Sains Malaysiana 43(7)(2014): 1095–1100

Sensitivity Analysis of the Refinement to the Mann-Whitney Test

(Analisis Kepekaan Penghalusan kepada Ujian Mann-Whitney)

ABDUL RAHMAN BIN OTHMAN & LAI CHOO HENG*

ABSTRACT

The aim of researchers when comparing two independent groups is to collect large normally distributed samples unless they lack the resources to access them. In these situations, there are a myriad of non-parametric tests to select, of which the Mann Whitney U test is the most commonly used. In spite of its great advantages of usage, the U test is capable of producing inflated Type I error when applied in situation of heterogeneity or distinct variances. This current study will present a viable alternative called the refined Mann-Whitney test (RMW). A Monte Carlo evaluation test is conducted on RMW using artificial data of various combinations of extreme test conditions. This study reviews that the RMW test justified its development by enhancing the performance of the U test. The RMW test is able to control well its Type I error rates even though it has a lower power.

Keywords: Mann-Whitney test; Monte Carlo; power; refined Mann-Whitney test; Type I error

ABSTRAK

Semasa membuat perbandingan dua kumpulan tak bersandar, matlamat penyelidik ialah untuk mengumpul sampel taburan normal yang besar. Jika menghadapi kekurangan sumber untuk mencapai matlamat tersebut, terdapat pelbagai pilihan ujian tak-berparameter dengan ujian U Mann-Whitney paling sering dipilih. Sungguhpun ujian Mann-Whitney mempunyai banyak kelebihan daripada segi kegunaan, tetapi ia berkemungkinan menghasilkan ralat jenis I yang tinggi semasa digunakan dalam situasi beza varians yang ketara. Kajian ini akan memperkenalkan satu ujian alternatif yang dikenali sebagai ujian pengubahsuaian Mann-Whitney (RMW). Satu ujian penilaian Monte Carlo dijalankan dengan menggunakan data tiruan daripada pelbagai gabungan situasi pengujian yang ekstrem. Kajian ini mendapati ujian RMW menepati langkah pengubahsuaian demi penambahbaikan prestasi ujian Mann-Whitney. Ujian RMW dapat mengawal ralat jenis I dengan baik sungguhpun kuasanya rendah.

Kata kunci: Kuasa; Monte Carlo; ralat jenis I; ujian Mann-Whitney; ujian pengubahsuaian Mann-Whitney

INTRODUCTION

The two-sample location problem assumes normality and homogeneity of variance. Even though the most widely used Welch test, designed to handle variance heterogeneity, requires the underlying populations be normal. The significance level of the Welch test is biased by variance heterogeneity especially when sample sizes are unequal (Overall et al. 1995; Scheffe 1959). Similarly, its non-parametric counterpart, the Mann-Whitney test requires the populations to be symmetrical. However, violations of these assumptions are common, thus rendering these statistical tests inappropriate. Subsequent efforts are focus on developing statistical test where its performance is not constraint by assumptions. Babu and Padmanabhan (2002) proposed a robust solution to the non-parametric Behren-Fisher problem. The proposed test is a refinement of the Mann Whitney, RMW that does not require the symmetrical assumption of the underlying population. Lai (2009) reviewed the general robustness of the RMW test in a combination of test conditions from symmetrical distributions. The refinement test is only appropriate in a balanced homogeneous group samples from skewed distributions like exponential distribution and lognormal distribution.

This study conducted a sensitivity analysis to gain a further insight into the robustness of the refinement test especially in extreme heterogeneity in a higher unbalanced group sample sizes. The sensitivity analysis was conducted by using computer simulated data to assess the distributional performance of the RMW test across a variety of test conditions (sample size, variance ratio and underlying distribution). A SPSS syntax program was design to generate independent samples from various types of population at various combinations of test conditions to be performed the RMW test, and Mann Whitney test. The random number seed used when generating the artificial samples for both procedures is the SPSS's default seed of 2000000. The outcomes of this sensitivity analysis will provide a comprehensive robust coverage of the RMW test that will enable practitioners effectively employed them. This study will also review the boundary test conditions for the classical and Mann-Whitney test.

Refinement of The Mann Whitney Test

This section will present the theoretical framework of the proposed refined Mann Whitney test. This refinement test incorporated a bootstrap test for determining the critical values of the test statistics. The formula for the refined Mann-Whitney tests is presented as follows.

With reference to the two-sample problems, let $X = (X_1, \ldots, X_m)$ and $Y = (Y_1, \ldots, Y_n)$ be the two independent samples from their respectively continuous distributions, F_1 and F_2. Briefly, $F_1(x) = F\left(\frac{x - M_X}{\sigma_X}\right)$ and $F_2(y) = F\left(\frac{y - M_Y}{\sigma_Y}\right)$ with median $M_{\backslash X}$ and M_Y, respectively.

Hence, F is an arbitrary continuous distribution with median zero and the variances, σ_X and σ_Y are possibly unequal.

Ideally, when $\sigma_X^2 = \sigma_Y^2$, the performance of the classical Mann-Whitney test is robust in both validity and effectiveness. The test is distribution free and the Mann-Whitney statistics, W is extensively tabulated. Even though when $\sigma_X^2 \neq \sigma_Y^2$, the Mann-Whitney test is still capable of maintaining its robust performance provided the underlying distributions are symmetrical. Under the null hypothesis of equal medians, the symmetry assumption of the underlying population distribution will ensure that $p = P(X_1 \leq Y_1) = 0.5$ and hence $E(W) = 0.5\ mn$.

The proposed RMW test is designed to do away with the symmetry assumption. When the symmetry assumption is violated, then value of p is no longer equal to 0.5 and its distribution is unknown. Consequently, W will not be centred at 0.5. The RMW test proposed to centre the value of $\left(W\middle/mn\right)$ at $\hat{p}$, an estimator of $p = P(X_1 \leq Y_1)$. Since the distribution property of $\hat{p}$ is unknown, the RMW test employed bootstrap percentile test to obtain its critical values. The RMW test proposed a different approach of computing the critical values.

Initially, the refinement test will align the two samples for location and scale. For location-alignment, being non-parametric in nature, the obvious choices were the sample medians, respectively, $\tilde{X}$ and $\tilde{Y}$. There were many choices for scale-alignment of the two samples and the current study uses the standard deviations, s_X and s_Y. The rational is that the best result for the statistics of U were obtained when the scales s_X and s_Y are used (Babu & Padmanabhan 2002). After alignment, both the aligned samples are combined and subsequently used to compute the test statistics, T.

Let the combined aligned sample $(z_1, \ldots, z_{m+n})$ be defined by:

$$z_i = \left(X_i - \tilde{X}\right)\Big/ s_X, \qquad if\ 1 \leq i \leq m$$

$$= \left(Y_{i-m} - \tilde{Y}\right)\Big/ s_Y, \qquad if\ m < i\ m + n.$$

By denoting $Q = m + n$, the estimator of p and the test statistics, T are determine using the following equations.

$$\hat{p} = \frac{1}{Q^2}\sum_{i=1}^{Q}\sum_{j=1}^{Q} I_{\{z_i s_X \leq z_j s_Y\}} \text{ and } T = \sqrt{n}\left(U - \hat{p}\right).$$

The RMW test proposed a bootstrap test to obtain the critical values for the T statistic.

Let $z_1^*, \ldots, z_Q^*$ be a bootstrap sample obtained from z_1, …, z_Q. Let $X_i^* = z_i^* s_X$, $i = 1, \ldots, m$ and $Y_j^* = z_{j+m}^* s_Y$, $j = 1, \ldots, n$ be the two samples obtained after splitting. Next the relevant values of U^*, p^* and T_1^* are computed using

$$U^* = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} I_{\left(X_i^* < Y_j^*\right)},$$

$$p^* = \frac{1}{Q^2}\sum_{i=1}^{n}\sum_{j=1}^{m} I_{\left(z_i^* s_X^* < z_j^* s_Y^*\right)} \text{ and}$$

$$T_1^* = \sqrt{n}\left(U^* - p^*\right),$$

with s_X^* and s_Y^* being the sample standard deviation of X^* and Y^* respectively.

Methods

The primary objective of this study was to evaluate the performance of the RMW test across various combinations of data conditions. This study is an extension of the earlier work on the RMW test. The selected conditions include higher unbalanced group sample sizes, extreme heterogeneity variance ratios and symmetrical underlying population distributions. For each distribution, there are four sets of sample sizes used ranging from 10 to 25, four different variance ratios. The underlying distributions identified are symmetrical distributions that are leptokurtic, mesokurtic and platykurtic. The two independent samples were obtained from population with the same mean and same variability. This study comprised of a thousand replications of the simulation test in order to obtain accurate estimates of the Type I error rates.

TYPE OF UNDERLYING DISTRIBUTIONS

This study involved the generation of samples/groups from distributions of specified shape to be used as synthetic data to study empirically the behaviours of refinement test in a controlled situation. The values of skewness and kurtosis (γ_1, γ_2) were used to show the different distributions and also allow examination of the extremes of the statistical tests in this study. Accordingly, Balakrishnan and Nevzoros (2003) indicates that distributions with $\gamma_2 > 3$ are leptokurtic distributions; those with $\gamma_2 < 3$ are platykurtic distributions; those with $\gamma_2 = 3$ are mesokurtic distributions (including the normal distribution). Distributions with $\gamma_1 = 0$ are symmetrical whereas $\gamma_1 > 0$ are considered as positively skewed and those with $\gamma_1 < 0$ are negatively skewed. The types of distributions proposed in this study were symmetrical distributions. There were three symmetrical distributions selected for the simulation study. The symmetrical distributions are the standard normal distribution, uniform distribution and laplace distribution. The values of skewness and kurtosis of the selected distributions are tabulated as in Table 1.

SAMPLE DESIGN

The independent sample sizes used in this study was (n_1, n_2) = (10, 10), (10, 15), (15, 25) and (25, 25). These sample sizes were then systematically manipulated to incorporate the various combinations of variance ratios. The variance ratios $\left(\sigma_1^2 : \sigma_2^2\right)$ identified were (1:9), (1:16) and (1:64) to study the sensitivity of the tests to variance heterogeneity. Crossing these two factors yielded 12 sample size combinations for each distribution. These sample sizes were generated from each of the three distributions using *RV* subroutine of the SPSS syntax. Thus the total numbers of data set are 3 × 12 = 36.

MONTE CARLO ASSESSMENT OF TYPE I ERROR

The extent to which a test controls its Type I error rate depends on how well the underlying assumptions of the test are satisfied by the data and the sensitivity of the test to the departure from these assumptions. If it is found that departures from the underlying assumptions do not seriously impair the distributional properties of the test, then the test is robust. The framework for examining these assumptions is how well the statistical model (i.e. the test) fits the observed data. A good fit implies the test should be able to control its Type I error at nominal level whereas a poor fit indicates otherwise. To evaluate the performance of a statistical test in controlling its Type I error rate, this research uses a method of reporting to which extend a test statistic disagrees with the null hypothesis. This measure of disagreement is called the observed significance level or simply *p*-value of the test. By definition, *p*-value for a specific test is the probability (assuming H_o is true) of observing a value of the test statistic that is as least as contradictory to the null hypothesis and supportive of the alternative hypothesis. In the Monte Carlo assessment of Type I error, the various types of data sets produced from the systematic manipulation of the group sample size, the degree of inequalities of the population variances and the underlying shape of the distribution is subsequently used to compute their respectively *p*-value. The two groups are sampled from populations that conformed to the null hypothesis. As there were 1000 vectors per data sets, there were 1000 replications of *p*-values. These *p*-values will be used to determine the empirical estimate of the Type I error for a non-directional two sample hypothesis test.

ROBUST CRITERIA

In order to quantify the performance of the refinement tests in controlling its probability of Type I error, this study uses the Bradley's (1978) liberal criterion for robustness. According to Bradley's (1978) liberal criterion of robustness, a test can be considered robust if its empirical rate of Type I error, α_e is within the interval $0.5\alpha \leq \alpha_e \leq 1.5\alpha$ where α is the significance level. Hence, for the non-directional Type I error rate, at the nominal level of significance, $\alpha = 0.05$ the empirical Type I error rate is acceptable if it is located within the bounds of 0.025 and 0.075. This represent good Type I error rate control and the test is considered robust. If the estimated error rates are outside the robust interval, they are either conservative if the rates are below the lower bound or otherwise liberal.

POWER ANALYSES

The power of a test is the ability of the test to detect an actual difference between the means of populations. Power analyses will investigate the power of a test for each of the combination of sample sizes and variance ratios. Hence, simulating datasets for power analysis will take into consideration the group sample size, variance ratio and effect size. For the power analysis, 1000 vectors per datasets will be generated for each combinations of test conditions. The power analysis is based on commonly applied effect size measurement that is the Cohen's standardized effect size, *d* (Cohen 1969). The mean difference between the two sample sizes is computed and the reported power rate is 0.5. Table 2 shows the means and effect sizes for the various combinations of group sample sizes and variance ratios. The corresponding values of *d* will have a medium to large size of effect.

RESULTS

The results obtained from Monte Carlo simulation on the tests are reported in Tables 3 - 4. These tables contained tabulated simulation results of Type I error and power analysis. The entries in Table 3 are empirical *p*-values for both tests under various combinations of test conditions. The shaded entries are indication that the particular empirical Type I error rates are within the Bradley's (1987) liberal criteria. Table 3 shows that the RMW test is capable of maintaining its Type 1 error rates across all combination of test conditions. The MW test is not as robust since there are instances where the rates are inflated.

TABLE 1. Skewness and kurtosis values of the distributions used in the simulation

Skewness (γ_1)	Kurtosis(γ_2)		
	Platykurtic	Mesokurtic	Leptokurtic
Symmetric	$\gamma_1 = 0, \gamma_2 = 1.8$ Uniform (-1,1)	$\gamma_1 = 0, \gamma_2 = 3$ Normal (0,1)	$\gamma_1 = 0, \gamma_2 = 6.0$ Laplace (0,1)

TABLE 2. Group sample means, effect sizes and sizes of effect for power analysis

Group sample size (n_1:n_2)	Variance ratio $\sigma_1^2:\sigma_2^2$	Shift parameter	*d*	Size of effect
10:10	1:9	2.07	0.92	Large
	1:16	2.70	0.93	
	1:64	5.28	0.93	
10:15	1:9	1.87	0.84	Large
	1:16	2.44	0.84	
	1:64	4.76	0.84	
15:25	1:9	1.47	0.66	Medium
	1:16	1.92	0.66	
	1:64	3.74	0.66	
25:25	1:9	1.27	0.57	Medium
	1:16	1.65	0.57	
	1:64	3.23	0.57	

TABLE 3. Empirical Type I error rates

Group sample size	Variance ratio	Distribution	RMW	MW
(10,10)	1:9	Normal	0.045	0.050
		Laplace	0.052	0.061
		Uniform	0.047	0.061
	1:16	Normal	0.033	0.063
		Laplace	0.055	0.067
		Uniform	0.041	0.067
	1:64	Normal	0.047	0.074
		Laplace	0.043	0.082
		Uniform	0.050	0.082
(10,15)	1:9	Normal	0.053	0.038
		Laplace	0.056	0.052
		Uniform	0.051	0.040
	1:16	Normal	0.053	0.040
		Laplace	0.058	0.049
		Uniform	0.052	0.044
	1:64	Normal	0.042	0.048
		Laplace	0.056	0.052
		Uniform	0.047	0.045
(15,25)	1:9	Normal	0.056	0.043
		Laplace	0.049	0.036
		Uniform	0.062	0.049
	1:16	Normal	0.049	0.044
		Laplace	0.054	0.037
		Uniform	0.052	0.056
	1:64	Normal	0.054	0.053
		Laplace	0.061	0.044
		Uniform	0.051	0.056
(25,25)	1:9	Normal	0.067	0.077
		Laplace	0.068	0.067
		Uniform	0.062	0.079
	1:16	Normal	0.059	0.080
		Laplace	0.061	0.066
		Uniform	0.061	0.082
	1:64	Normal	0.044	0.090
		Laplace	0.049	0.077
		Uniform	0.050	0.085

TABLE 4. Power rates

Group sample size	Variance ratio	Distribution	RMW	MW
(10,10)	1:9	Normal	0.253	0.507
		Laplace	0.215	0.396
		Uniform	0.489	0.835
	1:16	Normal	0.241	0.525
		Laplace	0.232	0.412
		Uniform	0.433	0.798
	1:64	Normal	0.225	0.528
		Laplace	0.211	0.453
		Uniform	0.376	0.774
(10,15)	1:9	Normal	0.261	0.518
		Laplace	0.211	0.422
		Uniform	0.361	0.801
	1:16	Normal	0.234	0.525
		Laplace	0.214	0.437
		Uniform	0.397	0.769
	1:64	Normal	0.214	0.496
		Laplace	0.211	0.458
		Uniform	0.456	0.765
(15,25)	1:9	Normal	0.290	0.516
		Laplace	0.234	0.445
		Uniform	0.413	0.738
	1:16	Normal	0.279	0.511
		Laplace	0.234	0.445
		Uniform	0.376	0.703
	1:64	Normal	0.284	0.507
		Laplace	0.249	0.463
		Uniform	0.542	0.765
(25,25)	1:9	Normal	0.340	0.513
		Laplace	0.175	0.450
		Uniform	0.542	0.765
	1:16	Normal	0.312	0.505
		Laplace	0.297	0.476
		Uniform	0.494	0.748
	1:64	Normal	0.280	0.514
		Laplace	0.301	0.510
		Uniform	0.467	0.721

Table 4 contains simulation results of power analysis. The power rates of RMW test are generally higher than RMW test across all the test conditions. Hence, the RMW test is relatively the more powerful test.

CONCLUSION

This research reviewed that the RMW test justified its development as it is capable of improving the performance of the original MW test. The MW test is found to be appropriate but only if there is no severe variance heterogeneity observed in the data. Past studies (Robert & Casella 2004; Zimmerman 1987) have also shown that the MW produced inflated Type I error rates when applied in situations of heterogeneity or distinct variances. Therefore, the MW test must be used with caution. The RMW test is a promising test as it is capable controlling its Type I error rates when handling multiple test assumptions violations. Despite the relatively lower power rates, the RMW test is a viable alternative. When severe variance heterogeneity is suspected, it is therefore more reliable to use the RMW test.

REFERENCES

Babu, G.J. & Padmanabhan, A.R. 2002. Re-sampling methods for the non-parametric Behrens-Fisher problem. *Sankhya Service A* 3: 678-692.

Balakrishnan, N. & Nevzoros, V.B. 2003. *A Primer on Statistical Distributions*. Hoboken, New York: A John Wiley & Sons, Inc.

Bradley, J.V. 1978. Robustness? *British Journal of Mathematics and Statistical Psychology* 31: 144-151.

Cohen, J. 1969. *Statistical Power Analyses for the Behavioral Sciences*. New York and London: Academic Press.

Lai Choo Heng. 2009. Evaluation of the two refined Mann-Whitney procedures. Doctoral dissertation, Universiti Sains Malaysia, Pulau Pinang, Malaysia (Unpublished).

Overall, J.E., Atlas, R.S. & Gibson, J.M. 1995. Tests that is robust against variance heterogeneity in k x 2 designs with unequal cell frequencies. *Psychological Reports* 76: 1011-1017.

Robert, C.P. & Casella, G. 2004. *Monte Carlo Statistical Methods*. 2nd ed. New York: Springer – Verlag.

Scheffe, H. 1959. *The Analysis of Variance*. New York: Wiley.

Zimmerman, D.W. 1987. Comparative power of student *t* test and Mann Whitney *U* test for unequal sample sizes and variances. *Journal of Experimental Education* 55: 172-174.

Robust Statistics Computational Laboratory
Pusat Pengajian Pendidikan Jarak Jauh
Universiti Sains Malaysia
11800 USM, Pulau Pinang
Malaysia

*Corresponding author; email: lchooheng@yahoo.com

Received: 5 February 2013
Accepted: 2 October 2013