Improvement of the trimmed mean procedure using bootstrap method

[1]Zahayu Md Yusof, [2]Abdul Rahman Othman, [3]Sharipah Soaad Syed Yahaya

[1,3] UUM College of Arts and Sciences, Science Quantitative Building,
Universiti Utara Malaysia
06010 UUM Sintok, Kedah, Malaysia
[1]School of Distance Education, Universiti Science Malaysia
11800 USM Pulau Pinang, Malaysia
E-mail: [1]zahayu@uum.edu.my, [2]oarahman@usm.my,[3]sharipah@uum.edu.my

Abstract: When the assumptions of normality and homoscedasticity are met, researchers should have no doubt in using classical test such as *t*-test and *ANOVA*, to test for the equality of central tendency measures for two and more than two groups, respectively. However, in real life this perfect situation is rarely encountered. When the problem of nonnormality and variance heterogeneity simultaneously arise, rates of Type I error are usually inflated resulting in spurious rejection of null hypotheses. In addition, the classical least squares estimators can be highly inefficient when assumptions of normality are not fulfilled. Thus, by substituting robust measures of location and scale such as trimmed means and Winsorized variances in place of the usual means and variances respectively, tests that are insensitive to the combined effects of nonnormality and variance heterogeneity can be obtained. In this study, we compared the performance of T_l statistic using bootstrap methods with the approximate trimmed F statistic (F_t). Both statistics used 15% symmetric trimming. The procedures examined generally resulted in good Type I error controlled. The F_t statistic shown good controlled of Type I error for balanced design. In contrast the T_l statistic gave better controlled of Type I error for the unbalanced design.

Keywords: trimming, Type I error, bootstrap

1. INTRODUCTION

In recent years, numerous methods were being studied in terms of finding better methods for controlling the rates of Type I error in the one-way independent group designs (Babu, Padmanabhan & Puri, 1999; Othman, Keselman, Padmanabhan, Wilcox & Fradette, 2004; Wilcox & Keselman, 2003). Through a combination of theoretical developments, more flexible statistical methods, and faster computers, serious practical problems that seemed insurmountable only a few years ago can now be addressed. One way to overcome the problems with controlling Type I error rates is by using robust statistics.

There are varieties of definitions for robust statistics that have been found in the literature and these unfortunately lead to the inconsistency of its meaning. Most of the definitions are based on the objective of the particular study by different researchers (Huber, 1981). The robust method is in fact an alternative to a classical method with the aim of producing estimators which cannot be influenced by the deviations from the given assumptions when hypothesis testing is being conducted.

When the homogeneity of variance assumption is not satisfied, particularly when the design is unbalanced, the classical test of mean equality can become seriously biased. The usual group means and variances are greatly influenced by the presence of outliers in the score distribution. Reduction in the power to detect differences between groups occurs because of the standard error for the usual mean can become seriously inflated when the underlying distribution is heavy tailed (Lix & Keselman, 1998).

When non-normality exist, classical least squares estimators could be highly inefficient. By substituting robust measures of location and scale such as trimmed means and Winsorized variances in place of the usual means and variances respectively, tests that were insensitive to the combined effects of non-normality and variance heterogeneity could be obtained (Lix & Keselman, 1998). In terms of power, Wilcox, Keselman and Kowalchuk (1998) stated that one was able to obtain test statistics that did not suffer losses in power due to non-normality by using trimmed means and variances based on Winsorized sum of squares.

Trimmed mean is a good measure of location because the standard error of the trimmed mean is less affected by departures from normality. This is due to the fact that the extreme values or outliers are removed (Lix & Keselman, 1998). According to Gross (1976), Winsorized variance is a consistent estimator of the variance of the corresponding trimmed mean. The trimmed mean and Winsorized variance are intuitively appealing because of their computational simplicity and good theoretical properties (Wilcox, 1995).

2. METHODS

This paper focuses on the T_1 and F_1 statistics with 15% symmetric trimming. These two methods were compared in terms of Type I error under conditions of normality and non-normality which will be represented by skewed *g*- and *h*- distributions.

2.1. T_1 method

When the distributions are symmetric, Babu, *et al.* (1999) recommended the use of T_1 statistic to compare differences between distributions. They used a refined version of calculating trimmed means (Rocke, Downs & Rocke, 1982).

Let $X_{(1)j} \leq X_{(2)j} \leq \ldots\ldots\ldots \leq X_{(n_j)j}$ represent the ordered observations associated with the j^{th} group.

In order to calculate the $100g\%$ sample trimmed mean, define

$$X_{Lj} = (1-r)X_{(k+1)j} + rX_{(k)j}$$

and

$$X_{Uj} = (1-r)X_{(n_j-k)j} + rX_{(n_j-k+1)j}$$

where

g represents the proportion of observations that are to be trimmed in each tail of the distribution.

$k = \lfloor gn_j \rfloor + 1$ which $\lfloor gn_j \rfloor$ is the largest integer $\leq gn_j$ and $r = k - gn_j$.

The trimmed mean is given by

$$\overline{X}_{tj} = \frac{1}{(1-2g)n_j}\left[\sum_{i=k+1}^{n_j-k} X_{(i)j} + r(X_{(k)j} + X_{(n_j-k+1)j})\right].$$

Its corresponding sample Winsorized mean is given by

$$\overline{X}_{wj} = \frac{1}{n_j}\left[\sum_{i=k+1}^{n_j-k} X_{(i)j} + k(X_{Lj} + X_{Uj})\right]$$

The squared sample Winsorized standard error is as follows:

$$\hat{v}_{tj} = \frac{1}{(1-2g)n_j(n_j-2n_jg-1)} \times$$

$$\left[\sum_{i=k+1}^{n_j-k}\left(X_{(i)j} - \overline{X}_{wj}\right)^2 + k\left(\left(X_{Lj} - \overline{X}_{wj}\right)^2 + \left(X_{Uj} - \overline{X}_{wj}\right)^2\right)\right].$$

Then the T_1 statistic is given by

$$T_1 = \sum_{1\leq j\leq j'\leq J} |t_{jj'}|,$$

where

$$t_{jj'} = \frac{\left(\overline{X}_{tj} - \overline{X}_{tj'}\right)}{\sqrt{\hat{v}_{tj} + \hat{v}_{tj'}}}$$

2.2. F_t Method

Lee and Fung (1985) introduced a statistic that was able to handle problems with sample locations when the variance for the population is equal. This statistic was named the trimmed F statistic, F_t. They also suggested this new statistic to be used for problem involving one-way ANOVA and they recommended this to be an alternative to the usual F method. This method had also been proven to be easy to program.

Let $X_{(1)j}, X_{(2)j}, \ldots, X_{(n_j)j}$ be an ordered sample of group j with size n_j. The g-Winsorized sum of squared deviations is then calculated as

$$SSD_{wj} = \sum_{i=k_j+1}^{n_j-k_j} \left(X_{ij} - \overline{X}_{wgj}\right)^2 + k_j\left[\left(X_{k_j+1,j} - \overline{X}_{wgj}\right)^2 + \left(X_{n_j-k_j,j} - \overline{X}_{wgj}\right)^2\right]$$

Hence the g-trimmed F is defined as

$$F_t(g) = \frac{\sum_{j=1}^{J}\left(\overline{X}_{tj} - \overline{X}_t\right)^2/(J-1)}{\sum_{j=1}^{J} SSD_{tj}/(H-J)},$$

where

J = number of groups,

$$h_j = n_j - g_{1j} - g_{2j}, \quad H = \sum_{j=1}^{J} h_j \quad \text{and} \quad \overline{X}_t = \sum_{j=1}^{J} h_j \overline{X}_{tj} / H.$$

2.3 Percentile Bootstrap and Approximation Method

Due to the intractability of the T_1 distribution, percentile bootstrap method was used to conduct the hypothesis test on the T_1 procedure while approximation method was considered for the F_t statistic. Babu *et al.* (1999) obtained the p-values for the S_1 and T_1 statistics by means of the percentile bootstrap method. They also discovered that the percentile bootstrap method produced better approximation than the one based on the normal approximation theory, and furthermore, this method works well especially when the samples are of moderate size.

While for the F_t statistic, Lee and Fung (1985) used the normal approximation theory to obtain the p-values. They reported that, when the sample in each group is of size 10 or more, the approximation method seems satisfactory.

3. Empirical Investigation

This paper focused on a balanced and unbalanced completely randomized design containing four groups with small samples. For unbalanced designs, unequal variances of 1:36 ratio will be considered. Variances and group sizes are both positively and negatively paired. For positive pairings, the largest n_j will be paired with the largest group variance and the smallest n_j will be paired with the smallest group variance, whereas for the negative pairings, the smallest n_j will be paired with the largest group variance and the largest n_j will be paired with the smallest group variance. The conditions were chosen since they typically produce conservative results for the positive pairings and liberal results for the negative pairings (Othman, *et al.*, 2004).

We have chosen two total sample sizes, namely $N = 60$ and $N = 80$. For balanced design, we set the samples at $n_1 = 15$, $n_2 = 15$, $n_3 = 15$ and $n_4 = 15$ for $N = 60$ and for $N = 80$, we set the samples at $n_1 = 20$, $n_2 = 20$, $n_3 = 20$ and $n_4 = 20$. As for the unbalanced design, when $N = 60$, we set the samples at $n_1 = 12$, $n_2 = 14$, $n_3 = 16$ and $n_4 = 18$ and for $N = 80$, we set the samples at $n_1 = 10$, $n_2 = 20$, $n_3 = 20$ and $n_4 = 30$. For both total sample sizes

under unbalanced design, we used heterogenous variances at 1, 1, 1 and 36 respectively for positive pairings and 36, 1, 1 and 1 respectively for negative pairings.

Table 1: Design Specifications for the Four Groups (balanced designs).

N	Group sizes				Group variances			
	1	2	3	4	1	2	3	4
60	15	15	15	15	1	1	1	1
80	20	20	20	20	1	1	1	1

Table 2: Design Specifications for the Four Groups (unbalanced designs).

Pairing	Group sizes				Group variances			
	(N = 60)							
	1	2	3	4	1	2	3	4
Positive	12	14	16	18	1	1	1	36
Negative	12	14	16	18	36	1	1	1
	(N = 80)							
	Group sizes				Group variances			
	1	2	3	4	1	2	3	4
Positive	10	20	20	30	1	1	1	36
Negative	10	20	20	30	36	1	1	1

Each method will be tested under three types of distributions which are g = 0.0 and h = 0.0 (normal), g = 0.5 and h = 0.0 (skewed normal tailed) and g = 0.5 and h = 0.5 (skewed leptokurtic). The g- and h- distributions were first proposed by Hoaglin (1985). These distributions are transformations of the standard normal distribution. By manipulating the g- parameter one can transform the standard normal distribution into a skewed distribution. In addition to this one can also transform the standard normal distribution into a heavy tailed distribution by changing the h- parameter. For each of the designs, 5000 datasets were simulated and for T_1 statistic, 599 bootstrap samples were generated. The random samples were drawn using SAS generator RANNOR (SAS Institute, 1999).

4. Results and Conclusion

The results for Type I error for the methods investigated were shown in Table 3 and Table 4. Based on Bradley's liberal criterion of robustness (Bradley, 1978), a test can be considered robust if rate of Type I error, is within the interval 0.5α and 1.5α. For the nominal level α = 0.05, the Type I error rate should be between 0.025 and 0.075.

Table 3 and Table 4 display the empirical Type 1 error rates for all the procedures across the three distributions under balanced and unbalanced designs. Values that fall within the Bradley's liberal criterion of robustness were highlighted, and the average values that satisfy the criterion were underlined.

Table 3: Type I Error Rates for Balanced Designs.

	N = 60 (15, 15, 15, 15)		N = 80 (20, 20, 20, 20)	
Distributions	T_1	F_t	T_1	F_t
g = 0.0 and h = 0.0	**0.0404**	**0.0476**	**0.0452**	**0.0488**
g = 0.5 and h = 0.0	**0.0366**	**0.0446**	**0.0402**	**0.0478**
g = 0.5 and h = 0.5	**0.0294**	**0.0350**	**0.0318**	**0.0384**
Average	0.0355	0.0424	0.0391	0.0450

Table 4: Type I Error Rates for Unbalanced Designs.

		$N = 60$ (12, 14, 16, 18)		$N = 80$ (10, 20, 20, 30)	
Distributions	Pairing	T_l	F_l	T_l	F_l
$g = 0.0$ and $h = 0.0$	Positive	**0.0438**	**0.0696**	**0.0504**	**0.0356**
	Negative	**0.0428**	0.1434	**0.0464**	0.2692
Average		0.0433	0.1065	0.0484	0.1524
$g = 0.5$ and $h = 0.0$	Positive	**0.0406**	**0.0688**	**0.0470**	**0.0358**
	Negative	**0.0454**	0.1408	**0.0506**	0.2720
Average		0.0430	0.1048	0.0488	0.1539
$g = 0.5$ and $h = 0.5$	Positive	**0.0376**	**0.0526**	**0.0424**	**0.0292**
	Negative	**0.0392**	0.1172	**0.0410**	0.2594
Average		**0.0384**	0.0849	**0.0417**	0.1443
Grand Average		0.0416	0.0987	0.0463	0.1502

The procedures examined generally resulted in good Type I error control. Approximate trimmed F statistic, F_l shown good controlled of Type I error for balanced designs compared to T_l statistic. However, for the unbalanced designs, the T_l statistic gave better controlled of Type I error compared to trimmed F statistic, F_l. This study has shown that, approximate method gave better results in controlling the Type I error rates for balanced designs and as for unbalanced designs, the bootstrap methods is the best solution in controlling the Type I error rates. However, there are improvements in rates of Type I error when we increase the sample size from 60 to 80.

REFERENCES

Babu, J. G., Padmanabhan, A. R., & Puri, M. P. (1999). Robust one-way ANOVA under possibly non-regular conditions. *Biometrical Journal,* 41(3), 321 – 339.

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

Gross, A. M. (1976). Confidence interval robustness with long-tailed symmetric distributions. *Journal of the American Statistical Association,* 71, 409 – 416.

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

Huber, P. J. (1981). *Robust statistics*. New York: Wiley.

Lee, H and Fung, K.Y. (1985). Behaviour of trimmed F and sine-wave F statistics in one-way ANOVA. *Sankhya: The Indian Journal of Statistics,* 47, 186-201.

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.

Othman, A. R., Keselman, H. J., Padmanabhan, A. R., Wilcox, R. R., & Fradette, K. (2004). Comparing measures of the 'typical' score across treatment groups. *British Journal of Mathematical and Statistical Psychology,* 215 – 234.

Rocke, D. M., Downs, G. W., & Rocke, A. J. (1982). Are robust estimators really necessary?.*Technometrics,* Vol. 24, No. 2, 95 – 101.

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

Wilcox, R. R. (1995). ANOVA: A paradigm for low power and misleading measures of effect size? *Review Of Educational Research,* 65(1), 51 – 77.

Wilcox, R. R., Keselman, H. J., & Kowalchuk, R. K. (1998). Can tests for treatment group equality be improved?: The bootstrap and trimmed means conjecture. *British Journal of Mathematical and Statistical Psychology,* 123 – 134.

Wilcox, R. R., & Keselman, H. J. (2003). Modern robust data analysis methods: Measures of central tendency. *Psychological Methods,* **8**, 254-274.