Malaysian Journal of Mathematical Sciences 4 (2): 195-207 (2010)

Comparison of Type I Error Rates Between T_1 and F_t Statistics for Unequal Population Variance Using Variable Trimming

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

[1,3]UUM College of Arts and Sciences, Bangunan Sains Kuantitatif, Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia

[2]Pusat Pengajian Pendidikan Jarak Jauh, Universiti Sains Malaysia, 11800 USM Pulau Pinang, Malaysia

E-mail: zahayu@uum.edu.my

ABSTRACT

Two robust procedures for testing the equality of central tendency measures, namely T_1 and trimmed F (noted as F_t) statistics are proposed in this paper. The T_1 and F_t statistics were modified using variable trimming with indeterminate percentage. The variable trimming percentages were based upon trimming criteria using robust scale estimators, MAD_n and T_n. Altogether there are four procedures investigated: T_1 with MAD_n, T_1 with T_n, F_t with MAD_n, and F_t with T_n. Concentrating on just balanced design and unequal population variances, the four procedures were tested for their Type I error under different types of distributional shapes and total sample sizes. This study used 5000 simulated data sets to generate the Type I error. Since T_1 distribution is unknown, bootstrap method was employed to test the hypothesis. The findings showed that T_1 statistic works well under normal tail distribution, while F_t statistic is good for extremely skewed distribution.

Keywords: variable trimming, robust scale estimators, extreme distribution.

INTRODUCTION

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.

A statistical method is considered robust if the inferences are not seriously invalidated by the violation of such assumptions, for instance non normality and variance heterogeneity (Scheffe, (1959)). Huber, (1981) defined robustness as a situation which is not sensitive to small changes in

assumptions while Brownlee, (1965) reported slight effects on a procedure when appreciable departures from the assumptions were observed.

The theory of robust statistics deals with deviations from the assumptions on the model and is concerned with the construction of statistical procedures which is still reliable and reasonably efficient in a neighborhood of the model (Ronchetti, (2006)). Hampel *et al.,* (1986), stated that in a broad informal sense, robust statistics is a body of knowledge, partly formalized into "theories of robustness" relating to deviations from idealized assumptions in statistics. As mentioned by Hoel *et al.,* (1971), a test that is reliable under rather strong modifications of the assumptions on which it was based is said to be robust.

Robust statistics has widely been used for many years now (Stigler, (1973)). Ronchetti, (2006) reported that research in robust statistics problems have been conducted since 40 years ago. However, there is no specific research on the robust statistics problems until the recent years (Staudte and Sheather, (1990)). Research about robust statistics is still active. In Ronchetti's , (2006) quick search in the Current Index of Statistics, 1617 papers on robust statistics between 1987 and 2001 in statistics journals and related fields were listed.

To date, there are several new procedures that were developed to deal with central tendency measures such as group trimmed means, group median or group *M*-measures of location. Among the latest procedures are the modified *MOM-H* statistic introduced by Syed Yahaya *et al.,* (2004) and a robust test due to Lee and Fung, (1985) based on a priori determined symmetric or asymmetric trimming strategies introduced by Keselman *et al.,* (2007). These methods used trimmed means as the central tendency measures and were proven to have good control of Type I error rates when comparing for the differences between distributions. In this study, we compare the performance of T_1 statistic developed by Babu *et al.,* (1999) and F_t statistic introduced by Lee and Fung, (1985) in asymmetric variable trimming. The trimming criterion is based on three robust scale estimators: MAD_n and T_n (Rousseeuw and Croux, (1993)). Unlike trimmed means, when using the aforementioned trimming criterion, no predetermined trimming percentage is needed.

TRIMMING

Trimmed mean is a central tendency measure to summarize data when trimming is carried out. By using the trimmed means, the effect of the

tails of the distribution is reduced by removing the extreme observations based on the predetermined amount. The common trimmed mean used the predetermined method for trimming. By using this method, amounts such as 10% or 20% of the observations from a distribution will be trimmed from both tails. In the case of a light-tailed distribution or the normal distribution, it may be desirable to trim a few observations or none at all. There is extensive literature regarding the trimming method that uses the predetermined amount of symmetric trimming. Among them are Lee and Fung, (1985), Keselman *et al.,* (2002) and Wilcox, (2003).

If we have skewed distributions then the amounts of trimming on both tails should be different, namely more should be trimmed from the skewed tail. However, if the predetermined symmetric trimming is used, regardless of the shape of the tails, the trimming is done symmetrically as set. A recent research by Keselman *et al.,* (2007) used asymmetric trimming and in particular, applying hinge estimators proposed by Reed and Stark, (1996) to determine the suitable amount of trimming on each tail of a distribution. However, their method still used predetermined trimming percentages.

The trimmed mean is not so robust because the breakdown point of trimmed mean is just as much as the percentage of trimming and this shows that trimmed mean cannot withstand large numbers of extreme value. Wilcox *et al.,* (2000) in their study stated that when comparing trimmed means versus means with actual data, the power of the trimmed mean procedure was observed to be greatly increased. They also discovered that there was improved control over the probability of a Type I error.

The question that always remains unanswered is "How can we determine the best percentage of trimming that would ensure good Type I error control and reasonable power?" A probable answer lies in trimming carried out for the calculation of modified one-step *M* - estimators (*MOMs*). Here trimming is based upon a trimming criterion that relies upon a robust scale estimator known as MAD_n (Wilcox and Keselman, (2002)). With this method of trimming we do not have to predetermine the amount of trimming required. The criterion will identify how many extreme values need to be removed from the distribution.

METHODS

This paper focuses on the T_1 and F_t statistics with variable trimming using several robust scale estimators as trimming criteria, namely MAD_n and T_n. These two statistics (T_1 and F_t) were compared in terms of Type I error under conditions of normality and non-normality which will be represented by the *g*- and *h*- distributions.

T_1 statistic

Let $X_{(1)j}, X_{(2)j}, ..., X_{(n_j)j}$ be an ordered sample of group *j* with size n_j. First, calculate the *g*-trimmed mean of group *j* by using:

$$\overline{X}_{tj} = \frac{1}{n_j - g_{1j} - g_{2j}} \left[\sum_{i=g_{1j}+1}^{n_j - g_{2j}} X_{(i)j} \right] \qquad \text{where}$$

g_{1j} = number of observations $X_{(i)j}$ such that

$$\left(X_{(i)j} - \hat{M}_j \right) < \text{ -2.24 (scale estimator)},$$

g_{2j} = number of observations $X_{(i)j}$ such that

$$\left(X_{(i)j} - \hat{M}_j \right) > \text{ 2.24 (scale estimator)},$$

$\hat{M}_j$ = median of group *j* and the scale estimator in the parentheses can be MAD_n or T_n.

Then, compute the sample Winsorized standard error. The squared sample Winsorized standard error is defined as

$$\hat{v}_{tj}^{\,2} = \frac{1}{(n_j - g_{1j} - g_{2j})(n_j - g_{1j} - g_{2j} - 1)} \times$$
$$\left[\sum_{i=g_{1j}+1}^{n_j - g_{2j}} \left(X_{(i)j} - \overline{X}_{tj} \right)^2 + g_{1j}(X_{(g_{1j}+1)j} - \overline{X}_{tj})^2 + g_{2j}(X_{(n_j - g_{2j})j} - \overline{X}_{tj})^2 \right]$$

The T_1 statistic (Babu *et al.,* 1999) is given by

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

where

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

T_1 is the sum of all possible differences of sample trimmed means from J distributions divided by their respective sample Winsorized standard errors. With J distributions, the number of $t_{jj'}$s is equal to J (J-1)/2. Note that trimmed means are used in the Winsorized standard errors formula instead of Winsorized means.

F_t statistic

Let $X_{(1)j}, X_{(2)j}, ..., X_{(n_j)j}$ be an ordered sample of group j with size n_j. The g-trimmed mean of group j is calculated by using the same formula as in the T_1 statistic i.e.:

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

The Winsorized sum of squared deviations for group j is then defined as,

$$\begin{aligned} SSD_{tj} = {} & \left(g_{1j}+1\right)\left(X_{\left(g_{1j}+1\right)j} - \overline{X}_{tj}\right)^2 + \left(X_{\left(g_{1j}+2\right)j} - \overline{X}_{tj}\right)^2 + \ldots \\ & + \left(X_{\left(n_j-g_{2j}-1\right)j} - \overline{X}_{tj}\right)^2 + \left(g_{2j}+1\right)\left(X_{\left(n_j-g_{2j}\right)j} - \overline{X}_{tj}\right)^2 \\ & - \left\{\left(g_{1j}\right)\left[X_{\left(g_{1j}+1\right)j} - \overline{X}_{tj}\right] + \left(g_{2j}\right)\left[X_{\left(n_j-g_{2j}\right)j} - \overline{X}_{tj}\right]\right\}^2 / n_j \end{aligned}$$

Note that we applied the trimmed means in SSD_{tj} formula instead of the Winsorized means.

Hence the trimmed F statistic (Lee and Fung, (1985)) is defined as

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

where

J = number of groups, $h_j = n_j - g_{1j} - g_{2j}$, $H = \sum_{j=1}^{J} h_j$ and $\overline{X}_t = \sum_{j=1}^{J} h_j \overline{X}_{tj} / H$. $F_t(g)$ will follow approximately an F distribution with ($J - 1$, $H - J$) degrees of freedom.

Robust scale estimators

The value of a breakdown point is the main factor to be considered when looking for a scale estimator (Wilcox, (2005)). Rousseeuw and Croux, (1993) have introduced several scale estimators with highest breakdown point, such as MAD_n and T_n. Due to their good performance in Rouseeuw and Croux, (1993) and Syed Yahaya *et al.,* (2004), the aforementioned scale estimators were chosen for this study. These scale estimators have 0.5 breakdown value and also exhibit bounded influence functions. These estimators were also chosen due to their simplicity and computational ease.

i. ***MAD_n***

MAD_n is the median absolute deviation about the median. It demonstrates the best possible breakdown value of 50%, twice as much as the interquartile range and its influence function is bounded with the sharpest possible bound among all scale estimators (Rousseeuw and Croux, (1993)).

This robust scale estimator is given by

$$MAD_n = b \ \text{med}_i \left| x_i - med_j x_j \right|$$

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

However, this estimator is not free from drawbacks. The efficiency of MAD_n is very low with only 37% at Gaussian distribution. It also takes a symmetric view on dispersion and does not seem to be a natural approach for problems with asymmetric distributions.

ii. T_n

T_n for asymmetric distribution, Rousseeuw and Croux, (1993) proposed T_n, a scale known for its highest breakdown point like MAD_n. However, this estimator has more plus points compared to MAD_n. It has 52% efficiency, making it more efficient than MAD_n. It also has a continuous influence function.

Given as

$$T_n = 1.3800 \frac{1}{h} \sum_{k=1}^{h} \{med_{j \neq i} | x_i - x_j | \}_{(k).} \quad \text{where } h = \left[\frac{n}{2} \right] + 1$$

T_n has a simple and explicit formula that guarantees uniqueness. This estimator also has 50% breakdown point.

EMPIRICAL INVESTIGATION

The asymptotic sampling distributions for T_1 and F_t are known and have been derived in Babu *et al.,* (1999) and Lee and Fung, (1985), respectively. Knowing this we are still unable to determine how these two statistics will perform when the group sample sizes are small, let alone the modifications of these two statistics to accommodate automatic trimming. Hence, this paper focuses on the performance of these modified statistics on two groups with small but equal sample sizes. Two groups of size $N = 30$ and $N = 40$ are chosen. Therefore, when $N = 30$, the sample sizes are set at $n_1 = n_2 = 15$ and when $N = 40$, they are set at $n_1 = n_2 = 20$.

The next consideration is the heterogeneity of variances of the two groups. Starting from the commonly use rule of thumb that variances between two groups are heterogeneous if one group variance is four times the other, we can say that ratios of 1:9, 1:16, etc., ensured heterogeneity. However, these ratios have been routinely used in past studies. Thus, we chose 1:36 conveniently and at the same time ensured extreme heterogeneity. Even though the selected ratio seemed large, based on the

previous literature, higher ratio than 1:36 had been used by other researchers in their study (Keselman *et al*., (2007)). In addition, it will also provide researchers with information regarding how well the methods hold up under any degree of heterogeneity they are likely to obtain in their data, thus providing a very generalizable result (Keselman *et al*., (2007)). Table 1 shows the design specifications used in this study.

The Tukey's *g*- and *h*- distribution (Hoaglin, (1985)) is based on a transformation of a standard normal variable *Z* to $Y_z = \frac{e^{gz}-1}{g} e^{hz^2/2}$ where *g* controls the skewness and *h* effects the tail weights. When $g = 0$, the random variable Y_z is symmetric with increasingly heavy tails as *h* increases. We compare the performance of the T_1 and F_t statistics under three types of *g*- and *h*- distributions: (i) $g = 0.0$, $h = 0.0$ (normal), (ii) $g = 0.5$, $h = 0.0$ (skewed distribution) and (iii) $g = 0.5$, $h = 0.5$ (skewed leptokurtic). 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. When testing for the T_1 and F_t procedures, 5000 datasets were simulated and 599 bootstrap samples were generated for each of the designs. The random samples were drawn using SAS generator RANNOR (SAS institute, 1989).

To obtain the *p*-value of the T_1 statistic by using the percentile bootstrap method, the steps are as follows:

(a) Calculate T_1 based on the available data.

(b) Generate bootstrap samples by randomly sampling with replacement n_j observations from the j^{th} group yielding $X^*_{(1)j}, X^*_{(2)j}, ..., X^*_{(n_j)j}$.

(c) Each of the sample points in the bootstrapped groups must be centered at their respective estimated trimmed means so that the sample trimmed mean is zero, such that $C^*_{ij} = X^*_{ij} - \overline{X}_{tj}$, $i = 1, 2, ..., n_j$. The empirical distributions are shifted so that the null hypothesis of the equal trimmed means among the *J* distributions is true. The strategy behind the bootstrap is to use the shifted empirical distributions to estimate an appropriate critical value.

(d) Let T_1^* be the value of T_1 test based on the C_{ij}^* values.

(e) Repeat Step (a) to Step (d) B times yielding $T_{(1)1}^*, T_{(1)2}^*, ..., T_{(1)B}^*$. B = 599 appears sufficient in most situations when $n_j \geq 12$ (Wilcox, 2005).

(f) Calculate the p-value as number of $\frac{T_{1B}^* > T_1}{B}$.

The calculated p-values are the estimated rates of Type I error for the procedures investigated under the T_1 statistic. Hall, (1986) also stated that it is advantageous to choose B such that the nominal level, α , is a multiple of $(B+1)^{-1}$. Efron and Tibshirani, (1993) suggested that B should be at least 500 or 1000 in order to make the variability of the estimated percentile acceptably low. In this study, B is set to be 599 with the reason that 599 is the lowest value that can make α a multiple of $(B+1)^{-1}$ based on suggestion by Efron and Tibshirani, (1993). Furthermore, trials on various numbers of bootstraps from B = 599 to 999 with the increment of 100 found that the p-values for different number of bootstraps are consistent. Thus, to save the running time, this study chose for the smallest B in the range.

The steps to obtain the p-value for the procedures under F_t statistic are enumerated below.

(a) Based on the available data, calculate the F_t statistic.
(b) Calculate the degree of freedom for the available data.
(c) Determine the p-value of the calculated F_t statistic for two groups cases.

The calculated p-value represents the estimated rate of Type I error for the procedures investigated under the F_t statistic.

TABLE 1: Design Specifications for the Two Groups.

N	***GROUP SIZES***		***GROUP VARIANCES***	
	1	2	1	2
30	15	15	1	36
40	20	20	1	36

RESULTS AND CONCLUSION

Table 2 displays the empirical Type I error rates for all the procedures across the three distributions. Based on Bradley's liberal criterion of robustness (Bradley, (1978)), a test can be considered robust if the rate of Type I error, $\hat{\alpha}$ is within the interval 0.5α and 1.5α. For the nominal level of $\alpha = 0.05$, the Type I error rates should be between 0.025 and 0.075. Values that fall within the Bradley's criterion were highlighted, and the average values that satisfy the criterion were also underlined.

TABLE 2: Type I error rates

	$N = 30$				$N = 40$			
	$N = 30(15, 15)$ Variance = (1:36)				$N = 40(20, 20)$ Variances = (1:36)			
Distribution	T_1 with scale estimator		F_t with scale estimator		T_1 with scale estimator		F_t with scale estimator	
	MAD_n	T_n	MAD_n	T_n	MAD_n	T_n	MAD_n	T_n
$g = 0.0$, $h = 0.0$	**0.0248**	**0.0272**	0.1152	0.1106	**0.0298**	**0.0320**	0.1112	0.1012
$g = 0.5$, $h = 0.0$	**0.0332**	**0.0326**	0.1460	0.1392	**0.0414**	**0.0448**	0.1320	0.1270
$g = 0.5$, $h = 0.5$	0.0164	0.0144	**0.0650**	**0.0614**	0.0192	0.0214	**0.0622**	**0.0558**
Average	**0.0248**	0.0247	0.1087	0.1037	**0.0301**	**0.0327**	0.1018	0.0947

According to the table, all of the p-values for T_1 statistic with variable trimming, T_n and MAD_n fell within the Bradley's liberal criterion of robustness for both total sample sizes. T_1 also works well with the MAD_n and T_n trimming criterion as long as the tail of the distribution is normal. When the tail became heavier ($g = 0.5$ $h = 0.5$), the results for the two procedures (MAD_n and T_n) become more conservative.

The F_t statistic shows better performance in controlling Type I error rates when the distribution is extremely skewed. All of the p-values for this distribution fell within the Bradley's interval. However, the performance diminishes when the distributions have normal tail. Under this condition all the other p-values for F_t procedure become liberal. We also observed some progress in the result for both statistics (F_t and T_1) when total sample size increased.

Comparing the performance of the two statistics, the rates of Type I error for T_1 statistic is better than the F_t statistic for the non-extreme distributions i.e. $g = 0.0$, $h = 0.0$ and $g = 0.5$ and $h = 0.0$, and vice versa for the extremely skewed distribution ($g = 0.5$ and $h = 0.5$).

To avoid unnecessary trimming, we suggest the T_1 statistic with T_n when the distribution is non-extreme as this procedure provided the nearest Type I error rates to the nominal level. Conversely, investigation on the F_t statistic discovered that F_t works very well with T_n under extremely skewed distribution. These trimming criteria serve as alternatives to predetermined trimming methods, especially in handling problems with non-normality and variance heterogeneity.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the work that led to this paper is partially funded by the Fundamental Research Grant Scheme of the Ministry of Higher Education, Malaysia.

REFERENCES

Babu, J. G., Padmanabhan, A. R., and 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 Journal of Mathematical and Statistical Psychology*, **31**:144 - 152.

Brownlee, K.A. 1965. *Statistical theory and methodology in science and engineering (2nd Ed.)*. New York: Wiley.

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

Hall, P. 1986. On the number of bootstrap simulations required to construct a confidence interval. *Annals of Statistics*, **14**:1431 – 1452.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. 1986. *Robust statistics*. New York: Wiley.

Hoel, P. G., Port, S. C. and Stone, C. J. 1971. *Introduction to statistical theory*. Boston: Houghton Mifflin.

Hoaglin, D.C. 1985. Summarizing shape numerically: The *g*- and *h*-distributions. In D. Hoaglin, F. Mosteller, and J. Tukey, (Eds.). *Exploring Data Tables, Trends, and Shapes*. New York: Wiley. 461 - 513.

Huber, P. J. 1981. *Robust Statistics*. New York: Wiley.

Keselman, H. J., Wilcox, R. R., 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 Methods,* **1**: 288 – 309.

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

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.

Reed III, J. F. and Stark, D. B. 1996. Hinge estimators of location: Robust to asymmetry. *Computer Methods and Programs in Biomedicine*, **49**:11-17.

Ronchetti, E. M. 2006. The historical development of robust statistics. *In Proceedings of the 7th International Conference on Teaching Statistics (ICOTS-7).* 2 – 7 July, 2006, Salvador, Brazil. International Statistical Institute: Working Cooperatively in Statistics Education. [Retrieve 20 February 2008 from http://www.stat.aucland.ac.nz/ ~iase/ publications/17/3B1_RONC.pdf.].

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

SAS Institute. 1989. *IML software: Usage and reference, version 6*, (1st ed.). Cary, NC: SAS Institute.

Scheffe, H. 1959. *The Analysis of Variance*. New York: Wiley.

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

Stigler, S. M. 1973. Simon Newcombe, Percy Daniell, and the history of estimation 1885-1920. *Journal of the American Statistical Association,* **68**: 872-879.

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.Hubert, G.Pison, A. Struyf and S. Van Aelst (Eds.), Theory and Applications of Recent Robust Methods, Series: Statistics for Industry and Technology, Birkhauser, Basel.* 319 – 328.

Wilcox, R. R. 2003. Multiple comparisons based on a modified one-step *M*-estimator. *Journal of Applied Statistics*, **30**(10):1231-1241.

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

Wilcox, R. R., Keselman, H. J., Muska, J. and Cribbie, R. 2000. Repeated measures ANOVA: Some new results on comparing trimmed means and means. *British Journal of Mathematical and Statistical Psychology*, **53**: 69-82.

Wilcox, R. R. and Keselman, H. J. 2002. Power analysis when comparing trimmed means. *Journal of Modern Applied Statistical Methods*, **1**(1):24-31.