Int. J. Anal. Appl. (2023), 21:36

International Journal of Analysis and Applications

A Comparison of Nonparametric Statistics and Bootstrap Methods for Testing Two Independent Populations with Unequal Variance

Wandee Wanishsakpong, Kantima Sodrung, Ampai Thongteeraparp*

Department of Statistics, Faculty of Science, Kasetsart University, Bangkok 10900, Thailand

**Corresponding author:* fsciamu@ku.ac.th

Abstract. The common parametric statistics used for testing two independent populations have often required the assumptions of normality and equal variances. Nonparametric tests have been used when assumptions of parametric tests cannot be achieved. However, some studies found nonparametric tests to be too conservative and less powerful than parametric tests. Bootstrap methods are also alternative tests when assumptions of parametric tests are violated, but they have small size limitations. Later, nonparametric tests when pooled with the bootstrap methods may overcome the powerful test and small sample sizes issue. Thus, the purpose of this study was to apply the bootstrap method together with nonparametric statistics and compare the efficiency of nonparametric tests and bootstraps methods when pooled with nonparametric tests for testing the mean difference between two independent populations with unequal variance. The Yuen Welch Test (YW), Brunner-Munzel Test (BM), Bootstrap Yuen Welch Test (BYW) and Bootstrap Brunner-Munzel Test (BBM) were studied via Monte Carlo simulation with non-normal population distributions. The results show that the probability of a type I error of all four test statistics could be controlled for all situations. The Brunner-Munzel test (BM) had the highest power and the best efficiency in the case of mean difference ratio increases. The Bootstrap Yuen Welch Test (BYW) had the highest power when the sample size was small.

1. Introduction

To test the mean differences between two independent populations, a well-known test statistic is the t-test. It is a parametric statistic that requires the assumption of normality (Fagerland & Sandvik, 2009). In practice, the analyzed data may violate the normality assumption. This will affect the performance of the test statistic. Nonparametric statistics are alternative method when the data

Received: Feb. 28, 2023.

2010 *Mathematics Subject Classification.* 62P30.

Key words and phrases. bootstrap method; nonparametric statistics; Yuen Welch test; Brunner-Munzel test.

https://doi.org/10.28924/2291-8639-21-2023-36
ISSN: 2291-8639

does not meet the assumption of normality. Three popular nonparametric statistics were used to test the population means: Yuen Welch test (Yuen, 1974), Rank Welch test (Zimmerman & Zumbo, 1993) and the Brunner-Munzel test (Brunner & Munzel, 2000).
In addition, the bootstrap method was combined with nonparametric statistics using resampling technique (Efron & Tibshirani, 1993). Noppadon & Chinnapong (2010) studied and compared the Brunner-Munzel test, the Bootstrap Brunner-Munzel test and the Bootstrap Rank Welch test for unequal variances with lognormal distributions. The Brunner-Munzel test had the ability to control for the probability of a type I error in all cases and had superior power. Dwivedi et al. (2017) found that the Nonparametric Bootstrap t test had power equal to or greater than the t-test, Welch t test, Wilcoxon rank sum test or the Permutation test.

In this study, the researcher is interested in applying the bootstrap method with nonparametric statistics and then comparing the power of the tests. The performance of four tests: Yuen Welch test (YW), Brunner-Munzel test (BM), Bootstrap Yuen Welch test (BYW) and Bootstrap Brunner-Munzel (BBM) are studied and compared under the Laplace distribution and the lognormal distribution for testing the mean differences between two independent populations with unequal variance.

2. Statistical Methods

2.1. **Yuen Welch test statistic (YW).** The Yuen Welch test statistic was proposed by Yuen (1974). It was developed from the t-test by calculating the trimmed mean. The amount of trimming for this study is 20% which is the best option (Wilcox, 2005). Thus the 20% smallest and 20% largest observations in each sample are removed (Fagerland & Sandvik, 2009).

The Yuen Welch test Statistic is

$$YW = \frac{\bar{X}_\gamma - \bar{Y}_\gamma}{\sqrt{d_X + d_Y}} \tag{2.1}$$

where $d_X = \frac{SW_X^2(n_1-1)}{h_X(h_X-1)}$, $d_Y = \frac{SW_Y^2(n_2-1)}{h_Y(h_Y-1)}$
with degree of freedom $df_{YW} = \frac{(d_X+d_Y)^2}{[\frac{d_X^2}{h_X-1}+\frac{d_Y^2}{h_Y-1}]^2}$, where

$\bar{X}_\gamma$ and $\bar{Y}_\gamma$ are the mean of the sample after trimming,
d_X and d_Y are estimates of the squared standard errors,
h_X and h_Y are the number of observations remaining in the samples x and y after trimming,
SW_X^2 and SW_Y^2 are the samples windsorized variances,
n_1 and n_2 are the sample sizes.

2.2. **Brunner-Munzel test Statistic (BM).** Brunner and Munzel (2000) proposed the Brunner-Munzel test which is associated with the midrank. It is an update of the Wilcoxon-Mann-Whitney

test statistic using the sample variance as a modifier and using the degrees of freedom as proposed by Satterwaite-Smith-Welz. The Brunner-Munzel test statistic is

$$BM = \frac{n_1 n_2(\bar{R}_2 - \bar{R}_1)}{n_1 + n_2\sqrt{n_1 S_1^2 + n_2 S_2^2}}. \tag{2.2}$$

The distribution of the Brunner-Munzel test can be approximated by a t-distribution with degrees of freedom

$$df_{BM} = \frac{(n_1 S_1^2 + n_2 S_2^2)^2}{\frac{(n_1 S_1^2)^2}{n_1 - 1} + \frac{(n_2 S_2^2)^2}{n_2 - 1}}$$

when

$\bar{R}_1$ and $\bar{R}_2$ are the mean of rank associated with sample X and Y when data are pooled,
n_1 and n_2 are the sample sizes,
S_1^2 and S_2^2 are the variance of rank associated with sample X and Y with replacement.

2.3. **The Bootstrap method.** The Bootstrap method is a statistical procedure that uses the principle of resampling to generate a new sample from a single random sample. (Efron & Tibshirani, 1993). In this research, two bootstrap methods with nonparametric statistics are applied as follows:

2.3.1. *Bootstrap Yuen Welch test (BYW).*

(i) Let $X = X_1, X_2, \cdots, X_{n_1}$ is the observed sample 1 of size n_1 and $Y = Y_1, Y_2, \cdots, Y_{n_2}$ is the observed sample 2 of size n_2.
(ii) Evaluate test statistic: YW as in equation (2.1).
(iii) Return sampling from X and Y of size n_1 and n_2 denoted by X^* and Y^* respectively.
(iv) Evaluate test statistic: YW^* as follows:

$$YW^* = \frac{\bar{X}_\gamma^* - \bar{Y}_\gamma^*}{\sqrt{d_X^* + d_Y^*}} \tag{2.3}$$

(v) Repeat step (iii) and (iv) for 1000 times
(vi) Approximate $p - value = \frac{\text{number of times}(|YW^*| \geq |YW|)}{1{,}000}$

2.3.2. *Bootstarp Bunner-Munzel Test (BBM).*

(i) Let $X = X_1, X_2, \cdots, X_{n_1}$ is the observed sample 1 of size n_1 and $Y = Y_1, Y_2, \cdots, Y_{n_2}$ is the observed sample 2 of size n_2.
(ii) Evaluate test statistic: BM as in equation (2.2).
(iii) Draw two bootstrap samples with replacement: one of size n_1 and n_2 denoted by X^* and Y^* respectively.
(iv) Evaluate test statistic: BM^* as follows:

$$BM^* = \frac{n_1^* n_2^*(\bar{R}_2^* - \bar{R}_1^*)}{n_1^* + n_2^*\sqrt{n_1^* s_1^{2^*} + n_2^* s_2^{2^*}}}. \tag{2.4}$$

(v) Repeat steps (iii) and (iv) for 1,000 times.

(vi) Approximate $p-value = \frac{\text{number of times}(|BM^*| \geq |BM|)}{1{,}000}$

3. Research Method

The simulation data are generated by R programming with Monte Carlo technique to compare the efficiency of four test statistics for the following situations:

(1) Generate two independent populations into two distributions.
   (a) Population with Laplace distribution
   (b) Population with Lognormal distribution
(2) Determine the sample sizes for both equal and unequal sizes:
   (a) Equal sample size (n_1, n_2) is (10,10), (20,20), (30,30), (50,50) and (100,100)
   (b) Unequal sample size (n_1, n_2) is (10,15), (10,20), (30,40), (45,50) and (50,100)
(3) Determine the means of two population groups as follows:
   (a) The means of two populations are not different for evaluating the probability of a type I error.
   (b) The means of two populations are different with the ratio 1.5:1 and 2:1 for evaluating the power of the test.
(4) Determine the variance ratios of two populations for unequal variances, which were 3:1 and 5:1 (Ratiwat, 2019).
(5) The significance level of the test is 0.05.
(6) The number of replications for each condition are 5,000 and bootstrap numbers are based on 1,000 replications.

4. Results

The results of the simulation are as follows:

4.1. **Probability of Type I Error.** The ability to control for the type I error of the four test statistics is considered based on criteria from Bradley (1978). At a significance level of 0.05, the test statistic with a probability of a type I error between [0.025 - 0.075] is considered as able to control for the probability of a type I error. The probability of a type I error of the four test statistics had the following details: From the tables 1 and 2, it would be found that when the population had Laplace and Lognormal distributions, and the ratios of variance are 3:1 and 5:1, all four test statistics were able to control for the probability of a type I error for both equal and unequal sample sizes.

Distribution	Variance ratio	Sample size	Probability of Type I error			
			Nonparametric		Bootstrap	
			YW	BM	BYW	BBM
Laplace	3:1	(10,10)	0.0384	0.0540	0.0344	0.0384
		(20,20)	0.0474	0.0572	0.0364	0.0410
		(30,30)	0.0472	0.0494	0.0464	0.0376
		(50,50)	0.0494	0.0492	0.0272	0.0428
		(100,100)	0.0464	0.0458	0.0368	0.0334
	5:1	(10,10)	0.0458	0.0620	0.0332	0.0418
		(20,20)	0.0486	0.0524	0.0322	0.0372
		(30,30)	0.0486	0.0548	0.0288	0.0358
		(50,50)	0.0478	0.0504	0.0376	0.0382
		(100,100)	0.0524	0.0530	0.0336	0.0344
Lognormal	3:1	(10,10)	0.0454	0.0550	0.0300	0.0458
		(20,20)	0.0492	0.0536	0.0290	0.0368
		(30,30)	0.0430	0.0488	0.0370	0.0382
		(50,50)	0.0552	0.0498	0.0376	0.0300
		(100, 100)	0.0580	0.0500	0.0348	0.0298
	5:1	(10,10)	0.0522	0.0600	0.0402	0.0462
		(20,20)	0.0406	0.0518	0.0412	0.0338
		(30,30)	0.0466	0.0522	0.0290	0.0298
		(50,50)	0.0536	0.0548	0.0388	0.0316
		(100, 100)	0.0450	0.0512	0.0450	0.0392

Table 1. Probability of type I error for four test statistics with equal sample size

Distribution	Variance ratio	Sample size	Probability of Type I error			
			Nonparametric		Bootstrap	
			YW	BM	BYW	BBM
Laplace	3:1	(10,15)	0.0428	0.0568	0.0394	0.0338
		(10,20)	0.0446	0.0562	0.0346	0.0310
		(30,40)	0.0490	0.0512	0.0310	0.0374
		(45,50)	0.0522	0.0566	0.0326	0.0376
		(50,100)	0.0486	0.0506	0.0308	0.0382
	5:1	(10,15)	0.0448	0.0566	0.0282	0.0364
		(10,20)	0.0478	0.0566	0.0362	0.0406
		(30,40)	0.0492	0.0508	0.0364	0.0394
		(45,50)	0.0506	0.0522	0.0366	0.0332
		(50,100)	0.0510	0.0508	0.0454	0.0280
Lognormal	3:1	(10,15)	0.0496	0.0520	0.0350	0.0344
		(10,20)	0.0580	0.0550	0.0408	0.0365
		(30,40)	0.0450	0.0520	0.0380	0.0334
		(45,50)	0.0516	0.0508	0.0420	0.0362
		(50,100)	0.0568	0.0506	0.0364	0.0350
	5:1	(10,15)	0.0534	0.0550	0.0368	0.0354
		(10,20)	0.0538	0.0498	0.0354	0.0358
		(30,40)	0.0450	0.0472	0.0426	0.0334
		(45,50)	0.0588	0.0546	0.0366	0.0338
		(50,100)	0.0582	0.0492	0.0508	0.0398

Table 2. Probability of type I error of four test statistics with unequal sample sizes

4.2. **Power of The Tests.** To study the power of the tests, the difference between the means of two populations is determined in two cases: (1) a difference with the ratio of 1.5:1 for the small mean difference and (2) a difference with the ratio of 2:1 for the moderate mean difference. To compare the efficiency of the four test statistics, the power of the test statistics can control for the probability of type I errors are considered. The details are as follow.

(1) The sample sizes are equal

(a) The variance ratio is 3:1:

For the Laplace distribution with the small mean difference, the Bootstrap Yuen Welch test (BYW) has a higher estimation power than the Yuen Welch test, except for sample sizes (50,50) and (100,100). The Yuen Welch test (YW), on the other hand, has the

higher test estimation power, when the mean difference is moderate, with the only exception being when the sample size is (10,10). In that case the Bootstrap Yuen Welch test (BYW) is marginally superior in test estimation power. (As shown in Figure 1)

For the Lognormal distribution with the small mean difference, the Brunner-Munzel test (BM) has the highest estimation power with the exception of a sample sizes equal to (10,10), (20,20) and (30,30). For those sample sizes, the Bootstrap Yuen Welch test (BYW) has the highest estimation power. When the mean difference is moderate, the Bruner-Munzel test (BM) has the highest test estimation power for all sample sizes. (As shown in Figure 1)

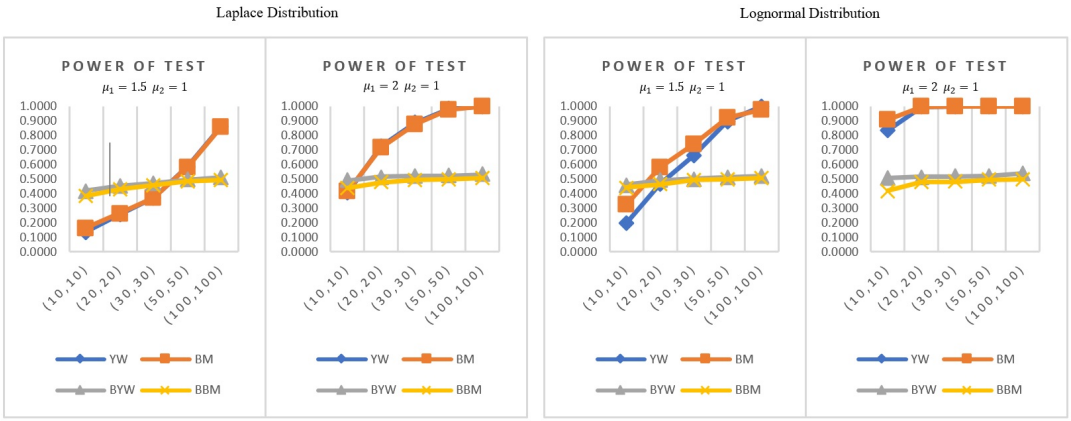

Figure 1. Test power estimation when given a variance ratio of 3:1 for equal sample sizes.

(b) The variance ratio is 5:1 with equal sample sizes:

For the Laplace distribution with a small mean difference, the Bootstrap Yuen Welch test (BYW) has the highest test estimation power, except for the case of a sample size (100,100) the Bruner-Munzel test is superior. The Yuen Welch test (YW) has the highest estimation power when the mean difference is moderate, except for a sample size equal to (10,10), the Bootstrap Yuen Welch test (BYW) has a higher test estimation power. (As shown in Figure 2)

For the Lognormal distribution with a small mean difference, the Brunner-Munzel test (BM) has the highest test estimation power except for sample sizes of (10,10) and (20,20) in which the Bootstrap Yuen Welch test (BYW) has superior test estimation power. When the mean difference is moderate, the Bruner-Munzel test (BM) has the highest test estimation power for all sample sizes (As shown in Figure 2).

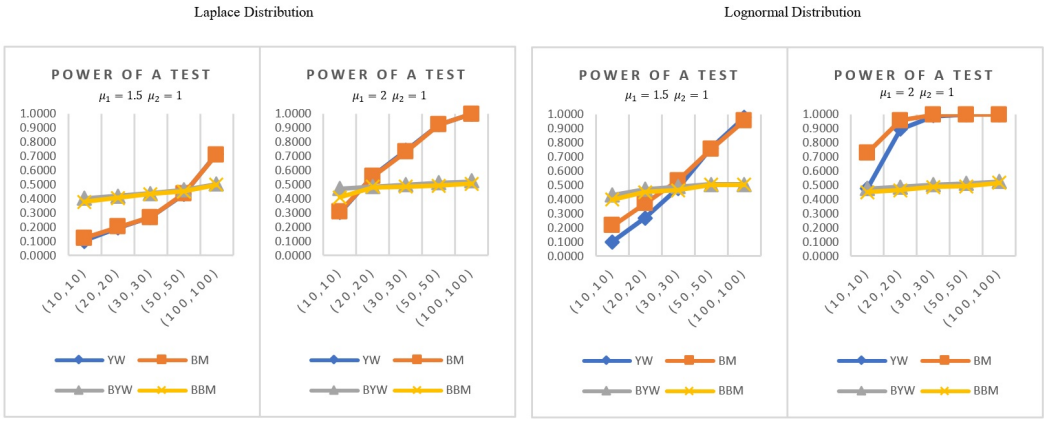

Figure 2. Test power estimation when given a variance ratio of 5:1 for the equal sample sizes.

(2) The sample sizes are not equal

(a) The variance ratio is 3:1 with unequal sample sizes:

For a Laplace distribution with the small mean difference, the Bootstrap Yuen Welch test (BYW) has the highest test estimation power, except for the cases of sample sizes of (45,50) and (50,100) in which the Bruner-Munzel statistic is superior. When the mean difference is moderate, the Yuen Welch test (YW) has the highest test estimation power. Except for sample sizes of (10,15) and (10,20), the Bootstrap Yuen Welch test (BYW) has the highest estimation power. (As shown in Figure 3)

For Lognormal distributions with small mean differences, the Brunner-Munzel test (BM) has the highest test estimation power except for sample sizes of (10,15) and (10,20) in which the Bootstrap Yuen Welch test (BYW) has a higher test estimation power. When the mean difference is moderate, the Bruner-Munzel test (BM) has the highest test estimation power for all sample sizes. (As shown in Figure 3)

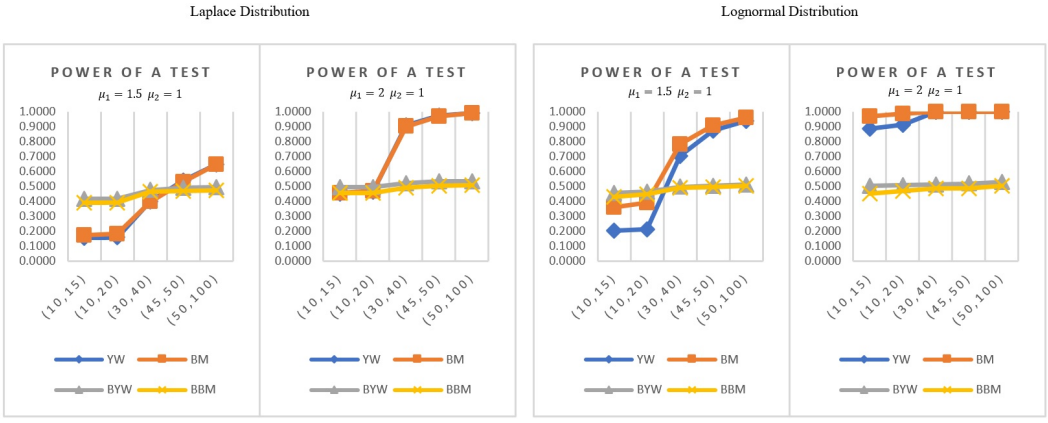

Figure 3. Test power estimation with a variance ratio of 3:1 for unequal sample sizes.

(b) The variance ratio is 5:1 with unequal sample sizes:

For the Laplace distribution with a small mean difference, the Yuen Welch test (BYW) has the highest test estimation power except for sample sizes of (45,50) and (50,100). When the mean difference is moderate, the Yuen Welch test (YW) has the highest estimation power except for sample sizes of (10,15) and (10,20), where the Bootstrap Yuen Welch test (BYW) has the greatest estimation power. (As shown in Figure 4)

For the Lognormal distribution with small mean differences, the Brunner-Munzel test (BM) has the highest test estimation power, except for sample sizes of (10,15) and (10,20) in which the Bootstrap Yuen Welch test (BYW) has the higher test estimation power. When the mean differences are moderate, the Bruner-Munzel test (BM) has the highest estimation power regardless of sample size. (As shown in Figure 4)

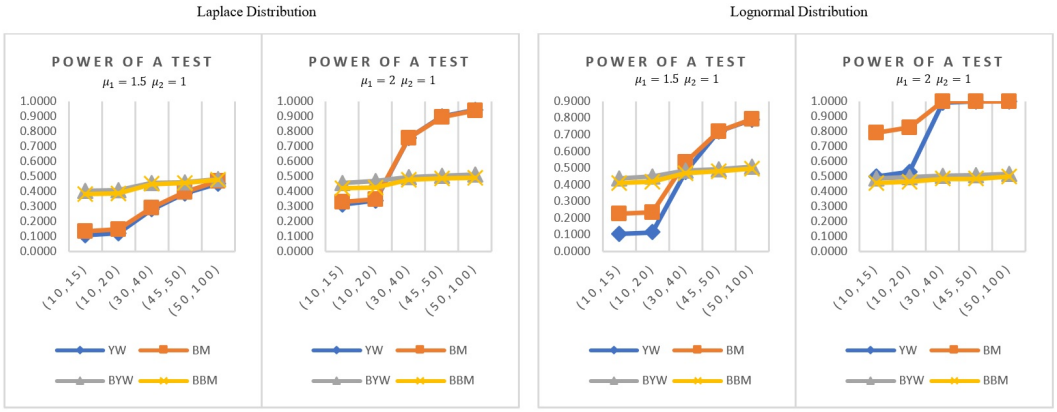

Figure 4. Test power estimation with a variance ratio of 5:1 for unequal sample sizes.

5. Conclusion and Discussion

In summary, it was found that under both the Laplace and Lognormal distributions with sample sizes being both equal and unequal, the probability of a type I error for the four test statistics could be controlled for in all cases according to Bradley's criteria.

The test statistics with the highest testing power in most situations for the Laplace distribution was the Bootstrap Yuen Welch test (BYW). However, as mean differences increased, the Yuen Welch test (YW) had the greatest strength. The YW test was the best with a sample size of more than 30 in which the sample were both equal and unequal sizes.

The test statistics with the highest testing power in most situations for the Lognormal distribution was the Brunner-Munzel test (BM). However, as mean differences increased, the Bootstrap Yuen

Welch (BYW) test statistic had the highest testing power.

In this study, the Bootstrap Yuen Welch test (BYW) method had the highest testing power in the skewness to the right when the mean difference was small and the sample size was less than 20. When the sample size was larger and the data less skewed, the Brunner-Munzel test (BM) had a greater testing power than the Bootstrap Yuen Welch test. This is consistent with research by Fagerland and Sandvik (2009) which found that the Yuen Welch test (YW) was appropriate for data with skewness to the right. Moreover, it was found the power of the test was higher as the sample size increased.

Acknowledge: The authors would like to thank the International SciKU Branding (ISB), Faculty of Science, Kasetsart University, Thailand for supporting this research.

Conflicts of Interest: The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] J.V. Bradley, Robustness?, Br. J. Math. Stat. Psychol. 31 (1978), 144-152. https://doi.org/10.1111/j.2044-8317.1978.tb00581.x.

[2] E. Brunner, U. Munzel, The Nonparametric Behrens-Fisher Problem: Asymptotic Theory and a Small-Sample Approximation, Biom. J. 42 (2000), 17-25. https://doi.org/10.1002/(sici)1521-4036(200001)42:1<17::aid-bimj17>3.0.co;2-u.

[3] A.K. Dwivedi, I. Mallawaarachchi, L.A. Alvarado, Analysis of Small Sample Size Studies Using Nonparametric Bootstrap Test With Pooled Resampling Method, Stat. Med. 36 (2017), 2187-2205. https://doi.org/10.1002/sim.7263.

[4] B. Efron, R.J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall, New York, (1993).

[5] M.W. Fagerland, L. Sandvik, Performance of Five Two-Sample Location Tests for Skewed Distributions With Unequal Variances, Contemp. Clinic. Trials. 30 (2009), 490-496. https://doi.org/10.1016/j.cct.2009.06.007.

[6] W. Noppadon, B. Chinnapong, A Comparison Study of Nonparametric Test Statistics for Two Populations in Case of Heterogeneity of Variances, J. Sci. Technol. 18 (2010), 60-69.

[7] S. Ratiwat, Comparison of Nonparametric Statistical Test for Two Independent Difference Medians in Case of Symmetrical and Skewed Distribution, Master Thesis, Bangkok: King Mongkut's Institute of Technology Ladkrabang Thailand, (2019).

[8] J. Reiczigel, I. Zakarias, L. Rozsa, A Bootstrap Test of Stochastic Equality of Two Populations, Amer. Stat. 59 (2005), 156-161. https://doi.org/10.1198/000313005x23526.

[9] R.R. Wilcox, Introduction to robust estimation and hypothesis testing (2nd ed.), Academic Press, San Diego, (2005).

[10] K.K. Yuen, The Two-Sample Trimmed T for Unequal Population Variances, Biometrika, 61 (1974), 165-170.

[11] D.W. Zimmerman, B.D. Zumbo, Rank Transformations and the Power of the Student t Test and Welch t' Test for Non-Normal Populations With Unequal Variances, Canadian J. Exper. Psychol. 47 (1993), 523-539. https://doi.org/10.1037/h0078850.