Robustness of Alternative and Classical Statistics in Two-sample Location Tests for Small Sample Sizes

Praphat Klubnual, Montri Sangthong and Wannaporn Suthon*

Division of Mathematics, Faculty of Science and Technology, Rajamangala University of Technology Suvarnabhumi, Ayutthaya, 13000 Thailand

* Corresponding author. E-mail address: wannaporn.ka@rmutsb.ac.th

Received: 4 April 2022; Revised: 27 September 2023; Accepted: 9 October 2023; Available online: 16 November 2023

Abstract

When the sample size is small, there is a possibility that the two population groups do not follow assumptions. This includes population distribution and variance. Thus, proper statistical techniques must be selected for generalisation. This article classifies statistical techniques into two types: First, classical statistics consisting of independent t-test, Welch t-test, and exact Wilcoxon-Mann-Whitney test (WMW) and, second, alternative statistics consisting of nonparametric bootstrap t-test (NBTT), nonparametric bootstrap Welch t-test (NBWT), nonparametric bootstrap Welch test based on rank (NBWR), and an exact permutation t-test (PTT). The objective of this study was to propose an alternative statistical method for a small sample size study. The data simulation tested both normal and non-normal distributions including equal and unequal variances. The results revealed that when the populations had normal or non-normal distribution and equal variances, almost all test statistics had robustness at a significance level of 0.05. For a significance level of 0.01, if at least one group had normal distribution, the Welch t-test was the most robust. If there were other distributions, the independent t-test was most robust. For unequal variance, when at least one group had a normal distribution with higher variance than other groups, the Welch t-test could control type I errors in all conditions at significance levels of 0.05 and 0.01. In other cases, it was non-robust. Therefore, if a small sample size is applied, the results must be carefully generalized.

Keywords: two-sample location tests, small sample sizes, bootstrap test, permutation test, parametric test, nonparametric test, robustness

Introduction

Quantitative research, in many fields, normally applies a statistical significance test based on empirical data for hypothesis testing. However, observed data frequently do not follow the preliminary assumptions (Keselman et al., 1998; Snyder & Thompson, 1998) For instance, the population must have a normal distribution or the variance of each population must be equal. It could be said that the widely applied statistical technique for comparing means between two populations is an independent t-test (Nguyen et al., 2016). The independent t-test is a parametric test on the assumptions of a population's normality and equal variance. This test was modified by Welch (1937), specifically known as the Welch t-test, and designed for unequal variance but the normality assumption remained (Welch, 1937). Moreover, many scholars have proposed a method when the distribution deviates from normality. The widely used non-parametric technique is the Wilcoxon-Mann Whitney test (WMW) (Fagerland & Sandvik, 2009). In some circumstances, these studies are based on small sample sizes. Ruthsatz and Urbach (2012) stated that there are many reasons why we have a small sample instead of a large sample such as limited budget, time, or ethical constraints. It may also be impossible for scientific research that works on rare animal species to have a large distribution. This situation occurs in biomedical research including experimental designs in laboratories and pilot randomized controls in clinical studies (Dwivedi,

Mallawaarachchi, & Alvarado, 2017) and nonparametric approaches are commonly recommended for analyzing data from small samples (Altman, Gore, & Gardner, 1983).

The p-values in a nonparametric test can be obtained in two ways: First, by calculating the exact probability from observed data under the null hypothesis which is proper for small sample sizes; Second, by calculating the p-value based on an asymptotic property, which is proper for large samples. Both methods are available for calculating p-values in most nonparametric tests when the assumptions of parametric tests are under suspicion (Siegal & Castellan, 1988; Mundry & Fischer, 1998). However, standard exact or asymptotic nonparametric methods do not yield appropriate results in many environments with small sample size studies (Dwivedi et al., 2017). For instance, if the first sample group contained 1000, 2000, and 3000 data and the second group contained 10000, 20000, and 30000 data, performing a WMW test to compare the location difference on a two-tailed test would not be possible and a significant difference would not be found. This conforms to the work of Dwivedi et al. (2017) who claimed that the use of exact nonparametric tests would show an insignificant p-value even when there are vast differences in small sample size studies.

Siegel (1956) authored the book titled "Nonparametric Statistics for Behavioural Sciences" which became the most highly cited and influential book in the statistical literature (Winter, 2013). According to this book, conventional parametric tests could not be applied with particularly small samples. This is because there are many assumptions underlying these tests. Specifically, the observation must be drawn from a normal distribution for both one-sample and independent two-sample t-tests but equal variances are needed for the two samples. Nonetheless, these assumptions could not be tested when the sample size is too small. Additionally, conflicting findings were reported when using parametric and nonparametric tests. Some nonparametric studies showed greater statistical power than parametric studies in small samples with non-normal distributions (Weber & Sawilowsky, 2009; Bridge & Sawilowsky, 1999; Tanizaki, 1997; Posten, 1982). On the other hand, some scholars also stated that nonparametric tests showed less or no power in small samples. Thus, the parametric method is recommended (Winter, 2013; Janusonis, 2009; Stonehouse & Forrester, 1998; Sawilowsky & Hillman, 1993; Zimermerman & Zumbo, 1992).

The other approaches for small sample sizes are resampling techniques such as nonparametric permutation and bootstrap tests. In detail, the test statistics are obtained by resampling without a replacement in the permutation method while they are obtained with a replacement in the bootstrap method. Some scholars stated that the permutation test does not perform appropriately in a small sample size. On the contrary, Efron and Tibshirani (1993) advised that the bootstrap method provides similar results to the permutation method when both data exist. They claimed that the bootstrap test is more commonly utilized although it is less accurate. Moreover, Barber and Thompson (2000) recommended using the bootstrap method for checking the robustness of parametric methods or for comparing the means in moderate or large samples with skewed data. Hall and Martin (1988) evaluated the properties of the bootstrap method and indicated that the bootstrap could be reliable when the sample size was eight or greater. The study of Dwivedi et al. (2017) revealed that the nonparametric pooled bootstrap t-test provided equal or greater power when comparing two means as compared to an unpaired t-test, Welch t-test, Wilcoxon rank sum test, and permutation test while keeping type I error probability for all conditions except for Cauchy and an extreme variable lognormal distribution.

It could be said that an alternative testing method is needed. The method would require minimal or no assumptions regarding the distribution for small samples, provide a considerable large or equal power to

parametric tests, and also be able to control type I errors. Such a method could be a nonparametric bootstrap and permutation test, which is named alternative statistics in this study. The previous method, called classical statistics, consists of the independent t-test, Welch t-test, and exact Wilcoxon-Mann Whitney test. In summary, there are uncertain results between using modern and classical statistics for comparing the location between two populations, especially for a small sample size. Therefore, the objective of this study was to propose an alternative statistical method for a small sample size study. The criteria for selecting robust statistics were based on the ability to control type 1 tolerances according to the established criteria.

Methods and Materials

Statistical Method

The testing statistics on two sample location tests for small sample sizes consist of two categories. First, the classical statistics consist of the independent t- test, Welch t- test and an exact Wilcoxon- Mann- Whitney test (WMW). Second, the modern statistics consisting of the nonparametric bootstrap t-test (NBTT), nonparametric bootstrap Welch t- test (NBWT), nonparametric bootstrap Welch test based on rank (NBWR), and an exact permutation t-test (PTT). The details of each test are as follows:

1) Independent sample t-test (t-test)

The independent sample t- test (t- test) is used for parametric statistics with assumptions consisting of normal distribution and equal variance populations. The details are: (Fagerland & Sandvik, 2009)

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_P^2}{n_1} + \frac{S_P^2}{n_2}}} \quad (1)$$

Where t is T-distribution

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

The reject H_0 is $t < -t_{\alpha/2,n_1+n_2-2}$ or $t > t_{\alpha/2,n_1+n_2-2}$ or the P-value = $P\left(T < \left|t_{df=n_1+n_2-2}\right|\right) < \alpha$

2) Welch t-test

The Welch t-test is for parametric statistics with the assumption of normal distribution. The Welch t-test was proposed by Welch (1937) for unequal variances. The details are:

$$\text{Welch} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \quad (3)$$

Where Welch is T-distribution

$$S_1^2 = \frac{\sum_{i=1}^{n_1}\left(X_{i1} - \bar{X}_1\right)^2}{n_1 - 1} \quad (4)$$

$$S_2^2 = \frac{\sum_{i=1}^{n_2}\left(X_{i2} - \bar{X}_2\right)^2}{n_2 - 1} \quad (5)$$

The reject H_0 is $Welch < -t_{\alpha/2,df}$ or $Welch > t_{\alpha/2,df}$ by $df = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$ or

$$P\left(T < \left|Welch\right|\right) < \alpha$$

3) Exact Wilcoxon-Mann-Whitney test (WMW)

The Wilcoxon-Mann-Whitney test (WMW) is mostly used for nonparametric statistics. The Wilcoxon-Mann-Whitney test was proposed by Wilcoxon (1945) and Mann and Whitney (1947) as follows:

$$WMW = n_1 n_2 + \frac{n_1\left(n_1 + 1\right)}{2} - R_X \quad (6)$$

where the WMW is the statistic test for the Wilcoxon-Mann-Whitney test.

R_X is the sum of rank for the n_1 by pooled rank between n_1 and n_2

The reject H_0 is $2*(\text{Probability of WMW}) < \alpha$

4) Nonparametric bootstrap t-test (NBTT)

The nonparametric bootstrap t-test (NBTT) is developed by the bootstrap concept, with replacement sampling, integrated with an independent sample t-test. The procedure is computed as follows: (Efron & Tibshirani, 1993)

(1) Evaluate test statistics: $t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_P^2}{n_1} + \frac{S_P^2}{n_2}}}$ (7)

(2) The observation of sample 1 and sample 2 is combined so that the sample size is $n_1 + n_2$

(3) Draw two bootstrap samples of sizes n_1^* and n_2^* from the sample size is $n_1 + n_2$

(4) Evaluate test statistics: $t^* = \frac{\bar{X}_1^* - \bar{X}_2^*}{\sqrt{\frac{S_P^{2*}}{n_1^*} + \frac{S_P^{2*}}{n_2^*}}}$ (8)

(5) Repeat steps 3 and 4 for B times (in this paper, B is 1,000 times)

(6) Calculate: P-value = $\frac{\text{number of times}\left(\left|t^*\right| \geq \left|t\right|\right)}{B}$ (9)

The reject H_0 is $P - value < \alpha$

5) Nonparametric bootstrap Welch t-test (NBWT)

The nonparametric bootstrap Welch t-test (NBWT) is integrated with the bootstrap and Welch t-tests. It is described as follows: (Efron & Tibshirani, 1993)

(1) Evaluate test statistics: $$\text{Welch} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \quad (10)$$

(2) The observation of sample 1 and sample 2 is combined so that the sample size is $n_1 + n_2$

(3) Draw two bootstrap samples are sample sizes n_1^* and n_2^* from the sample size is $n_1 + n_2$

(4) Evaluate test statistics: $$\text{Welch}^* = \frac{\bar{X}_1^* - \bar{X}_2^*}{\sqrt{\frac{S_1^{2*}}{n_1^*} + \frac{S_2^{2*}}{n_2^*}}} \quad (11)$$

(5) Repeat steps 3 and 4 for B times (this paper, B is 1,000 times)

(6) Calculate: $$\text{P-value} = \frac{\text{number of times}\left(\left|\text{Welch}^*\right| \geq \left|\text{Welch}\right|\right)}{B} \quad (12)$$

The reject H_0 is $P - \text{value} < \alpha$

6) Nonparametric bootstrap welch t-test based on rank (NBWR)

The nonparametric bootstrap Welch t- test based on rank (NBWR) is integrated with the bootstrap test and the Welch t- test based on rank. The procedure is computed as follows: (Efron & Tibshirani, 1993; Reiczigel, Zakarias, & Rozsa, 2005)

(1) The observation of sample 1 and sample 2 is combined so that sample size is $n_1 + n_2$, combined data rank in ascending order. An average of ranks is given for tied values.

(2) Evaluate test statistics: $$\text{NBWR} = \frac{\bar{R}_1 - \bar{R}_2}{\sqrt{\frac{S_{\bar{R}_1}^2}{n_1} + \frac{S_{\bar{R}_2}^2}{n_2}}} \quad (13)$$

(3) Draw two bootstrap samples are sample sizes n_1^* and n_2^* from the sample size $n_1 + n_2$, the combined data rank is in ascending order. An average of ranks is given for tied values.

(4) Evaluate test statistics: $$\text{NBWR}^* = \frac{\bar{R}_1^* - \bar{R}_2^*}{\sqrt{\frac{S_{\bar{R}_1}^{2*}}{n_1^*} + \frac{S_{\bar{R}_2}^{2*}}{n_2^*}}} \quad (14)$$

(5) Repeat steps 3 and 4 for B times (in this paper, B is 1,000 times)

(6) Calculate: $$\text{P-value} = \frac{\text{number of times}\left(\left|\text{NBWR}^*\right| \geq \left|\text{NBWR}\right|\right)}{B} \quad (15)$$

The reject H_0 is $P - \text{value} < \alpha$

(7) Exact Permutation t-test (PTT)

The exact Permutation t- test (PTT) is integrated without replacement sampling and the Welch t-test. The procedure is computed as follows: (Efron & Tibshirani, 1993)

(1) Evaluate test statistics: $$\text{Welch} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \quad (16)$$

(2) The observation of sample 1 and sample 2 is combined so that the sample size is $n_1 + n_2$.

(3) Draw two samples, without replacement sampling, with sample sizes n_1^* and n_2^* from the sample size $n_1 + n_2$

(4) Evaluate test statistics: $$\text{Welch}^* = \frac{\bar{X}_1^* - \bar{X}_2^*}{\sqrt{\frac{S_1^{2*}}{n_1^*} + \frac{S_2^{2*}}{n_2^*}}} \quad (17)$$

(5) Repeat steps 3 and 4 for $\binom{n_1 + n_2}{n_1}$ times

(6) Calculate: $$\text{P-value} = \frac{\text{number of times}\left(|\text{Welch}^*| \geq |\text{Welch}|\right)}{\binom{n_1 + n_2}{n_1}} \quad (18)$$

The reject H_0 is $P - \text{value} < \alpha$

Methodology

Program R version 3.2.1 was designed to generate random data sets obtained from two populations. Those populations were different in several conditions but had equal and unequal means. The data set was tested with classical statistics including an independent t-test, Welch t-test, exact Wilcoxon-Mann-Whitney test (WMW) and with modern statistics including a nonparametric bootstrap t-test (NBTT), nonparametric bootstrap Welch t-test (NBWT), nonparametric bootstrap welch t-test based on rank(NBWR) and exact permutation t-test (PTT). The data sets were varied in three different conditions, which were population distribution, sample sizes, and variances. The details were:

1) Population distributions were a normal distribution using the function $\text{Norm}(\mu, \sigma^2)$, lognormal distribution using the function $\log\text{Norm}(\mu, \sigma^2)$ and gamma distribution using the function $\text{Gamma}(\alpha, \beta)$ both identically distributed and non-identically distributed.

2) Population variances were equal and unequal variances. For the unequal variance, the ratio was 1:9 given the F-ratio between the two data sets was also 9, which showed to be statistically significant for all used sample sizes.

3) Small sample sizes. For each data set, sample sizes were (3,3) (3,5) (5,3) (5,5) (7,7) (7,10) (10,7), and (10,10).

4) The robustness of testing statistics considered from maintained type I errors. All combinations of conditions were tested and each case was repeated 10,000 times. After that, results from the seven testing statistics were compared by considering type I errors.

- The probability of type I error (alpha) was calculated by focusing on the frequency of rejection H_0 by fixing the same mean ratio for both groups (the fraction of the P-value formula) divided 10,000 times. Bradley's (1978) standard criterion was applied to evaluate the differences between the actual type I error rate and the nominal significance level. This criterion has been applied to the robustness of investigations, such as Haidous

and Sawilowsky (2013) and Nguyen et al. (2016). The actual rate of probability of type I errors between 0.05 and 0.01 was considered acceptable. In detail, if the probability of type I error was between 0.025-0.075 or 0.005-0.015, it could be said that the test statistics could control the type I error at a significant level of 0.05 or 0.01 or the statistical test is robust.

Results

The results of the simulation are presented in the type I error at significance levels of 5% and 1% with the details as follows:

Table 1 shows the type I error probability (significance level of 5%) for the independent t-test, Welch t-test, exact Wilcoxon-Mann-Whitney test, nonparametric bootstrap t-test, nonparametric bootstrap Welch t-test, nonparametric bootstrap welch t-test based on rank, and permutation t-test. Almost all tests could control type I errors when variances were equal for both populations. However, when the sample size was smallest (3,3), it was found that the Wilcoxon-Mann-Whitney test and permutation t-test could not control type I errors.

Table 2 shows the type I error probability (significance level of 1%) that if at least one group of the population has a normal distribution, the Welch t-test could control the type I error on almost all conditions. When considered in detail, it was found that both populations had a normal distribution. Both an independent sample t-test and a Welch t-test therefore could control the type I error on all conditions. However, if the first group had normal distribution and the second had lognormal, the Welch t-test was more prominent than other test statistics. In the case where the first group had a normal distribution and the second had a gamma distribution, the Welch t-test could control the type I error under almost all conditions. On the other hand, if both groups had non-normal distribution, their distributions could be either lognormal distribution, gamma distribution, or lognormal distribution and gamma distribution. It was found that the independent sample t-test was the best for controlling the type I error followed by a nonparametric bootstrap t-test. However, when the sample size was smallest (3,3), it was found that the Wilcoxon-Mann-Whitney test nonparametric bootstrap t-test, nonparametric bootstrap Welch t-test, nonparametric bootstrap Welch t-test based on rank, and permutation t-test could not control type I error.

Table 1 The type I error probability (nominal level=0.05) for location difference testing statistics between two populations under equal variances and various distributions and sample sizes

Sample size	Normal Distribution vs. Normal Distribution							Normal Distribution vs. Lognormal Distribution						
	t	Welch	WMW	NBTT	NBWT	NBWR	PTT	t	Welch	WMW	NBTT	NBWT	NBWR	PTT
3,3	**0.0503**	**0.0364**	0.0000	**0.0603**	**0.0596**	0.0076	0.0000	**0.0692**	**0.0434**	0.0000	0.0812	0.0806	**0.0084**	0.0000
3,5	**0.0489**	**0.0462**	**0.0368**	**0.0510**	**0.0506**	**0.0404**	**0.0362**	0.0999	**0.0545**	0.0761	0.1089	0.0774	0.0817	0.0902
5,3	**0.0513**	**0.0472**	**0.0358**	**0.0529**	**0.0515**	**0.0403**	**0.0367**	**0.0368**	**0.0502**	**0.0377**	**0.0371**	**0.0531**	**0.0416**	**0.0261**
5,5	**0.0507**	**0.0433**	**0.0314**	**0.0522**	**0.0529**	**0.0459**	**0.0462**	**0.0638**	**0.0498**	**0.0490**	**0.0679**	**0.0692**	**0.0634**	**0.0662**
7,10	**0.0479**	**0.0468**	**0.0416**	**0.0488**	**0.0487**	**0.0490**	**0.0473**	0.0780	**0.0490**	**0.0724**	0.0818	**0.0615**	**0.0708**	0.0853
10,7	**0.0500**	**0.0486**	**0.0409**	**0.0504**	**0.0500**	**0.0495**	**0.0436**	**0.0426**	**0.0562**	**0.0489**	**0.0454**	**0.0586**	**0.0672**	**0.0534**
10,10	**0.0529**	**0.0509**	**0.0431**	**0.0534**	**0.0528**	**0.0505**	**0.0537**	**0.0628**	**0.0576**	**0.0655**	**0.0662**	**0.0658**	**0.0738**	0.0818
	Normal Distribution vs. Gamma Distribution							Lognormal Distribution vs. Lognormal Distribution						
3,3	**0.0591**	**0.0409**	0.0000	**0.0710**	**0.0706**	0.0074	0.0000	**0.0314**	0.0188	0.0000	**0.0411**	**0.0399**	0.0058	0.0000
3,5	**0.0685**	**0.0457**	**0.0487**	**0.0730**	**0.0580**	**0.0538**	**0.0567**	**0.0325**	0.0215	**0.0328**	**0.0389**	**0.0300**	**0.0369**	**0.0330**

Table 1 (Cont.)

Sample size	Normal Distribution vs. Normal Distribution							Normal Distribution vs. Lognormal Distribution						
	t	Welch	WMW	NBTT	NBWT	NBWR	PTT	t	Welch	WMW	NBTT	NBWT	NBWR	PTT
	Normal Distribution vs. Gamma Distribution							Lognormal Distribution vs. Lognormal Distribution						
5,3	**0.0458**	**0.0528**	**0.0391**	**0.0468**	**0.0534**	**0.0437**	**0.0349**	**0.0353**	0.0237	**0.0353**	**0.0435**	**0.0333**	**0.0399**	**0.0383**
5,5	**0.0585**	**0.0488**	**0.0411**	**0.0612**	**0.0608**	**0.0588**	**0.0583**	**0.0304**	0.0186	**0.0300**	**0.0396**	**0.0397**	**0.0490**	**0.0470**
7,10	**0.0607**	**0.0488**	**0.0581**	**0.0638**	**0.0532**	**0.0644**	**0.0627**	**0.0315**	**0.0307**	**0.0402**	**0.0420**	**0.0413**	**0.0495**	**0.0506**
10,7	**0.0513**	**0.0590**	**0.0504**	**0.0535**	**0.0619**	**0.0619**	**0.0545**	**0.0309**	**0.0282**	**0.0394**	**0.0420**	**0.0404**	**0.0468**	**0.0405**
10,10	**0.0523**	**0.0482**	**0.0564**	**0.0554**	**0.0539**	**0.0647**	**0.0563**	**0.0338**	**0.0278**	**0.0434**	**0.0419**	**0.0423**	**0.0489**	**0.0501**
	Lognormal Distribution vs. Gamma Distribution							Gamma Distribution vs. Gamma Distribution						
3,3	**0.0495**	**0.0322**	0.0000	**0.0614**	**0.0606**	0.0079	0.0000	**0.0418**	**0.0256**	0.0000	**0.0505**	**0.0512**	0.0070	0.0000
3,5	**0.0342**	**0.0255**	**0.0281**	**0.0394**	**0.0308**	**0.0325**	**0.0351**	**0.0402**	**0.0300**	**0.0343**	**0.0455**	**0.0396**	**0.0387**	**0.0354**
5,3	**0.0609**	**0.0504**	**0.0705**	**0.0668**	**0.0684**	0.0769	**0.0557**	**0.0424**	**0.0334**	**0.0363**	**0.0485**	**0.0416**	**0.0418**	**0.0367**
5,5	**0.0406**	**0.0289**	**0.0413**	**0.0478**	**0.0476**	**0.0594**	**0.0554**	**0.0404**	**0.0292**	**0.0326**	**0.0475**	**0.0458**	**0.0549**	**0.0503**
7,10	**0.0314**	**0.0320**	**0.0418**	**0.0367**	**0.0418**	**0.0577**	**0.0341**	**0.0417**	**0.0373**	**0.0430**	**0.0472**	**0.0473**	**0.0515**	**0.0494**
10,7	**0.0533**	**0.0404**	**0.0640**	**0.0613**	**0.0532**	**0.0642**	**0.0519**	**0.0402**	**0.0377**	**0.0436**	**0.0460**	**0.0456**	**0.0513**	**0.0418**
10,10	**0.0420**	**0.0363**	**0.0600**	**0.0479**	**0.0476**	**0.0690**	**0.0400**	**0.0442**	**0.0383**	**0.0458**	**0.0494**	**0.0500**	**0.0519**	**0.0509**

*Bold entries indicate a type I error that is controlled.

Table 2 The type I error probability (nominal level=0.01) for location difference testing statistics between two populations under equal variances and various distributions and sample sizes

Sample size	Normal Distribution vs. Normal Distribution							Normal Distribution vs. Lognormal Distribution						
	t	Welch	WMW	NBTT	NBWT	NBWR	PTT	t	Welch	WMW	NBTT	NBWT	NBWR	PTT
3,3	**0.0109**	**0.0064**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0196	**0.0103**	0.0000	0.0001	0.0000	0.0000	0.0000
3,5	**0.0118**	**0.0090**	0.0000	**0.0103**	**0.0077**	0.0000	0.0000	0.0360	**0.0147**	0.0000	0.0352	0.0195	0.0001	0.0000
5,3	**0.0107**	**0.0090**	0.0000	**0.0096**	**0.0074**	0.0000	0.0000	**0.0079**	**0.0113**	0.0000	**0.0070**	**0.0094**	0.0003	0.0000
5,5	**0.0094**	**0.0073**	**0.0076**	**0.0099**	**0.0095**	**0.0095**	**0.0076**	0.0177	**0.0110**	0.0164	0.0186	0.0179	0.0200	0.0164
7,10	**0.0088**	**0.0097**	**0.0094**	**0.0107**	**0.0099**	**0.0106**	**0.0090**	0.0240	**0.0101**	0.0214	0.0282	0.0167	0.0229	0.0291
10,7	**0.0098**	**0.0102**	**0.0087**	**0.0103**	**0.0104**	**0.0102**	**0.0083**	**0.0092**	**0.0133**	**0.0133**	**0.0103**	0.0176	0.0178	0.0155
10,10	**0.0100**	**0.0090**	**0.0099**	**0.0108**	**0.0116**	**0.0112**	**0.0096**	0.0171	**0.0134**	0.0190	0.0190	0.0188	0.0216	0.0291
	Normal Distribution vs. Gamma Distribution							Lognormal Distribution vs. Lognormal Distribution						
3,3	**0.0150**	**0.0090**	0.0000	0.0000	0.0000	0.0000	0.0000	**0.0065**	0.0029	0.0000	0.0000	0.0000	0.0000	0.0000
3,5	0.0215	**0.0111**	0.0000	0.0182	**0.0106**	0.0002	0.0000	**0.0066**	0.0036	0.0000	**0.0074**	0.0036	0.0000	0.0000
5,3	**0.0087**	**0.0124**	0.0000	**0.0076**	**0.0092**	0.0000	0.0000	**0.0068**	0.0041	0.0000	**0.0060**	0.0041	0.0003	0.0000
5,5	0.0165	**0.0113**	**0.0117**	0.0163	0.0157	**0.0144**	**0.0117**	**0.0038**	0.0019	**0.0067**	0.0034	0.0037	**0.0089**	**0.0067**
7,10	0.0158	**0.0105**	**0.0150**	0.0172	**0.0136**	0.0172	0.0162	**0.0040**	0.0020	**0.0086**	**0.0064**	0.0047	**0.0108**	**0.0089**
10,7	**0.0122**	0.0167	**0.0144**	**0.0133**	0.0176	0.0174	**0.0134**	**0.0033**	0.0020	**0.0083**	**0.0051**	0.0044	**0.0097**	**0.0060**
10,10	**0.0132**	**0.0110**	**0.0130**	**0.0142**	**0.0143**	0.0153	0.0166	**0.0047**	0.0026	**0.0083**	**0.0075**	**0.0080**	**0.0105**	**0.0098**
	Lognormal Distribution vs. Gamma Distribution							Gamma Distribution vs. Gamma Distribution						
3,3	**0.0114**	**0.0059**	0.0000	0.0000	0.0000	0.0000	0.0000	**0.0107**	**0.0062**	0.0000	0.0000	0.0000	0.0000	0.0000
3,5	**0.0067**	**0.0051**	0.0000	**0.0064**	0.0048	0.0000	0.0000	**0.0078**	0.0046	0.0000	**0.0077**	0.0044	0.0002	0.0000
5,3	0.0156	**0.0094**	0.0000	**0.0140**	**0.0099**	0.0002	0.0000	**0.0103**	**0.0051**	0.0000	**0.0095**	**0.0053**	0.0000	0.0000
5,5	**0.0077**	0.0048	**0.0116**	**0.0085**	**0.0080**	**0.0142**	**0.0116**	**0.0070**	0.0039	**0.0080**	**0.0070**	**0.0072**	**0.0103**	**0.0080**
7,10	**0.0053**	0.0035	**0.0088**	**0.0071**	**0.0063**	**0.0114**	**0.0054**	**0.0060**	0.0035	**0.0096**	**0.0090**	**0.0070**	**0.0096**	**0.0106**
10,7	**0.0104**	**0.0079**	0.0161	**0.0147**	**0.0117**	0.0169	**0.0106**	**0.0070**	**0.0059**	**0.0101**	**0.0085**	**0.0099**	**0.0112**	**0.0079**
10,10	**0.0083**	**0.0054**	**0.0141**	**0.0114**	**0.0109**	0.0168	**0.0080**	**0.0072**	0.0049	**0.0104**	**0.0100**	**0.0107**	**0.0118**	**0.0112**

*Bold entries indicate a type I error that is controlled.

Tables 3 and 4 also show the interaction between heteroscedastic and non-normal distributions, which affects type I errors. For instance, when both populations are lognormal distributions, none of the test statistics could control type I errors at significance levels of 5% and 1%. Moreover, the Wilcoxon-Mann-Whitney test and nonparametric bootstrap Welch t-test Based on Rank have a high type I error rate (24.41%-25.76%) and 10.12%-11.33% at significance levels of 5% and 1% respectively when the sample size is highest in a small group (10,10). Regarding the least population to have a normal distribution and a higher variance than other groups, the Welch t-test could control type I errors in all conditions at significance levels of 5% and 1%. When considered specifically at a significance level of 0.05, both groups had a normal distribution. It was found that the Welch t-test could control the type I error on all conditions followed by the independent t-test and nonparametric bootstrap Welch t-test respectively. In the case where the first group had lognormal and the second had a normal distribution with higher variance, it was found that the Welch t-test could control the type I error on all conditions followed by independent t-test, the Wilcoxon-Mann-Whitney test and nonparametric bootstrap Welch t-test respectively, and in the case where the first group had gamma distribution and the second had normal distribution with higher variance, it was found that Welch t-test could control the type I error on all conditions followed by independent t-test, exact Wilcoxon-Mann-Whitney test, nonparametric bootstrap t-test, nonparametric bootstrap Welch t-test, and permutation t-test. When considered specifically at significance level 0.01, the first condition was both groups had a normal distribution. The second condition was that the first group had a lognormal distribution and the second group had a normal distribution with higher variance. The last condition was the first group had a gamma distribution and the second group had a normal distribution with higher variance. The results showed that the Welch t-test could control the type I error on almost all conditions. By contrast, if the populations had other distributions, most-test statistics could not control the type I error both at the significance levels of 0.05 and 0.01. Finally, when the distributions and the variance ratio were reversed, the outcome was different. For instance, when the first population had a normal distribution and the second had a lognormal distribution with higher variance, Welch t-test statistics could not control type I errors. However, when the condition was reversed the Welch t-test could control type I errors on all conditions at 5% and 1% significance levels. Nevertheless, if the sample sizes were unequal, the smaller group that had the higher variance was more capable of controlling the type I error.

Table 3 The type I error probability (nominal level=0.05) for location difference testing statistics between two populations under unequal variances and various distributions and sample sizes

Sample size	Normal Distribution vs. Normal Distribution							Normal Distribution vs. Lognormal Distribution						
	t	Welch	WMW	NBTT	NBWT	NBWR	PTT	t	Welch	WMW	NBTT	NBWT	NBWR	PTT
3,3	**0.0744**	**0.0543**	0.0000	0.0865	0.0848	0.0104	0.0000	0.1158	0.0846	0.0000	0.1314	0.1297	0.0109	0.0000
3,5	**0.0272**	**0.0430**	**0.0272**	**0.0291**	**0.0429**	**0.0309**	0.0180	0.1034	0.0790	**0.0667**	0.1044	0.0874	**0.0732**	0.0782
5,3	0.1424	**0.0680**	0.1002	0.1581	0.1036	0.1080	0.1333	0.1394	0.1395	0.1082	0.1416	0.1423	0.1164	0.1040
5,5	**0.0689**	**0.0538**	**0.0471**	0.0757	**0.0750**	**0.0680**	**0.0717**	0.1240	0.1137	0.0893	0.1270	0.1262	0.1170	0.1197
7,10	**0.0312**	**0.0517**	**0.0434**	**0.0328**	**0.0547**	**0.0657**	**0.0256**	0.0988	0.1005	0.1387	0.0980	0.1013	0.1755	0.0858
10,7	0.1055	**0.0513**	0.0796	0.1103	**0.0710**	**0.0741**	**0.0696**	0.1457	0.1365	0.1602	0.1477	0.1398	0.1593	0.1248
10,10	**0.0589**	**0.0488**	**0.0597**	**0.0603**	**0.0604**	**0.0666**	**0.0277**	0.1244	0.1223	0.1799	0.1261	0.1252	0.1969	0.0926
	Normal Distribution vs. Gamma Distribution							Lognormal Distribution vs. Normal Distribution						
3,3	0.1174	0.0887	0.0000	0.1293	0.1283	0.0099	0.0000	0.0860	**0.0528**	0.0000	0.1059	0.1043	0.0121	0.0000
3,5	0.0726	**0.0690**	**0.0472**	**0.0709**	**0.0707**	**0.0518**	**0.0506**	**0.0289**	**0.0464**	**0.0382**	**0.0302**	**0.0508**	**0.0402**	0.0209
5,3	0.1705	0.1321	0.1247	0.1767	0.1515	0.1339	0.1398	0.1672	**0.0590**	0.1441	0.1888	0.1180	0.1504	0.1728

Table 3 (Cont.)

Sample size	Normal Distribution vs. Normal Distribution							Normal Distribution vs. Lognormal Distribution						
	t	Welch	WMW	NBTT	NBWT	NBWR	PTT	t	Welch	WMW	NBTT	NBWT	NBWR	PTT
	Normal Distribution vs. Gamma Distribution							Lognormal Distribution vs. Normal Distribution						
5,5	0.1085	0.0970	**0.0736**	0.1089	0.1105	0.0957	0.1019	**0.0698**	**0.0501**	**0.0489**	0.0774	0.0794	**0.0614**	0.0797
7,10	**0.0668**	0.0814	0.1033	**0.0684**	0.0824	0.1398	**0.0552**	**0.0316**	**0.0484**	**0.0531**	**0.0329**	**0.0553**	0.0789	0.0238
10,7	0.1446	0.1091	0.1513	0.1485	0.1182	0.1417	0.1120	0.1126	**0.0501**	0.0937	0.1201	**0.0747**	0.0859	**0.0707**
10,10	0.0997	0.0942	0.1474	0.1011	0.1012	0.1603	**0.0636**	**0.0554**	**0.0460**	**0.0701**	**0.0591**	**0.0586**	0.0780	0.0240
	Lognormal Distribution vs. Lognormal Distribution							Lognormal Distribution vs. Gamma Distribution						
3,3	0.1442	0.1069	0.0000	0.1618	0.1621	0.0143	0.0000	0.1383	0.1047	0.0000	0.1578	0.1562	0.0155	0.0000
3,5	0.1014	0.0848	0.0796	0.1061	0.0934	0.0856	0.0849	**0.0748**	0.0795	**0.0667**	0.0764	0.0815	**0.0699**	**0.0589**
5,3	0.2018	0.1581	0.2099	0.2180	0.2024	0.2153	0.1934	0.2134	0.1312	0.2040	0.2318	0.1890	0.2079	0.2127
5,5	0.1447	0.1289	0.1131	0.1525	0.1507	0.1391	0.1539	0.1180	0.1036	0.0840	0.1237	0.1227	0.1019	0.1239
7,10	0.1010	0.1046	0.1909	0.1047	0.1122	0.2434	0.0884	**0.0664**	0.0825	0.1315	**0.0680**	0.0856	0.1673	**0.0533**
10,7	0.1699	0.1431	0.2285	0.1779	0.1590	0.2142	0.1479	0.1466	0.0992	0.1632	0.1547	0.1183	0.1529	0.1104
10,10	0.1281	0.1227	0.2441	0.1330	0.1346	0.2576	0.0884	0.0939	0.0877	0.1622	0.0976	0.0970	0.1736	**0.0540**
	Gamma Distribution vs. Normal Distribution							Gamma Distribution vs. Lognormal Distribution						
3,3	0.0784	**0.0544**	0.0000	0.0942	0.0943	0.0093	0.0000	0.1090	0.0761	0.0000	0.1220	0.1241	0.0125	0.0000
3,5	**0.0299**	**0.0461**	**0.0331**	**0.0306**	**0.0486**	**0.0367**	0.0192	0.0956	**0.0612**	**0.0662**	0.1014	0.0778	**0.0714**	0.0808
5,3	0.1446	**0.0646**	0.1165	0.1625	0.1058	0.1215	0.1433	0.1525	0.1515	0.1615	0.1627	0.1737	0.1674	0.1300
5,5	**0.0676**	**0.0521**	**0.0496**	**0.0743**	0.0754	**0.0641**	**0.0742**	0.1263	0.1134	0.0997	0.1331	0.133	0.1318	0.1323
7,10	**0.0327**	**0.0516**	**0.0482**	**0.0344**	**0.0567**	**0.0746**	**0.0259**	0.0902	0.0878	0.1550	0.0930	0.0944	0.2008	0.0770
10,7	0.1045	**0.0510**	0.0874	0.1116	**0.0709**	0.0789	**0.0711**	0.1480	0.1347	0.1992	0.1561	0.1419	0.1868	0.1303
10,10	**0.0619**	**0.0531**	**0.0677**	**0.0637**	**0.0643**	0.0755	**0.0289**	0.1201	0.1168	0.2148	0.1244	0.1236	0.2292	0.0837
	Gamma Distribution vs. Gamma Distribution													
3,3	0.1225	0.0881	0.0000	0.1372	0.1388	0.0115	0.0000							
3,5	**0.0742**	**0.0690**	**0.0542**	**0.0743**	**0.0715**	**0.0590**	**0.0544**							
5,3	0.1863	0.1338	0.1717	0.2010	0.1713	0.1775	0.1759							
5,5	0.1124	0.1017	0.0810	0.1156	0.1159	0.1040	0.1126							
7,10	**0.0617**	**0.0734**	0.1138	**0.0634**	0.0775	0.1517	**0.0523**							
10,7	0.1360	0.0990	0.1575	0.1407	0.1114	0.1437	0.1046							
10,10	0.0870	0.0825	0.1476	0.0910	0.0903	0.1590	**0.0533**							

* Bold entries indicate a type I error that is controlled.

Table 4 The type I error probability (nominal level=0.01) for location difference testing statistics between two populations under unequal variances and various distributions and sample sizes

Sample size	Normal Distribution vs. Normal Distribution							Normal Distribution vs. Lognormal Distribution						
	t	Welch	WMW	NBTT	NBWT	NBWR	PTT	t	Welch	WMW	NBTT	NBWT	NBWR	PTT
3,3	0.0200	**0.0109**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0384	0.0219	0.0000	0.0001	0.0000	0.0000	0.0000
3,5	**0.0056**	**0.0068**	0.0000	0.0042	0.0047	0.0002	0.0000	0.0389	0.0235	0.0000	0.0349	0.0236	0.0002	0.0000
5,3	0.0500	0.0234	0.0000	0.0500	0.0275	0.0005	0.0000	0.0426	0.0475	0.0000	0.0370	0.0395	0.0004	0.0000
5,5	0.0167	**0.0122**	0.0164	0.0179	0.0179	0.0202	0.0164	0.0499	0.0406	0.0343	0.0481	0.0461	0.0402	0.0343
7,10	**0.0062**	**0.0110**	**0.0097**	**0.0071**	**0.0116**	**0.0136**	0.0041	0.0445	0.0395	0.0490	0.0437	0.0414	0.0601	0.0315
10,7	0.0330	**0.0116**	0.0225	0.0378	0.0220	0.0212	0.0175	0.0666	0.0674	0.0610	0.0665	0.0678	0.0613	0.0463
10,10	**0.0136**	**0.0095**	**0.0140**	0.0151	0.0160	0.0155	0.0039	0.0574	0.0546	0.0685	0.0568	0.0568	0.0759	0.0318
	Normal Distribution vs. Gamma Distribution							Lognormal Distribution vs. Normal Distribution						
3,3	0.0396	0.0237	0.0000	0.0000	0.0000	0.0000	0.0000	0.0297	**0.0139**	0.0000	0.0000	0.0002	0.0000	0.0000
3,5	0.0256	0.0195	0.0000	0.0213	0.0178	0.0000	0.0000	**0.0055**	**0.0080**	0.0000	0.0040	**0.0065**	0.0002	0.0000

Table 4 (Cont.)

Sample size	Normal Distribution vs. Normal Distribution							Normal Distribution vs. Lognormal Distribution						
	t	Welch	WMW	NBTT	NBWT	NBWR	PTT	t	Welch	WMW	NBTT	NBWT	NBWR	PTT
	Normal Distribution vs. Gamma Distribution							Lognormal Distribution vs. Normal Distribution						
5,3	0.0648	0.0557	0.0000	0.0587	0.0498	0.0004	0.0000	0.0723	0.0199	0.0000	0.0795	0.0400	0.0005	0.0000
5,5	0.0453	0.0385	0.0315	0.0410	0.0425	0.0364	0.0315	0.0207	**0.0104**	0.0284	0.0234	0.0219	0.0313	0.0284
7,10	0.0276	0.0319	0.0341	0.0271	0.0304	0.0428	0.0171	**0.0052**	**0.0083**	**0.0137**	**0.0058**	**0.0129**	0.0188	0.0040
10,7	0.0685	0.0536	0.0540	0.0709	0.0577	0.0520	0.0386	0.0377	**0.0119**	0.0271	0.0483	0.0266	0.0222	0.0212
10,10	0.0438	0.0407	0.0511	0.0427	0.0439	0.0565	0.0175	**0.0130**	**0.0079**	0.0178	0.0166	0.0161	0.0215	0.0035
	Lognormal Distribution vs. Lognormal Distribution							Lognormal Distribution vs. Gamma Distribution						
3,3	0.0524	0.0298	0.0000	0.0000	0.0000	0.0000	0.0000	0.0586	0.0356	0.0000	0.0002	0.0002	0.0000	0.0000
3,5	0.0361	0.0260	0.0000	0.0312	0.0239	0.0002	0.0000	0.0278	0.0273	0.0000	0.0226	0.0221	0.0001	0.0000
5,3	0.0896	0.0678	0.0000	0.0910	0.0782	0.0006	0.0000	0.1069	0.0606	0.0000	0.1130	0.0812	0.0009	0.0000
5,5	0.0668	0.0525	0.0699	0.0676	0.0672	0.0746	0.0699	0.0537	0.0416	0.0561	0.0541	0.0550	0.0579	0.0561
7,10	0.0437	0.0422	0.0708	0.0460	0.0468	0.0922	0.0297	0.0272	0.0315	0.0413	0.0274	0.0336	0.0532	0.0166
10,7	0.0834	0.0718	0.1014	0.0931	0.0861	0.0806	0.0701	0.0694	0.0444	0.0649	0.0764	0.0600	0.0479	0.0444
10,10	0.0594	0.0552	0.1012	0.0661	0.0662	0.1133	0.0309	0.0391	0.0350	0.0568	0.0416	0.0427	0.0629	0.0160
	Gamma Distribution vs. Normal Distribution							Gamma Distribution vs. Lognormal Distribution						
3,3	0.0233	**0.0122**	0.0000	0.0002	0.0002	0.0000	0.0000	0.0300	0.0172	0.0000	0.0001	0.0000	0.0000	0.0000
3,5	**0.0056**	**0.0076**	0.0000	0.0045	**0.0058**	0.0000	0.0000	0.0315	**0.0148**	0.0000	0.0282	**0.0150**	0.0003	0.0000
5,3	0.0565	0.0219	0.0000	0.0567	0.0288	0.0005	0.0000	0.0458	0.0486	0.0000	0.0392	0.0430	0.0004	0.0000
5,5	0.0212	**0.0131**	0.0222	0.0224	0.0222	0.0244	0.0222	0.0417	0.0315	0.0445	0.0444	0.0446	0.0488	0.0445
7,10	**0.0066**	**0.0110**	**0.0113**	**0.0067**	**0.0130**	0.0152	**0.0055**	0.0354	0.0268	0.0536	0.0409	0.0326	0.0686	0.0241
10,7	0.0349	**0.0105**	0.0267	0.0427	0.0233	0.0212	0.0172	0.0595	0.0631	0.0796	0.0672	0.0701	0.0693	0.0512
10,10	0.0153	**0.0105**	0.0180	0.0175	0.0176	0.0212	0.0043	0.0509	0.0483	0.0825	0.0548	0.0561	0.0924	0.0265
	Gamma Distribution vs. Gamma Distribution													
3,3	0.0377	0.0213	0.0000	0.0000	0.0000	0.0000	0.0000							
3,5	0.0256	0.0192	0.0000	0.0228	0.0155	0.0002	0.0000							
5,3	0.0775	0.0558	0.0000	0.0758	0.0610	0.0004	0.0000							
5,5	0.0463	0.0367	0.0425	0.0463	0.0468	0.0457	0.0425							
7,10	0.0219	0.0242	0.0332	0.0214	0.0249	0.0457	**0.0123**							
10,7	0.0639	0.0486	0.0577	0.0688	0.0571	0.0462	0.0400							
10,10	0.0374	0.0348	0.0524	0.0384	0.0397	0.0564	**0.0142**							

* Bold entries indicate a type I error that is controlled.

Discussion

In this research, location testing between two populations was performed when the sample size was small with normal and non-normal distributions and equal variance. The results revealed that the test statistics based on a bootstrap could control a type I error in the same way as classical statistics. This is consistent with Dwivedi et al. (2017) who analyzed small sample sizes utilizing the nonparametric pooled bootstrap test with a pooled resampling method. The type I error probability for the Independent t-test was examined in this study. The test statistics were the Welch t-test, exact Wilcoxon rank sum test, asymptotic permutation test, and nonparametric bootstrap t-test. All the tests practically controlled type I error probability when variances were equal between groups, regardless of normal or skew-normal distribution. Nonetheless, the permutation t-test could not perform well enough compared to bootstrap test statistics. (Ahad, Abdullah, Lai, & Ali, 2012).

Nevertheless, with the same criteria but unequal variance, the outcome of the highest-test statistics is the Welch t-test. This is similar to Boos and Brownie (1988) and Dwivedi et al. (2017) who both claimed that the Welch t-test has the highest test outperformance followed by the nonparametric bootstrap t-test under a normal population distribution. Also, the highest outperformance of the Welch t-test was greater than the nonparametric bootstrap t (pooled t statistic and Welch t statistics) under a non-normal population distribution.

Conclusion and Suggestions

In summary, firstly, when a population has a normal distribution, lognormal distribution, gamma distribution and homogeneity variances, classical statistics (Independent t-test, Welch t-test, Wilcoxon-Mann-Whitney test) and alternative statistics (nonparametric Bootstrap t-test, nonparametric bootstrap welch t-test, nonparametric bootstrap Welch t based on rank, Permutation t-test) could control type I errors. Secondly, when at least one population has a normal distribution and higher variance than another group, the Welch t-test could best control type I errors. Thirdly, if the small sample size was (3,3), the results showed that many test statistics could not reject a null hypothesis which is affected by type I errors. Finally, it could be concluded that hypothesis testing with small sample sizes must be careful when generalizing the results of the research. This is because some cases may present a problem in maintaining a type I error, especially in cases of interaction between non-normal distribution and unequal variances.

References

Ahad, N. A., Abdullah, S., Lai, C. H., & Ali, N. M. (2000). Relative power performance of t-test and bootstrap procedure for two samples. *Pertanika Journal of Science & Technology*, *20*, 43-52.

Altman, D. G., Gore, S. M., & Gardner, M. J. (1983). Statistical guidelines for contributors to medical journals. *British Medical Journal (Clinical Research Ed.)*, *286*, 1489-1493.

Barber, J. A., & Thompson, S. G. (2000). Analysis of cost data in randomized trials: an application of the non-parametric bootstrap. *Statistics in Medicine*, *19*, 3219-3236.

Boos, D. D., & Brownie, C. (1988). *Bootstrap p-values for tests of nonparametric hypotheses.* Institute of Statistics Mimeo Series No. 1919, North Carolina State University.

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

Bridge, P. D., & Sawilowsky, S. S. (1999). Increasing physicians' awareness of the impact of statistics on research outcomes: comparative power of the t-test and Wilcoxon rank-sum test in small samples applied research. *Journal of Clinical Epidemiology*, *52*, 229-235.

Dwivedi, A. K., Mallawaarachchi, I., & Alvarado, L. A. (2017). Analysis of small sample size studies using nonparametric bootstrap test with pooled resampling method. *Statistics in Medicine*, *36*, 2187-2205.

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

Fagerland, M. W., & Sandvik, L. (2009). Performance of five two-sample location tests for skewed distributions with unequal variances. *Contemporary Clinical Trials*, *30*, 490-496.

Haidous, N. H., & Sawilowsky, S. (2013). Robustness and power of the Kornbrot rank difference, signed ranks, and dependent samples t-test. *American Journal of Applied Mathematics and Statistics*, *1*, 99-102.

Hall, P., & Martin, M. (1998). On the bootstrap and two sample problems. *Australian Journal of Statistics*, *30A*, 179-192.

Janusonis, S. (2009). Comparing two small samples with an unstable, treatment-independent baseline. *Journal of Neuroscience Methods*, *179*, 173-178.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., & Donahue, B. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, *68*, 350-386.

Mann, H. B., & Whitney, D. R. (1947). On a Test of Whether One of Two Random Variables is Stochastically Larger than the other. *Annals of Mathematical Statistics*, *18*, 50-60

Mundry, R., & Fischer, J. (1998). Use of statistical programs for nonparametric tests of small samples often leads to incorrect P values: examples from Animal Behaviour. *Animal Behaviour*, *56*, 256-259.

Nguyen, D. T., Kim, E. S., Gil, P. R., Kellermann, A., Chen, Y. H., & Kromrey, J. D. (2016). Parametric Tests for Two Population Means under Normal and Non-Normal Distribution. *Journal of Modern Applied StatisticalMethods*, *15*, 141-159.

Posten, H. O. (1982). Two-sample Wilcoxon power over the Pearson system and comparison with t-test. *Journal of Statistical Computation and Simulation*, *16*, 1-18.

Reiczigel, J., Zakarias, I., & Rozsa, L. (2005). A bootstrap test of stochastic equality of two populations. *The American Statistician*, *59*, 1-6.

Ruthsatz, J., & Urbach, J. B. (2012). Child prodigy: A novel cognitive profile places elevated general intelligence, exceptional working memory and attention to detail at the root of prodigiousness. *Intelligence*, *40*, 419-426.

Sawilowsky, S. S., & Hillman, S. B. (1993). Power of the independent samples t-test under a prevalent psychometric measure distribution. *Journal of Consulting and Clinical Psychology*, *60*, 240-243.

Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

Siegel, S., & Castellan, N. J. (1998). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

Snyder, P. A., & Thompson, B. (1998). Use of tests of statistical significance and other analytic choices in a school psychology journal: Review of practices and suggested alternatives. *School Psychology Quarterly*, *13*, 335-348.

Stonehouse, J. M., & Forrester, G. J. (1998). Robustness of the t and U tests under combined assumption violations. *Journal of Applied Statistics*, *25*, 63-74.

Tanizaki, H. (1994). Power comparison of non-parametric tests: small-sample properties from Monte Carlo experiments. *Journal of Applied Statistics*, *24*, 603-632.

Welch, B. L. (1937). The significance of the difference between two means when the population variances are unequal. *Biometrika*, *29*, 350-362.

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics*, *1*, 80-83.

Weber, M., & Sawilowsky, S. (2009). Comparative power of the independent t, permutation t, and Wilcoxon tests. *Journal of Modern Applied Statistical Methods*, *8*, 10-15.

Winter, J. C. F. (2013). Using the Student's t-test with extremely small sample sizes. *Practical Assessment, Research, and Evaluation*, *18*, 1-12.