Malaysian Journal of Science **32** (1): 73-77 (2013)

ROBUSTNESS OF TRIMMED F STATISTIC WHEN HANDLING NON-NORMAL DATA

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

[1,3] School of Science Quantitative, UUM College of Arts and Sciences, Universiti Utara Malaysia, 06010 UUM Sintok, Kedah.
[2] Pusat Pengajian Pendidikan Jarak Jauh, Universiti Sains Malaysia, 11800 USM Penang, Pulau Pinang.

ABSTRACT When the assumptions of normality and homoscedasticity are met, researchers should have no doubt in using classical test such as *t*-test, to test for the equality of central tendency measures for two groups. 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. The effect of non-normality on the trimmed *F* statistic was demonstrated in this study. We propose the modifications of the trimmed *F* statistic mentioned by using (1) a priori determined 15% symmetric trimming and (2) empirically determined trimming using robust scale estimators such as MAD_n, T_n and LMS_n. The later trimming method will trim extreme values without prior trimming percentage. Based on the rates of Type I error, the procedures were then compared. Data from *g*- and *h*- distributions were considered in this study. We found the trimmed *F* statistic using robust scale estimator LMS_n as trimming criterion provided good control of Type I error compared to the other methods.

ABSTRAK Apabila andaian normal dan homokedastik dipenuhi, penyelidik tidak perlu ragu untuk menggunakan ujian klasik seperti ujian-*t* bagi menguji kesamaan sukatan kecenderungan memusat untuk dua kumpulan. Walau bagaimanapun, dalam kehidupan sebenar situasi yang sempurna ini jarang dijumpai. Apabila masalah ketaknormalan dan varians heterogen berlaku serentak, ini akan memberi kesan kepada kadar ralat Jenis I dan seterusnya menyebabkan berlakunya penolakan terhadap hipotesis nol. Di samping itu, penganggar kuasa dua terkecil boleh menjadi sangat tidak cekap apabila andaian kenormalan tidak dipenuhi. Kesan ketidaknormalan pada statistik *F* terpangkas telah dibuktikan dalam kajian ini. Kami mencadangkan pengubahsuaian statistik *F* terpangkas menggunakan (1) penentuan awal 15% pemangkasan secara simetri dan (2) pemangkasan secara empirikal menggunakan penganggar skala teguh seperti MAD_n, T_n dan LMS_n. Kaedah pemangkasan yang terkemudian, akan memangkas nilai ekstrem tanpa penentuan awal peratusan pemangkasan. Berdasarkan kadar ralat Jenis I, prosedur-prosedur ini dibandingkan. Data dari taburan *g*- dan *h*- dipertimbangkan dalam kajian ini. Kami mendapati statistik *F* terpangkas menggunakan penganggar skala kukuh LMS_n sebagai kriteria pemangkasan mempunyai kawalan ralat Jenis I yang baik berbanding dengan kaedah lain.

(Keywords: Trimming criterion, robust scale estimators, Type I error)

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 [1, 2, and 3]. 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 were several definitions of robust statistics readily found in the literature and these unfortunately led to the inconsistency of its meaning. Most of the definitions were based on the objective of the particular study by different researchers [4]. When one is searching for a procedure which cannot be influenced by the deviations from assumptions while conducting hypothesis testing, a robust statistics based method provided an alternative to the classical method. [4] gave a definition for robustness as a situation which is not sensitive to small changes in assumptions. While [5] in his study reported that a

robust procedure is a procedure that was affected only slightly by appreciable departures from assumptions. Regardless of the definition provided, robust method in general offers substantial improvement over classical method [6 and 7]. Robust statistics have been used in statistical problems for the past 40 years. However, specific robust statistics based research on one-way ANOVA began two decades ago [8, 9 and 10].

In a non-normal model, 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 [11]. 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 [11]. According to [12], 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 [13].

Trimming can also be very beneficial in terms of efficiency and achieving high power. According to literature, the optimal amount of trimming is between 0 and 0.25. A good value would be 0.20 [3]. When using 20% trimming, we can expect more accurate probability coverage of confident interval regarding differences between means when distributions are skewed [14]. In [7], it is stated that the more we trim, the less effect skewness had on these probability coverage. However when n is small, the optimal amount of trimming is yet to be determined. While [15] in their paper concluded that the best results are obtained with 20% to 25% symmetric trimming. [2] found that one can achieve a slightly better Type I error control with a 15% symmetric trimming than with a 20% symmetric trimming. [16] demonstrated that a good control of Type I error can be achieved with only modest amounts of trimming, namely 15% or 10% from each tail of the distribution. To empirically determine the amount of trimming was difficult, and not always obvious.

METHODS

This paper focuses on the trimmed F statistic methods with 15% symmetric trimming and several trimming criteria using robust scale estimators MAD_n, T_n and LMS_n. These four 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.

Trimmed F statistic

Let $X_1, X_2,..., X_n$ be an ordered sample of size n and let $k = [gn]+1$ where $[x]$ is the largest integer $\leq x$. The g-trimmed mean of the sample is defined as,

$$\overline{X}_{tg} = \frac{\sum_{j=k}^{n-k+1} X_j}{(n-2gn)},$$

Prior to deriving the g-Winsorized sum of squared deviations, we defined g- Winsorized mean as,

$$\overline{X}_{wg} = \frac{\sum_{j=k+1}^{n-k} X_j + k(X_k + X_{n-k+1})}{n},$$

The g-Winsorized sum of squared deviations is then

$$SSD_{wg} = \sum_{j=k+1}^{n-k} (X_j - \overline{X}_{tg})^2 + k\left[(X_k - \overline{X}_{wg})^2 + (X_{n-k+1} - \overline{X}_{wg})^2\right]$$

And the g-trimmed F is defined as

$$F_t(g) = \frac{\sum_{i=1}^{c} h_i (\overline{X}_{itg} - \overline{X}_{tg})^2 / (c-1)}{\sum_{i=1}^{c} SSD_{iwg} / (H-c)}$$

where

c = number of groups.

$h_i = n_i(1-2g)$, $H = \sum_i h_i = N(1-2g)$, $(N = \sum_i n_i)$

$\overline{X}_{itg}$ = the g-trimmed mean of the i-th group $\overline{X}_{tg} = \sum_i h_i \overline{X}_{itg} / H$,

SSD_{iwg} = the g-Winsorized sum of squared deviation of the i-th group.

Trimming criterion

The trimmed mean was calculated by using:

$$\overline{X}_{(t)j} = \frac{1}{n_{j-i_1-i_2}}\left[\sum_{i=i_1+1}^{n_j-i_2} X_{(i)j}\right]$$

where

i_1 = number of observations X_{ij} such that $\left(X_{ij} - \hat{M}_j\right) < -2.24$ (scale estimator)

i_2 = number of observations X_{ij} such that $\left(X_{ij} - \hat{M}_j\right) > 2.24$ (scale estimator)

To arrive at the $F_t(g)$ statistic for these methods, the g-Winsorized mean is given by,

$$\overline{X}_{(w)j} = \frac{1}{n_{j-i_1-i_2}}\left[\sum_{i=i_1+1}^{n_j-i_2} X_{(i)j} + i_1 X_{(i_1+1)j} + i_2 X_{(n_j-i_2)j}\right]$$

The g-Winsorized sum of squared deviations is then

$$SSD_{(w)j} = \sum_{i=i_1+1}^{n_j-i_2} (X_{(i)j} - \overline{X}_{(w)j})^2 + i_1(X_{(i_1+1)j} - \overline{X}_{(w)j})^2 + i_2(X_{(n_j-i_2)j} - \overline{X}_{(w)j})^2$$

Robust scale estimators

Scale measure is a quantity that explains the dispersion of a distribution. The value of a breakdown point is a main factor to be considered when looking for a scale estimator [17]. [18] have introduced several scale estimators by considering their breakdown point.

MAD_n, T_n and LMS_n are three robust scale estimators used in this study. These estimators have 0.5 breakdown value and also have bounded influence functions. These estimators were chosen because of their simplicity and computational ease.

***MAD*$_n$**

MAD_n is the median absolute deviation about the median. It has the best possible breakdown value and its influence function is bounded with the sharpest possible bound among all scale estimators [18]. There are also some drawbacks about this scale estimator. The efficiency of MAD_n is very low with only 37% at Gaussian distribution. MAD_n takes a symmetric view on dispersion and also does not seem to be a natural approach for asymmetric distributions. This robust scale estimator is given by

$$MAD_n = b \text{ med}_i |x_i - \text{med}_j x_j|$$

where the constant b is needed to make the estimator consistent for the parameter of interest.

***T*$_n$**

Another scale estimator proposed by [18] is T_n, which has highest breakdown point like MAD_n. The scale estimator is given as

$$T_n = 1.3800 \frac{1}{h}\sum_{k=1}^{h} \{med_{j\neq i} | x_i - x_j |\}_{(k).}$$

where $h = \left[\frac{n}{2}\right] + 1$. T_n was proven to have 50% breakdown point and an efficiency of 52%. It is more efficient than MAD_n.

***LMS*$_n$**

LMS_n is also a scale estimator with a 50% breakdown point which is based on the length of the shortest half sample as shown below:

$$LMS_n = c' \min_i \left|x_{(i+h-1)} - x_{(i)}\right|$$

where $x_{(1)} \leq x_{(2)} \leq .... \leq x_{(n)}$ are the ordered data. The default value of *c'* is 0.7413 which achieves consistency at Gaussian distributions.

EMPIRICAL INVESTIGATION

This paper focused on a balanced completely randomized design containing two and four groups with small samples. We have chosen two population sizes, N = 30 and N = 40. For N = 30, the samples are set at n_1 = 15 and n_2 = 15 while for N = 40, the setting is n_1 = 20 and n_2 = 20. For both sizes we used homogeneous variances at 1:1. For four groups, 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. Each method was tested under three types of distributions with 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). For each of the designs, 5000 datasets were simulated. The random samples were drawn using SAS generator RANNOR [19].

Table 1 Design specifications for balanced design.

N	Group sizes		Group variances	
	1	2	1	2
30	15	15	1	1
40	20	20	1	1

Table 2 Design specifications for unbalanced design.

N	Group sizes		Group variances	
	1	2	1	2
30	12	18	1	1
40	15	25	1	1

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 [20], 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 I 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 displays the empirical Type I error rates for all the procedures across the three distributions. Values that fall within the Bradley's criterion were highlighted, and the average values that satisfy the criterion were underlined.

Across the distributions, all the values for 15% symmetric trimming are robust. However for extreme case, g = 0.5 and h = 0.5, trimming criterion using MAD_n and T_n works better than 15% symmetric trimming. On the average, trimmed F statistic with trimming criterion using robust scale estimator, LMS_n perform better in controlling Type I error rate as compared to all the other methods for smaller group size (N = 30). .While for larger group size (N = 40), 15% symmetric trimming seem to have better control of Type I error rates. From this finding, we would like to suggest using this method as the alternative to the traditional methods especially when the sample size is small. For extreme cases, trimmed F statistic with robust estimators MAD_n and T_n are recommended regardless of group size.

Table 3 Empirical Type I Error Rates (balanced design).

Distributions	Trimmed F statistic with robust scale estimator, N = 30 (15, 15)				Trimmed F statistic with robust scale estimator, N = 40 (20, 20)			
	MAD_n	T_n	LMS_n	$\hat{\nu}$ (15%)	MAD_n	T_n	LMS_n	$\hat{\nu}$ (15%)
g=0.0 h=0.0	0.0912	0.0886	**0.0628**	**0.0456**	0.0956	0.0858	**0.0614**	**0.0532**
g=0.5 h=0.0	0.1080	0.1050	**0.0428**	**0.0426**	0.1172	0.1164	**0.0472**	**0.0506**
g=0.5 h=0.5	**0.0462**	**0.0462**	0.0200	**0.0314**	**0.0472**	**0.0438**	0.0240	**0.0408**
Average	0.0818	0.0799	**0.0419**	**0.0399**	0.0867	0.0820	**0.0442**	**0.0482**

Table 4 Empirical Type I Error Rates (unbalanced design).

Distributions	Trimmed F statistic with robust scale estimator, N = 30 (12,18)				Trimmed F statistic with robust scale estimator, N = 40 (15,25)			
	MAD_n	T_n	LMS_n	$\hat{\nu}$ (15%)	MAD_n	T_n	LMS_n	$\hat{\nu}$ (15%)
g=0.0 h=0.0	0.0912	0.0810	0.1134	**0.0474**	0.0914	0.0834	0.1242	**0.0488**
g=0.5 h=0.0	0.1078	0.1060	0.1252	**0.0450**	0.1162	0.1124	0.1286	**0.0484**
g=0.5 h=0.5	**0.0518**	**0.0474**	**0.0510**	**0.0370**	**0.0492**	**0.0454**	**0.0520**	**0.0404**
Average	0.0836	0.0781	0.0965	**0.0431**	0.0856	0.0804	0.1016	**0.0459**

REFERENCES

1. 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.

2. 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.

3. Wilcox, R. R., & Keselman, H. J. (2003). Modern robust data analysis methods: Measures of central tendency. *Psychological Methods*, *8*, 254-274.
4. Huber, P. J. (1981). *Robust Statistics*. Wiley & Sons Inc.

5. Brownlee, K.A. (1965). *Statistical Theory and Methodology in Science and Engineering 2nd Ed*, John Wiley & Sons Inc.

6. Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust Statistics*. New York: Wiley.
7. Wilcox, R. R. (1998). The goals and strategies of robust methods. *British Journal of Mathematical and Statistical Psychology,* 51, 1-39.

8. Stigler, S. M. (1973). Simon Newcombe, Percy Daniell, and The History of Estimation 1885-1920. *Journal of the American Statistical Association,* 68, 872-879.

9. Ronchetti, E. M. (2006). The Historical Development of Robust Statistics. *Proceedings of the 7th International Conference on Teaching Statistics (ICOTS-7).*

10. Staudte, R. G., & Sheather, S. J. (1990). *Robust estimation and testing*. New York: Wiley.

11. 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*, Vol. 58, No. 3, 409 – 429. Mann, P. S. (2004). Introductory Statistics. Wiley & Sons Inc.

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

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

14. Wilcox, R. R. (1996). A note on testing hypotheses about trimmed means. *Biometrical Journal*, 38, 173-180.

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

16. H.J.Keselman, A.R.Othman, R.R.Wilcox and K.Fradette (2004). The New and Improved Two-Sample t Test. *American Psychological Society*, **15**, 47 - 51.

17. Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing (2nd ed)*. San Diego, CA: Academic Press.

18. Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88, 1273 – 1283.

19. SAS Institute Inc. (1999). SAS/IML User's Guide Version 8. Cary, NC: New York.

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