Malaysian Journal of Mathematical Sciences 4(1): 85-94 (2010)

Performance of the Traditional Pooled Variance *t*-Test against the Bootstrap Procedure of Difference Between Sample Means

[1,3]Teh Sin Yin, [2]Zahayu Md Yusof,
[3]Che Rohani Yaacob, [1,4]Abdul Rahman Othman
[1]School of Distance Education, Universiti Sains Malaysia
11800 USM Pulau Pinang, Malaysia
[2]Kolej Sastera dan Sains, Universiti Utara Malaysia
06010 Sintok, Kedah, Malaysia
[3]School of Mathematical Sciences, Universiti Sains Malaysia
11800 USM Pulau Pinang, Malaysia
[4]Institute of Postgraduate Studies, Universiti Sains Malaysia
11800 USM Pulau Pinang, Malaysia
E-mail: [1,3]syin.teh@gmail.com, [2]zahayu@uum.edu.my, [3]rohani@cs.usm.my,
[1,4]oarahman@usm.my

ABSTRACT

In this study, two methods of comparing the means of two samples were conducted. The first method used the traditional pooled variance *t*-test while the second method used the bootstrap method of comparing means. In the second method, pseudo sampling distributions were generated for the normal and non-normal distributions. For the non-normal we used the χ_3^2 and a *g*=.5, *h*=.5 distribution. Group sizes {5, 15} and {15, 25} with equal and unequal variances were generated and for each method Type I error rates were evaluated. We found that about 20% and 33% of the study conditions using the pooled variance *t*-test were robust when group sizes are {5, 15} and {15, 25}, respectively. Whereas, about 33% and 47% of the study conditions in bootstrapped procedure were robust when group sizes are {5, 15} and {15, 25}, respectively.

Keywords: *t*-test, Monte Carlo, bootstrap, Type I error

INTRODUCTION

The traditional *t*-test is usually affected by nonnormality and variance heterogeneity. Departures from normality originate from two problems, that is, skewness and the existence of outliers. These problems can be remedied by using transformations such as exponential and logarithm but sometimes, even after the transformation, problems with nonnormal data still occur. Simple transformations of the data such as the logarithm can reduce skewness but not for complex transformations such as those in the the class of Box-Cox transformations (Wilcox & Keselman, 2003).

The recent study of Teh and Othman (2009) showed when the pooled variance *t*-test Type I error rates get closer to the nominal value of 0.05 and when they are not. Their findings indicate that unbalanced groups with small sample sizes and slight departure from variance homogeneity produced liberal or conservative Type I error rates. For unbalanced groups of large sample sizes, the test was considered not robust based on the criterion used when there is a slight departure from variance homogeneity, regardless of distributions. Their findings also showed that when the distribution was non-normal, Type I error of pooled variance *t*-test of 0.05 was not achievable when one of the group variance is larger than the other by about 10 units.

Another problem which researchers always encounter when using the classical methods is heteroscedasticity. Some of the parametric methods that can handle this problem are those proposed by James (1951), Welch (1951), and Alexander and Govern (1994). Unfortunately, all of these methods have difficulty in dealing with nonnormal data. Therefore, one way deal with the non-normalily problem is to use trimmed mean as the central tendency measure (Abdullah, *et al.*, 2008).

Some researchers sought for alternatives in the non-parametric methods, such as the Mann Whitney. However, these methods have low power (Wilcox, 1992). Even though the Mann Whitney test is distribution free, the distribution is assumed to be symmetric.

Another alternative is to use a Monte Carlo approach to deal with the problems of nonnormality and heteroscedasticity. The bootstrapped version of the test of difference between two group means was applied to evaluate the performance of a test procedure in terms of Type I error. Even with the bootstrap version of the existing methods stated earlier, the test statistic need to be defined. There is a simpler version of bootstrap test of difference between the two groups means which was introduced by Efron and Tibshirani (1993, p. 202). The reasons for using this simpler version are because it does not require any test statistic and it deals with differences between samples from a single population. Thus, the bootstrap scheme involves pooling the two samples together. Essentially, this action ensures that we are bootstrapping over a homogeneous variance situation.

In this study, our aim is to examine whether these two procedures, which are handicapped to work when the homogeneity of variance condition exists, will work effectively when the data are nonnormal, and group variances and sizes are unequal.

DESIGN SPECIFICATIONS AND METHODS

To evaluate the performance of the test procedures, two variables were manipulated. They were: (1) the type of distribution – normal or non-normal and (2) the nature of pairing of variances. We used unequal group sizes (5, 15) and (15, 25) with total sample sizes of the two groups being N = 20 and 40. For nonnormal distributions, we chose the chi-square distribution with three degrees of freedom (χ_3^2) to represent a mild skewness distribution and a g=.5, h=.5 (Hoaglin, 1985) distribution which is extremely skewed and heavy tailed. These two distributions were widely studied by researchers (Keselman, *et al.*, 2007; Othman, *et al.*, 2004; Syed Yahaya, *et al.*, 2004). In terms of variance heterogeneity, the ratios of 1:9 and 1:36 are used in this study. They are considered extreme variance conditions under which the efficacy of the tests should be examined (Alexander & Govern, 1994; Keselman, *et al.*, 2007; Neuhäuser & Hothorn, 2000).

Unequal group sizes, when paired with unequal variances, can affect Type I error control for tests that compare the typical score across groups (Keselman, *et al.*, 1998; Keselman, *et al.*, 2002; Othman, *et al.*, 2004; Syed Yahaya, *et al.*, 2006). Therefore, we positively and negatively paired the sample sizes and variances. A positive pairing occurs when the largest group size is associated with the largest group variance, while the smallest group size is associated with the smallest group variance. On the other hand, in a negative pairing, the largest group size is paired with the smallest group variance and the smallest group size is paired with the largest group variance.

This study was based on simulated data. In terms of data generation, we used the SAS generator RANDGEN (SAS Institute, 2004) to obtain pseudo-random standard normal variates (RANDGEN(Y, ‘NORMAL’)) and to generate the chi-squared variates with three degrees of freedom we used RANDGEN(Y, ‘CHISQUARE’, 3).

To generate data from a g and h distribution, standard normal random numbers Z were converted to g and h distributed random numbers via

$$Y = \frac{e^{gZ} - 1}{g} e^{\frac{hZ^2}{2}} \tag{1}$$

where both g and h are non-zero. The Z values were generated using the generator RANDGEN with the normal distribution option.

We then compared the performance of the traditional pooled variance t-test with the bootstrap procedure of difference between sample means. Here, the group means are set to {0, 0} as to reflect the null hypothesis of equal means (H$_0$: $\mu_1 = \mu_2$). The study conditions that we considered are as follows:

Group variances: {1, 1}, {1, 9}, {1,36}, {9, 1}, {36,1}
Group sample sizes: {5, 15}, {15, 25}
Distribution: Normal, Chi-square with 3 degree of freedom (χ_3^2) and g=.5, h=.5.

The design specifications are presented in Table 1.

Based on these conditions there are 3 distributions × 5 design specifications or 15 Monte Carlo studies for both methods (refer to Table 1). For each condition examined, 1000 bootstrap samples were obtained. The nominal level of significance was set at α=0.05.

TABLE 1: Design specification for two groups

Pairing	**Group Sizes**		**Population Variances**	
	1	**2**	**1**	**2**
Negative	5	15	9	1
	15	25		
	5	15	36	1
	15	25		
Equal	5	15	1	1
	15	25		
Positive	5	15	1	9
	15	25		
	5	15	1	36
	15	25		

MONTE CARLO STUDY ALGORITHM

The algorithm of the Monte Carlo study of Type I error rates of the traditional pooled variance t-test is given as Algorithm 1. The algorithm to obtain the p-value of the bootstrap method of comparing two means is given as Algorithm 2. In order to carry out a Monte Carlo study of Type I error rates of the bootstrap method of comparing two means, we replace the pooled variance *t*-test steps in Algorithm 1 with the all of the steps in Algorithm 2.

Algorithm 1

The algorithm for getting the Type I error rates of the traditional pooled variance *t*-test procedure is as follows:

1) Initialize a variable, count = 0
2) Generate data to reflect the null hypothesis of equal mean is true.
3) Calculate *t*-test statistic based on data generated in step 2.
4) Determine *p*-value of calculated *t*-test statistic in step 3.
5) If *p*-value ≤ 0.05, then increase count by one (count = count + 1).
5) Repeat steps 2 to step 5 for 1000 times.
6) Obtain the average Type I error rates by dividing count by 1000.
7) Repeat this simulation for 15 different conditions (3 distributions × 5 design specifications).

The following algorithm is used to determine the *p*-values of the bootstrap procedure of differences between two means.

Algorithm 2

These are the steps used to get the *p*-values of the bootstrap method.

1) Based on the two samples, calculate $\hat{\theta} = \overline{X} - \overline{Y}$.
2) Pool the sample points from both samples. We can do so because in H_0: $F = G$ where F and G are the population distributions of samples X and Y, respectively.
3) Let n_1 and n_2 be the sample sizes of the two samples and $n_1 + n_2 = N$. Draw randomly with replacement from this pool n_1 sample points. This will be x_1^*. The remainder N-n_1 will form y_1^*.
4) From step 3, obtain $\hat{\theta}_1^* = \overline{X}_1^* - \overline{Y}_1^*$.
5) Repeat step 3 and step 4 for 1000 times.
6) Obtain the Type I error rate of the test by calculating the number of $\hat{\theta}_b^* \geq \hat{\theta}$ and then dividing by 1000 where b =1, 2, …, 1000.

As mentioned earlier, in order to do the Monte Carlo study of Type I error rates of the replace steps 3 to 6 in Algorithm 1 with steps 1 to 6 in Algorithm 2.

RESULTS AND DISCUSSIONS

Table 2 and Table 3 show the Type I error rates of the *t*-test and bootstrap method, respectively, when group sizes are {5,15} or *N*=20. Table 4 and Table 5 show the Type I error rates of the two methods when group sizes are {15, 25} or *N*=40.

According to Bradley's (1978) liberal criterion of robustness, a test can be considered robust if its empirical rate of Type I error, $\hat{\alpha}$, is within the interval $0.5\alpha \le \hat{\alpha} \le 1.5\alpha$. Thus, if the nominal level is α = 0.05, the empirical Type I error rate should be within the interval $0.025 \le \hat{\alpha} \le 0.075$.

TABLE 2: Type I error rates of two sample *t*-test with pooled variance when group sizes {5, 15}.

Distribution	Group Variances				
	(1, 9)	(1,36)	(1, 1)	(9, 1)	(36,1)
Normal	0.003	0.001	**0.043***	0.209	0.274
χ_3^2	0.021	0.014	**0.049**	0.240	0.314
g=0.5, *h*=0.5	0.010	0.007	**0.047**	0.175	0.254

Note: *Bold values indicate Type I error within $0.025 \le \hat{\alpha} \le 0.075$.

TABLE 3: Type I error rates of the bootstrapped comparison of mean when group sizes {5, 15}.

Distribution	Group Variances				
	(1, 9)	(1,36)	(1, 1)	(9, 1)	(36,1)
Normal	0.006	0.003	**0.059**	0.159	0.186
χ_3^2	**0.031**	**0.033**	**0.072**	0.112	0.129
g=0.5, *h*=0.5	0.018	0.008	**0.061**	0.143	0.178

Note: *Bold values indicate Type I error within $0.025 \le \hat{\alpha} \le 0.075$.

TABLE 4: Type I error rates of two sample *t*-test with pooled variance when group sizes {15, 25}

Distribution	Group Variances				
	(1, 9)	(1,36)	(1, 1)	(9, 1)	(36,1)
Normal	0.011	0.009	**0.049**	0.110	0.115
χ_3^2	**0.025**	0.023	**0.043**	0.132	0.152
g=0.5, *h*=0.5	0.014	0.014	**0.033**	**0.074**	0.102

Note: *Bold values indicate Type I error within $0.025 \le \hat{\alpha} \le 0.075$.

TABLE 5: Type I error rates of the bootstrapped comparison of mean when group sizes {15, 25}.

Distribution	Group Variances				
	(1, 9)	(1,36)	(1, 1)	(9, 1)	(36,1)
Normal	0.017	0.014	**0.052**	0.083	0.095
χ_3^2	**0.044**	**0.045**	**0.045**	**0.059**	**0.062**
g=0.5, *h*=0.5	0.017	0.012	**0.045**	0.090	0.137

Note: *Bold values indicate Type I error within $0.025 \le \hat{\alpha} \le 0.075$.

Based on this criterion of robustness, about 20% (3 out of 15) and 33% (5 out of 15) of the study conditions using the *t*-test are robust when *N* = 20 and *N* = 40, respectively. Whereas, there are about 33% (5 out of 15) and 47% (7 out of 15) of the study conditions using the bootstrap method are considered robust when *N* =20 and *N* = 40, respectively.

In both methods, samples with equal variances have *p*-values that are within the Bradley's (1978) liberal criterion of robustness. However, the values nearest to the nominal level ($\alpha = 0.05$) are from the two sample *t*-test with pooled variance when the distribution is normal.

For the *t*-test method, current results reveal that positive pairings have conservative results and negative pairings have liberal results. This is in accord with findings in Othman, *et al.* (2004), and Teh and Othman (2009), that positive pairings produced conservative values, while negative pairings generated liberal values.

In the bootstrap method, the χ_3^2 -distribution with positive pairing showed Type I error rates which are robust for group sizes {5, 15}. Whereas, the same distribution with group sizes {15, 25} showed Type I error rates which are robust for both negative and positive pairing. In other words, the *p*-values of these conditions are within Bradley's (1978) liberal criterion of robustness. On the other hand, the Type I error rates of *g*=.5, *h*=.5 distribution are not robust when group variances are unequal, for group sizes {5, 15} and {15, 25}. However, the error rate is closer to the nominal level compared to normal distribution when group sizes are {5, 15}.

CONCLUSIONS

When comparing two means, our study showed that the pooled variance *t*-test method is slightly better than the bootstrap method for the equal variance case. Whereas, for the unbalanced design (unequal sample sizes with unequal variances), the bootstrap method produced better Type I error rate than the pooled variance *t*-test method. Thus, when assumptions are untenable and sampling distribution of test statistics unknown, the Monte Carlo remedy, that is, the simple version of bootstrap (Efron & Tibshirani, 1993) construction of the pseudo sampling distributions would enable us to conduct tests of hypotheses.

ACKNOWLEDGEMENT

The authors would like to acknowledge the work that led to this paper publication funded by the Fundamental Research Grant Scheme of the Malaysian Ministry of Higher Education, and supported by the Universiti Sains Malaysia Fellowship.

REFERENCES

Abdullah, S., Syed Yahaya, S. S. and Othman, A. R. 2008. A power investigation of Alexander Govern test using modified one-step m-estimator as a central tendency measure. *Proceedings of Joint Meeting of 4th. World Conference of IASC and 6th Conference of the Asian Regional Section of the IASC on Computational Statistics and Data Analysis*, Yokohama.

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

Bradley, J. V. 1978. Robustness?. *British J. Math. Statist. Psych.*. **31**: 321-339.

Efron, B., and Tibshirani, R. J. 1993. *An introduction to the bootstrap*. New York: Chapman and Hall.

Hoaglin, D. C. 1985. Summarizing shape numerically: The *g*- and *h*-distributions. In D. C. Hoaglin, F. Mosteller & J. Tukey (Eds.), *Exploring data tables, trends, and shapes* (pp. 461-513). New York: Wiley.

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

Keselman, H. J., Huberty, C. J., Lix, L. M., Oleijnik, S., Cribbie, R. A., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C., & Levin, J. R. 1998. Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA and ANCOVA analyses. *Review of Educational Research*. **68**(3): 350-386.

Keselman, H. J., Wilcox, R. R., Algina, J., Othman, A. R., and Fradette, K. 2002. Trimming, transforming statistics, and bootstrapping: Circumventing the biasing effects of heteroscedasticity and nonnormality. *Journal of Modern Applied Statistical Method*. **1**(2): 288-309.

Keselman, H. J., Wilcox, R. R., Lix, L. M., Algina, J., and Fradette, K. 2007. Adaptive robust estimation and testing. *British Journal of Mathematical and Statistical Psychology*, **60**: 267-293.

Neuhäuser, M., and Hothorn, L. A. 2000. Parametric location-scale and scale trend tests based on. Levene's transformation. *Computational Statistics and Data Anaysis*, **33**: 189-200.

Othman, A. R., Keselman, H. J., Padmanabhan, A. R., Wilcox, R. R., Algina, J., and Fradette, K. 2004. Comparing measures of the "typical" score across treatment groups. *British Journal of Mathematical and Statistical Psychology*: **57**(2): 215-234.

SAS Institute Inc. 2004. *SAS OnlineDoc® 9.1.2.* Cary, NC: SAS Institute Inc.

Syed Yahaya, S. S., Othman, A. R., and Keselman, H. J. 2004. Testing the equality of location parameters for skewed distributions using S1 with high breakdown robust scale estimators. In M. Huber, G. Pison, A. Struyf & S. V. Aelst (Eds.), *Theory and Applications of Recent Robust Methods* (pp. 319-328). Series: Statistics for Industry and Technology, Birkhauser, Basel.

Syed Yahaya, S. S., Othman, A. R., and Keselman, H. J. 2006. Comparing the "typical score" across independent groups based on different criteria for trimming. *Metodološki Zvezki-Advances in Methodology and Statistics*. **3**(1): 49-62.

Teh, S. Y., and Othman, A. R. 2009. When does the pooled variance *t*-test fail? *African Journal of Mathematics and Computer Science Research.* **2**(4): 056-062.

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

Wilcox, R. R. 1992. Why can methods for comparing means have relatively low power, and what can you do to correct the problem? *Psychological Science*, **1**(3): 104 – 105.

Wilcox, R. R., and Keselman, H. J. 2003. Repeated measures one-way ANOVA based on a modified one-step M-estimator. *British Journal of Mathematical and Statistical Psychology*, **56**: 15-25.