ISSN: 0128-7680
Pertanika J. Sci. & Technol. **20 (1)**: 43 – 52 (2012)

Relative Power Performance of *t*-test and Bootstrap Procedure for Two-Sample

Nor Aishah Ahad[1*], Suhaida Abdullah[1], Lai Choo Heng[2] and Nazihah Mohd. Ali[3]
[1]UUM College of Arts and Sciences, Universiti Utara Malaysia, 06010 Sintok, Kedah
[2]School of Distance Education, Universiti Sains Malaysia, 11800, Pulau Pinang
[3]Jabatan Matematik, Fakulti Sains, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia
[*]E-mail: aishah@uum.edu.my

ABSTRACT

The classical procedures of comparing two groups, such as *t*-test are, usually restricted with the assumptions of normality and equal variances. When these assumptions are violated, the rates of the Type I errors of the independent samples *t*-test are affected, particularly when the sample sizes are small. In this situation, the bootstrap procedure has an advantage over the parametric *t*-test. In this study, the performances of the bootstrap procedure and the independent sample *t*-test were investigated. The investigation focused on the power of both the test procedures to compare the two groups under different design specifications for normal and chi-square distributions. The results showed that the bootstrap procedure has a slight edge over the conventional *t*-test in term of the rate of achieving the benchmark level for both the distributions. In fact, the bootstrap procedure consistently outperformed the conventional *t*-test across all the combinations of the test conditions.

Key words: Bootstrap, power, *t*-test

INTRODUCTION

The classical procedures of comparing two groups such as *t*-test are usually restricted by the assumptions of normality and equal variances. Moreover, in the real world, these assumptions are not always fulfilled. Over the years, many procedures arose to handle the violation of these assumptions. Note that nonparametric procedures are viable alternatives that can be used when the distribution is not normal. Meanwhile, robust hypothesis testing procedures, such as James (1951), Welch (1951) and also Alexander and Govern (1994), are some examples of the procedures that have been developed to handle the problem of unequal variances.

With the increase of computational power, the statistical technique has also improved consistently. Computational intensive techniques have benefited from these and made a come back in terms of their usage. Bootstrapping is one of the recently revived techniques used for making certain kinds of statistical inferences. The essence of bootstrapping is the idea that, in the absence of any other knowledge about a population, the distribution of values found in a random sample of size *n* from the population is the best guide to the distribution in the population (Manly, 1997). In situation where classical test assumptions are not met, the bootstrap procedure offers a viable

[*]Corresponding Author

alternative as it uses computational power to perform intricate calculation. More importantly, the bootstrap procedure does not rely on a theoretical sampling distribution (such as central limits theorem that requires large samples) as in the classical tests. Othman, Keselman, Padmanabhan, Wilcox and Fradette (2003) listed out a practical advantage of using the bootstrap procedure. They noted that certain variations of the bootstrap procedure do not require the knowledge of the sampling distribution of the test statistic, and thereby, not requiring explicit expressions for standard errors of estimators. This condition makes hypothesis testing quite flexible.

Meanwhile, Krishnamoorthy, Lu and Mathew (2007) made a comparison between the bootstrap procedure and the general *F* test, Welch test and James test. The researchers suggested using the bootstrap method for the reason that it is the only procedure that performs satisfactorily, regardless of the sample sizes, values of the error variances, and the number of means being compared under unequal variances.

In the same vein, Higgins (2005) adopted and simplified the process of bootstrapping. The bootstrap method relies on Monte Carlo random number generators to generate bootstrap samples. In addition to carrying out hypothesis testing with the test statistic of unknown sampling distribution, the bootstrap procedure is also capable of assessing the performance of statistical test in terms of Type I error and power. Therefore, this facilitates the design of the new test procedure without resorting to proving them analytically first. Moreover, the bootstrap procedure is also capable of conducting sensitivity analyses on the performance of known parametric, nonparametric or Monte Carlo based methods under usual and extreme operating conditions. Hence, these sensitivity analyses are also comparative analyses if the analyses are performed on both the new and known procedures.

This study was conducted to evaluate the performance of the bootstrap method as compared to the classical pooled variance *t*-test in terms of their statistical power for the two groups. These tests were evaluated at various combinations of test conditions (namely, sample sizes, variances ratios and underlying distributions).

DESIGN OF THE STUDY

For comparison purposes, this study considered unbalanced groups with the sample sizes of 5 and 15 and as well as investigated the performance of the procedures under normal and chi-square distributions. The total sample size was set to be equivalent to 20 so as to maintain the small sample size (for each sample size and both) because the effect of non-normality would when dealing with small sample size (i.e. less than 30). In this study, the ratio of the group sample size (1:3) was to reflect the unbalanced design. Other researchers also had their own preference, such as Lix and Keselman (1998) whereby they used a group sample size ratio of 1:2. In order to examine the effect of distributional shapes on power value, this study used data from normal distribution as well as from skewed distribution. Chi-square with three degrees of freedom has skewness of 1.63 and kurtosis 4.00 was chosen to simulate a positively skewed distribution with a minimum at zero (Soong, 1981).

Meanwhile, the variance ratios used were 1:1 and 1:9. For heterogeneous variances, the ratio 1:9 was chosen as it reflects extreme variance heterogeneity. Hess, Olejnik and Huberty (2001), in their study, considered any variance ratio larger than 1:8 as extreme variance heterogeneity. Accordingly, this study sought to investigate how well both the tests performed under more extreme condition. These unequal variances were paired with unbalanced sample sizes positively and negatively. Positive pairing was done when the group with the smallest sample size was paired with the smallest variance while the group with the largest sample size was paired with the largest variance. For the negative pairing, the group with the largest sample size was paired with the smallest variance and the group with the smallest sample size was paired with the largest variance.

METHODS

This study was based on the simulated data generated by subroutine RANDGEN from SAS (1999). The nominal level of significance was set at α=0.05. For each condition examined, 1000 data sets were generated and for the bootstrap procedure, 1000 more bootstrap samples were generated for each of the data set. In the real world studies, the usual recommended bootstrap replicate is 100 and this was done with the trade-off between computational cost and accuracy. Meanwhile, Pattengale, Alipour, Bininda-Emonds, Moret and Stamatakis (2010) recommended bootstrap replicates of 100 to 500 range for the real world data where a practitioners would no longer have to enter a guess nor worry about the quality of estimates. Most conservative criteria may for several thousand replicates which only reduce the effects of random sampling errors. Wilcox (2005) stated that the choice of the number of bootstrap replicates has to be sufficiently large so that if the seed in the random number generator is altered, the same conclusions would essentially still be obtained. Chernick (2008) urged practitioners to choose the number of bootstrap replications that make the sampling variance sufficiently small so as to ensure that the bootstrap approximation is close to the actual bootstrap estimate. He further suggested at least 1000 bootstrap replications in the case of confidence interval estimation and hypothesis testing problem.

The power of a test is the probability that a test will yield statistically significant results or reject the false null hypothesis. It can be denoted as $1 - \beta$. As reported by Yin and Othman (2009), the power of a test depends on three factors, namely, significance criterion (α), sample size (n) and effect size (EF). α, which is also known as Type I error, is positively related to power. The power values increase as the standard for determining significance increases. The sample size, n, is also positively related to power. The larger the sample size, the higher the power value. The effect size, EF, is the magnitude of the effect under alternate hypothesis. The power values increase when the effect size increases.

For the two groups, the effect size index is the effect size that is to be detected. Thus, the effect size index (d) is defined as:

$$d = \frac{|\mu_1 - \mu_2|}{\sigma}$$

Where,

d = effect size index

$\mu_1 - \mu_2$ = population means

σ = the standard deviation of either population (if they are equal)/the smallest standard deviation (if they are unequal).

According to Cohen (1988), the effect size is considered as small when $d = 0.2$, medium when $d = 0.5$, and large when $d = 0.8$. The effect size index is used to choose the values of population means for the true alternative hypothesis. Assuming that $\mu_1 > \mu_2$ and $\mu_2 = 0$, we can have a range of values to represent μ_1. In order to cover all the conditions, the value of the shift parameter can systematically varied from 0.5 to 7.5 with the increments of 0.5. Table 1 represents the possible choices for μ_1 and μ_2 for all the conditions and methods. Murphy and Myors (1998) noted that the power of a test is usually judged to be adequate if the value is 0.8 and above. *Fig. 1* outlines the steps to determine the power value for both the *t*-test and bootstrap procedure.

RESULTS AND DISCUSSION

The entries in Table 2 and Table 3 are power rates for the *t*-test and bootstrap procedure, respectively, on both the test distributions. Similarly, Tables 4 and 5 show the power values of both the procedures

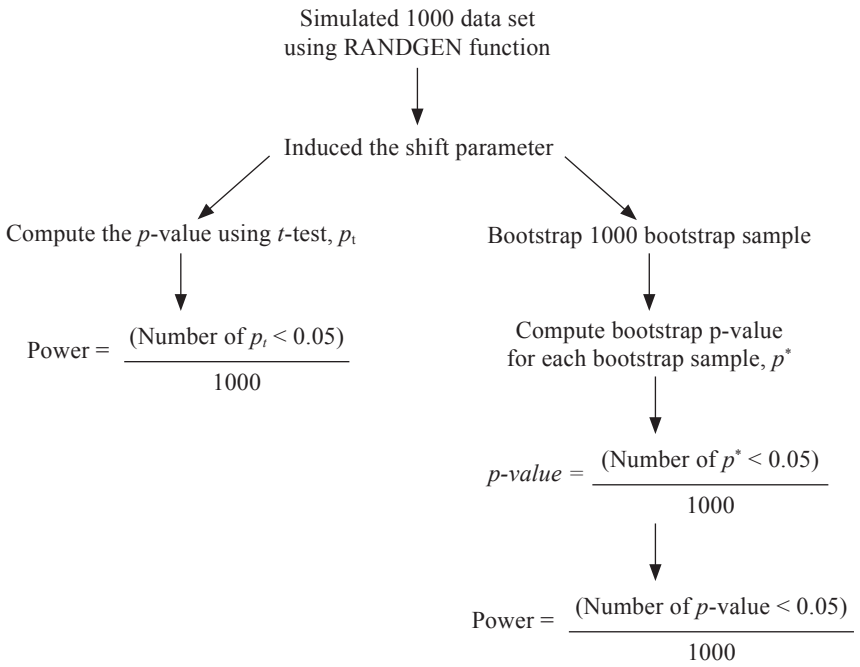

Fig. 1: A diagrammatic outline to determine the power values for the t-*test and the bootstrap procedure*

TABLE 1
Possible choices of shift parameter for all conditions and methods

Group sample sizes	Group variances	Effect size	Range for shift parameter
{5, 15}	{1, 1}	$\frac{\mu_1 - \mu_2}{0.547}$	0.5, 1, 1.5, …
{5, 15}	{1, 9}	$\frac{\mu_1 - \mu_2}{1.449}$	1.5, 3, 4.5, …
{5, 15}	{9, 1}	$\frac{\mu_1 - \mu_2}{0.949}$	1, 2, 3, …
Bootstrap	difference	$\mu_1 - \mu_2$	1, 2, 3, …

for the normal and chi-square distributions. On the other hand, *Fig. 2* and *Fig. 3* displayed the power curves of the *t*-test and bootstrap procedure. The results and discussion for the *t*-test method and bootstrap procedure are separately presented.

The t-test Method

Table 2 shows the power rates of the pooled variance *t*-test under normal and chi- square distributions. The bolded entries are power rates that are above the benchmark power rate of 0.80. For a clear viewing, *Fig. 2* displays the power curves of the pooled variance *t*-test under normal and chi-square distributions with all the variance ratios that have been considered.

As shown in Table 2 and *Fig. 2*, the homogeneous variances produced higher power rates under normal distribution than the heterogeneous variances, whereas in terms of pairing, the negative pairing produced higher power rates than positive pairing. The power rates of the pooled variance *t*-test behave similarly under the chi-square distribution, where the highest power rates are in homogeneous variances while in terms of pairing, the negative pairing is higher than positive. Nonetheless, the performance of the *t*-test does not vary between the two distributions. The power rate reaches the benchmark point of 0.8 at almost the same level. However, the chi-square distribution achieved the benchmark rate at a lower shift parameter when the variances are negatively paired.

The Bootstrap Procedure

Table 3 and *Fig. 3* show the power values and power curves for the bootstrap procedure, respectively. The bolded entries in Table 3 are power rates that exceed the benchmark power rate of 0.80. The bootstrap procedure produces similar trends as that of the pooled variance *t*-test. Under normal distribution, the high power rate was achieved the fastest in the presence of variance homogeneity, while the slowest was when the pairing of variance and sample size was positive.

For the chi-square distribution, once again just like in the pooled variance *t*-test, the results of power rates were found as not incredibly different with that of the normal distribution where groups with homogeneous variances attained the high power the fastest. While in terms of pairing, the negative pairing reached the highest power earlier than the positive pairing.

When the two distributions were compared under normal distribution with homogeneous variances and positive pairing, the results showed that the bootstrap procedure reached the power of 0.80 much faster as compared to the chi-square distribution. However, for the groups with the negative pairing, the performance of the bootstrap procedure under the chi-square distribution was shown to be better as compared to the normal distribution because at shift parameter = 2.5, the power is 0.81 whereas the power for normal distribution is just 0.765.

Comparative Study

For the comparative study, an attempt was done to determine which of the two procedures would be the fastest in term of achieving the benchmark power. The detail of the comparative study is shown in Table 4 for the normal distribution and Table 5 for the chi-square distribution.

By comparison, the bootstrap procedure has a slight edge over the conventional *t*-test in term of the rate of achieving the benchmark level with the normal distribution. The entries for the bootstrap procedures across various variance ratios exceeded the benchmark level relatively faster as compared to the *t*-test. Similarly, with the chi-square distribution, the bootstrap procedure is capable of achieving a faster rate. Conclusively, the bootstrap procedure consistently outperformed the conventional *t*-test across all the combinations of the test conditions.

CONCLUSIONS

The bootstrap procedure has a slight edge over the conventional *t*-test in terms of the rate of achieving the benchmark level for both the normal and chi-square distributions. In particular, the bootstrap procedure consistently outperformed the conventional *t*-test across all the combinations of the test conditions.

Evidently, both the tests are capable of achieving the benchmark level at a faster rate in the situation of homogeneity of variances across distributions. Similarly, they are sensitive to effect size

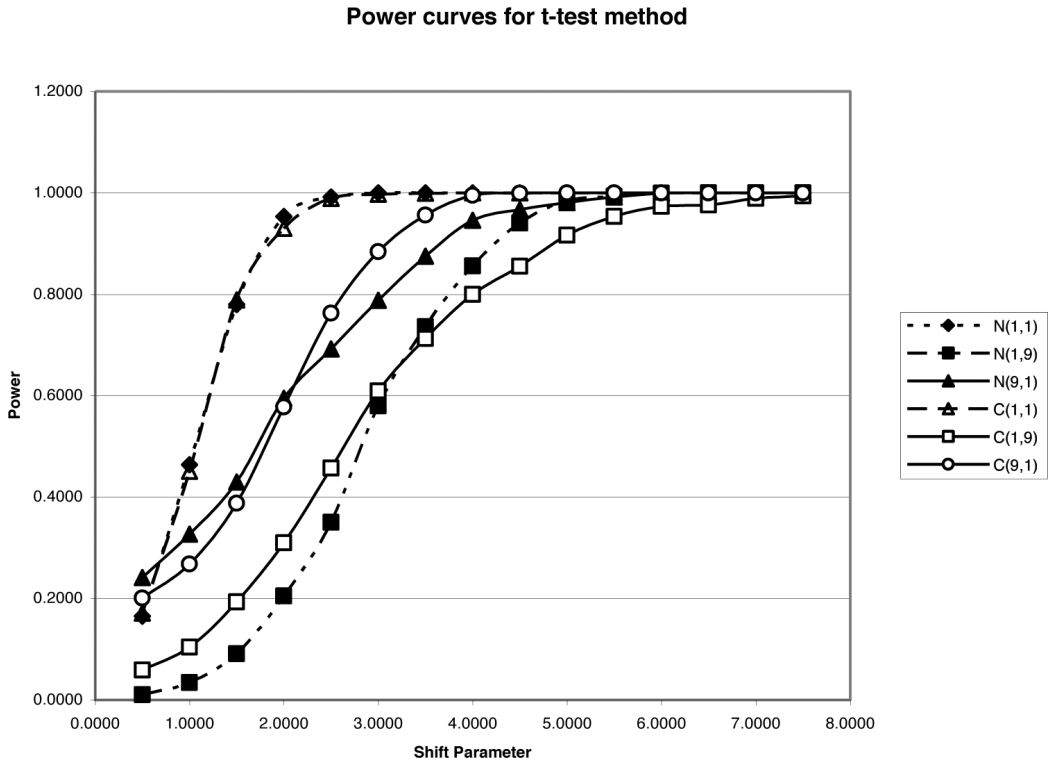

Fig. 2: Power curves for the pooled variance t-test

TABLE 2
Power values for the two sample *t*-test with pooled variances

Shift parameter	Normal			Chi-square (3)		
	Var (1,1)	Var (1,9)	Var (9,1)	Var (1,1)	Var (1,9)	Var (9,1)
0.5	0.165	0.010	0.241	0.171	0.059	0.201
1.0	0.464	0.034	0.327	0.451	0.104	0.268
1.5	0.780	0.091	0.430	0.789	0.193	0.388
2.0	**0.953**	0.205	0.595	**0.930**	0.310	0.578
2.5	**0.992**	0.350	0.692	**0.989**	0.457	0.763
3.0	**1.000**	0.580	0.788	**0.997**	0.609	**0.884**
3.5	**1.000**	0.736	**0.875**	**0.999**	0.713	**0.956**
4.0	**1.000**	**0.856**	**0.946**	**1.000**	**0.800**	**0.995**
4.5	**1.000**	**0.940**	**0.967**	**1.000**	**0.855**	**0.999**
5.0	**1.000**	**0.984**	**0.981**	**1.000**	**0.916**	**1.000**
5.5	**1.000**	**0.993**	**0.992**	**1.000**	**0.953**	**1.000**
6.0	**1.000**	**0.999**	**0.999**	**1.000**	**0.973**	**1.000**
6.5	**1.000**	**1.000**	**1.000**	**1.000**	**0.976**	**1.000**
7.0	**1.000**	**1.000**	**1.000**	**1.000**	**0.989**	**1.000**
7.5	**1.000**	**1.000**	**1.000**	**1.000**	**0.994**	**1.000**

Note: Values indicated in bold show power rate ≥ 0.80.

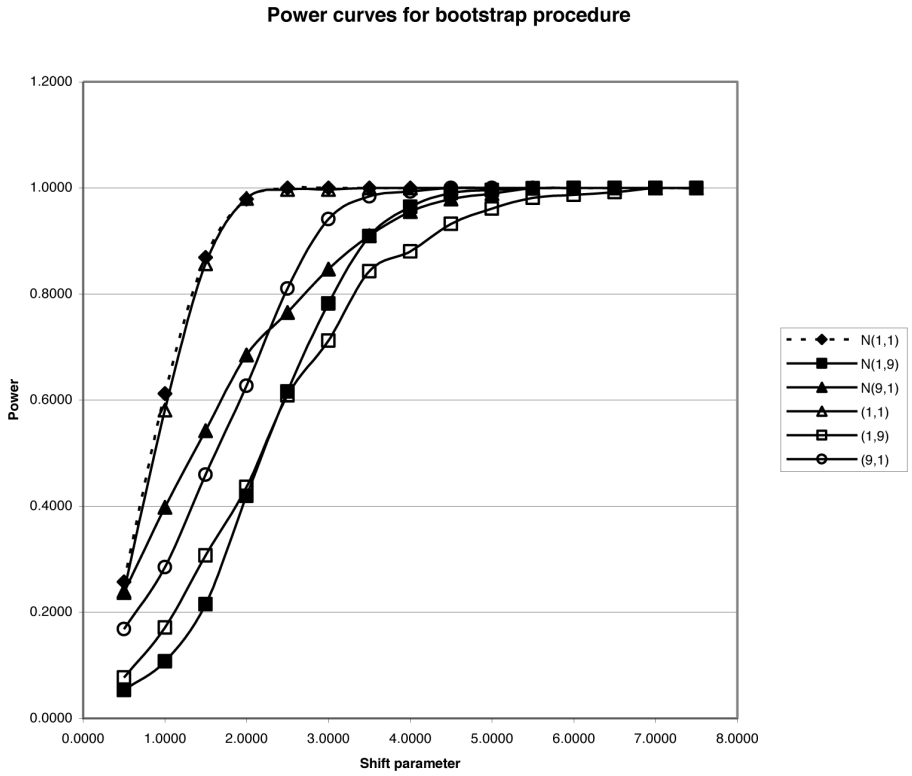

Fig. 3: Power curves for the bootstrap procedure

TABLE 3
Power values for the bootstrap procedure

Shift parameter	Normal			Chi-square (3)		
	Var (1,1)	Var (1,9)	Var (9,1)	Var (1,1)	Var (1,9)	Var (9,1)
0.5	0.257	0.053	0.237	0.240	0.077	0.168
1.0	0.612	0.107	0.398	0.581	0.171	0.285
1.5	**0.869**	0.215	0.542	**0.857**	0.307	0.459
2.0	**0.979**	0.419	0.685	**0.980**	0.436	0.627
2.5	**1.000**	0.616	0.765	**0.997**	0.609	**0.810**
3.0	**1.000**	0.782	**0.847**	**0.997**	0.712	**0.941**
3.5	**1.000**	**0.909**	**0.910**	**1.000**	**0.843**	**0.984**
4.0	**1.000**	**0.964**	**0.955**	**1.000**	**0.880**	**0.993**
4.5	**1.000**	**0.990**	**0.979**	**1.000**	**0.932**	**1.000**
5.0	**1.000**	**0.996**	**0.989**	**1.000**	**0.961**	**1.000**
5.5	**1.000**	**0.999**	**0.999**	**1.000**	**0.981**	**1.000**
6.0	**1.000**	**1.000**	**1.000**	**1.000**	**0.987**	**1.000**
6.5	**1.000**	**1.000**	**1.000**	**1.000**	**0.992**	**1.000**
7.0	**1.000**	**1.000**	**1.000**	**1.000**	**0.999**	**1.000**
7.5	**1.000**	**1.000**	**1.000**	**1.000**	**0.999**	**1.000**

Note: Values indicated in bold show power rate ≥ 0.80.

TABLE 4
Power values for the normal distribution

Shift parameter	Var (1,1)		Var (1,9)		Var (9,1)	
	t-test	Bootstrap	t-test	Bootstrap	t-test	Bootstrap
0.5	0.165	0.257	0.010	0.053	0.241	0.237
1.0	0.464	0.612	0.034	0.107	0.327	0.398
1.5	0.780	**0.869**	0.091	0.215	0.430	0.542
2.0	**0.953**	**0.979**	0.205	0.419	0.595	0.685
2.5	**0.992**	**1.000**	0.350	0.616	0.692	0.765
3.0	**1.000**	**1.000**	0.580	0.782	0.788	**0.847**
3.5	**1.000**	**1.000**	0.736	**0.909**	**0.875**	**0.910**
4.0	**1.000**	**1.000**	**0.856**	**0.964**	**0.946**	**0.955**
4.5	**1.000**	**1.000**	**0.940**	**0.990**	**0.967**	**0.979**
5.0	**1.000**	**1.000**	**0.984**	**0.996**	**0.981**	**0.989**
5.5	**1.000**	**1.000**	**0.993**	**0.999**	**0.992**	**0.999**
6.0	**1.000**	**1.000**	**0.999**	**1.000**	**0.999**	**1.000**
6.5	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
7.0	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
7.5	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**

Note: Values indicated in bold show power rate ≥ 0.80.

TABLE 5
Power Values for the Chi-Square Distribution

Shift parameter	Var (1,1)		Var (1,9)		Var (9,1)	
	t-test	Bootstrap	t-test	Bootstrap	t-test	Bootstrap
0.5	0.171	0.240	0.059	0.077	0.201	0.168
1.0	0.451	0.581	0.104	0.171	0.268	0.285
1.5	0.789	**0.857**	0.193	0.307	0.388	0.459
2.0	**0.930**	**0.980**	0.310	0.436	0.578	0.627
2.5	**0.989**	**0.997**	0.457	0.609	0.763	**0.810**
3.0	**0.997**	**0.997**	0.609	0.712	**0.884**	**0.941**
3.5	**0.999**	**1.000**	0.713	**0.843**	**0.956**	**0.984**
4.0	**1.000**	**1.000**	**0.800**	**0.880**	**0.995**	**0.993**
4.5	**1.000**	**1.000**	**0.855**	**0.932**	**0.999**	**1.000**
5.0	**1.000**	**1.000**	**0.916**	**0.961**	**1.000**	**1.000**
5.5	**1.000**	**1.000**	**0.953**	**0.981**	**1.000**	**1.000**
6.0	**1.000**	**1.000**	**0.973**	**0.987**	**1.000**	**1.000**
6.5	**1.000**	**1.000**	**0.976**	**0.992**	**1.000**	**1.000**
7.0	**1.000**	**1.000**	**0.989**	**0.999**	**1.000**	**1.000**
7.5	**1.000**	**1.000**	**0.994**	**0.999**	**1.000**	**1.000**

Note: Values indicated in bold show power rate ≥ 0.80.

in this situation. Conversely, both the tests produced lower rate when large variance was associated with the larger sample, signifying a low sensitivity to effect size. Despite the slight difference in the rate, both the *t*-test and bootstrap procedures produced a consistent power performance across the two distributions. For each test, the rates of achieving the benchmark level were found to be similar for both the normal and chi-square distributions, but the bootstrap procedure achieved it at a faster rate.

REFERENCES

Alexander, R. A., & Govern, D. M. (1994). A new and simpler approximation for *ANOVA* under variance heterogeneity. *Journal of Educational Statistics, 19*, 91-101.

Chernick, M. R. (2008). *Bootstrap methods. A guide for practitioners and researchers (2nd ed)*. New York: John Wiley & Sons.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* New York: Academic Press.

Hess, B., Olejnik, S., & Huberty, C. J. (2001). The efficacy of two improvement-over-chance effect sizes for two-group univariate comparisons under variance heterogeneity and nonnormality. *Educational and Psychological Measurement*, *61*, 909-936.

Higgins, G. E. (2005). Statistical significance testing: The bootstrapping method and an application to self-control theory. *The Southwest Journal of Criminal Justice*, *2*(1), 54-75.

James, G. S. (1951). The comparison of several groups of observations when the ratios of the population variances are unknown. *Biometrika*, *38*, 324-329.

Krishnamoorthy, K., Lu, F., & Mathew, T. (2007). A parametric bootstrap approach for ANOVA with unequal variances: Fixed and random models. *Computational Statistics & Data analysis*, *51*, 5731-5742.

Lix, L. M., & Keselman, H. J. (1998). To trim or not to trim: Tests of location equality under heteroscedasticity and nonnormality. *Educational and Psychological Measurement*, *58*(3), 409-429.

Manly, B. F. J. (1997). *Randomization, bootstrap and Monte Carlo methods in biology (2nd ed)*. London: Chapman & Hall.

Murphy, K. R., & Myors, B. (1998). *Statistical power analysis*: *A simple and general model for traditional and modern hypothesis tests*. Mahwah, NJ: Lawrence Erlbaum.

Othman, A. R., Keselman, H. J., Padmanabhan, A. R., Wilcox, R. R., & Fradette, K. (2003). An improved robust Welch-James test statistic. *In the proceeding of the Regional Conference on Integrating Technology in the Mathematical Sciences, 2003.* Universiti Sains Malaysia, Pulau Pinang, Malaysia.

Pattengale, N. D., Alipour, M., Bininda-Emonds, O. R., Moret, B. M., & Stamatakis, A. (2010). How many bootstrap replicates are necessary? *Journal of Computational Biology, 17*(3), 337-354.

SAS Institute Inc. (1999). *SAS/IML User's Guide version 8.* Cary, NC: SAS Institute Inc.

Soong, T. T. (1981). *Probabilistic modeling and analysis in science and engineering*. New York: John Wiley & Sons.

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika*, *38*, 330-336.

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing* (2nd ed). New York: Academic Press.

Yin, T. S., & Othman, A. R. (2009). When does the pooled variance *t*-test fail? *African Journal of Mathematics and Computer Science Research*, *2*(4), 56-62.