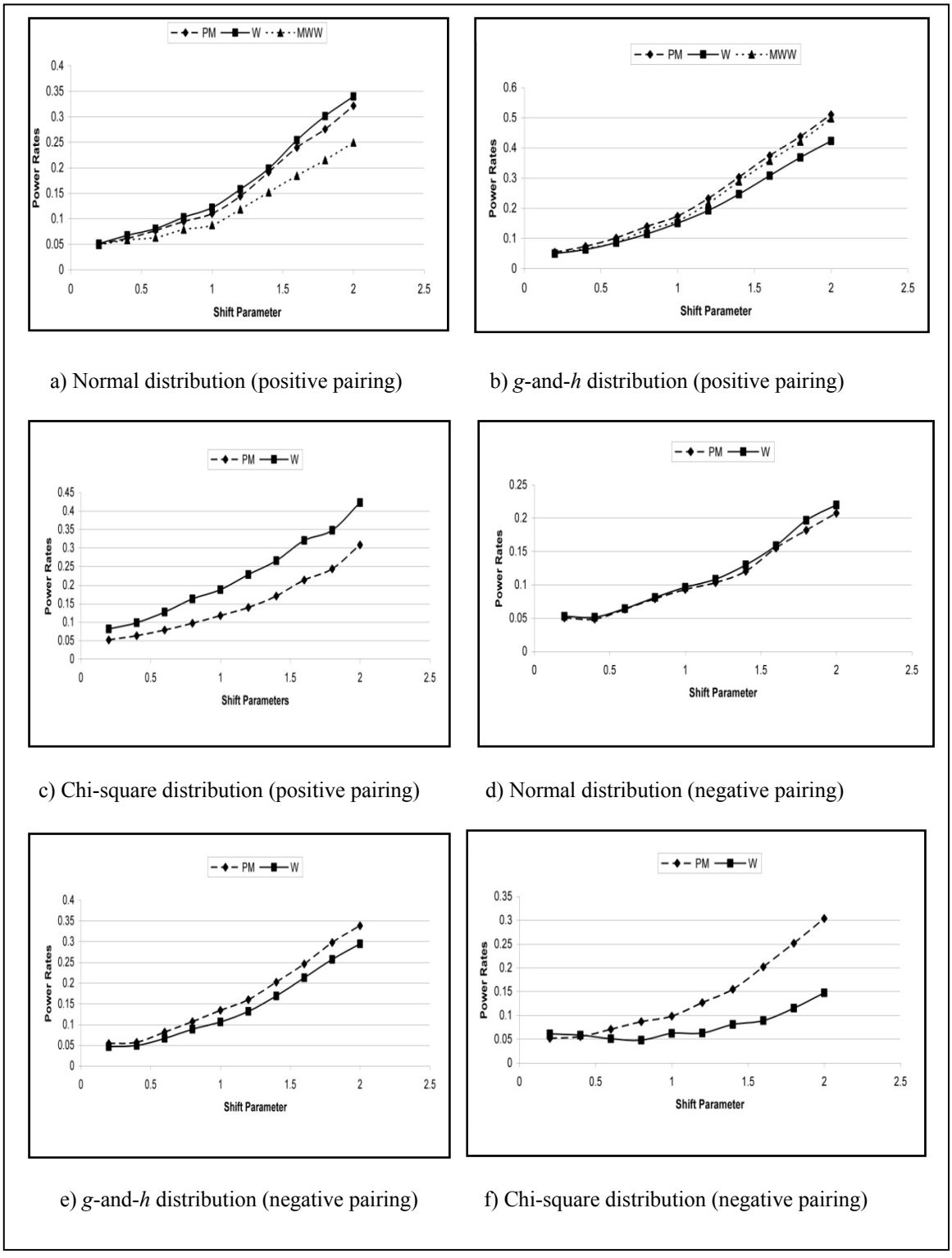
Bold values indicate Type I error within [0.025, 0.075]

Table 2. Power rates for all procedures under specific pairings

Shift Parameter	Group sizes (15, 25) and group variances (1:36) -- Positive pairing							
	Normal			*g*-and-*h*			χ_3^2	
	PM	W	MWW	PM	W	MWW	PM	W
0.2	0.0502	0.0508	0.0484	0.0546	0.0498	0.0500	0.0520	0.0816
0.4	0.0614	0.0672	0.0582	0.0738	0.0630	0.0660	0.0632	0.0986
0.6	0.0766	0.0804	0.0630	0.1016	0.0858	0.0862	0.0790	0.1274
0.8	0.0944	0.1026	0.0786	0.1395	0.1150	0.1292	0.0968	0.1630
1.0	0.1098	0.1216	0.0872	0.1740	0.1506	0.1576	0.1176	0.1880
1.2	0.1438	0.1576	0.1180	0.2322	0.1924	0.2162	0.1398	0.2288
1.4	0.1912	0.1982	0.1516	0.3028	0.2468	0.2884	0.1704	0.2660
1.6	0.2392	0.2542	0.1842	0.3746	0.3080	0.3574	0.2138	0.3210
1.8	0.2754	0.3014	0.2150	0.4374	0.3684	0.4210	0.2440	0.3484
2.0	0.3212	0.3396	0.2494	0.5104	0.4226	0.4976	0.3086	0.4230

Shift Parameter	Group sizes (15, 25) and group variances (36:1) -- Negative pairing					
	Normal		*g*-and-*h*		χ_3^2	
	PM	W	PM	W	PM	W
0.2	0.0504	0.0528	0.0544	0.0464	0.0520	0.0614
0.4	0.0486	0.0514	0.0574	0.0496	0.0558	0.0586
0.6	0.0634	0.0642	0.0818	0.0668	0.0708	0.0512
0.8	0.0792	0.0812	0.1072	0.0890	0.0870	0.0482
1.0	0.0932	0.0962	0.1344	0.1064	0.0986	0.0622
1.2	0.1032	0.1082	0.1600	0.1318	0.1270	0.0632
1.4	0.1202	0.1300	0.2022	0.1690	0.1552	0.0812
1.6	0.1554	0.1588	0.2462	0.2132	0.2022	0.0894
1.8	0.1820	0.1968	0.2978	0.2576	0.2518	0.1154
2.0	0.2076	0.2198	0.3384	0.2948	0.3036	0.1474

ISSN 1913-1844 E-ISSN 1913-1852



a) Normal distribution (positive pairing)

b) *g*-and-*h* distribution (positive pairing)

c) Chi-square distribution (positive pairing)

d) Normal distribution (negative pairing)

e) *g*-and-*h* distribution (negative pairing)

f) Chi-square distribution (negative pairing)

Figure 1. Power Curves for all distributions under specific pairing