Far East Journal of Mathematical Sciences (FJMS)
Volume 69, Number 1, 2012, Pages 37-50
Published Online: August 2012
Available online at http://pphmj.com/journals/fjms.htm
Published by Pushpa Publishing House, Allahabad, INDIA

TYPE I ERROR AND POWER RATES OF F_t STATISTIC WITH TRIMMED MEAN

Zahayu Md Yusof, Suhaida Abdullah, Sharipah Soaad Syed Yahaya and Abdul Rahman Othman*

School of Quantitative Sciences
UUM College of Arts and Sciences
Universiti Utara Malaysia
06010 UUM Sintok, Kedah
Malaysia
e-mail: zahayu@uum.edu.my
suhaida@uum.edu.my
sharipah@uum.edu.my

*Pusat Pengajian Pendidikan Jarak Jauh
Universiti Sains Malaysia
11800 USM Penang
Pulau Pinang, Malaysia
e-mail: oarahman@usm.my

Abstract

Achieving nominal type I error rates and having high power values simultaneously will produce good test statistics. In order to identify a good test statistic which is able to satisfy both aforementioned criteria, a study is done on F_t statistic with trimming strategies using robust scale estimators, namely, MAD_n, T_n and LMS_n. To test for the

2010 Mathematics Subject Classification: 62G05, 62G35.
Keywords and phrases: type I error, power, robust scale estimators, robustness.
Received April 4, 2012

robustness of the procedures towards the violation of the assumptions, several variables are manipulated. The variables are types of distributions, heterogeneity of variances, sample sizes, nature of pairings of group sample sizes and group variances, and number of groups. This study is based on simulated data with each procedure simulated 5000 times. When testing for the hypothesis of the equality of central tendency measures, approximation method is used on F_t statistic. Type I error and power rates on $J = 4$ groups are then compared. Normal and skewed data from *g*- and *h*-distributions are considered in this study. Generally, all trimming strategies produce good type I error rates with high power values concurrently.

Introduction

Two sample *t*-test and analysis of variance (ANOVA) are two common statistical methods used to locate treatment effects in a one way independent group design. However, in using these two statistics, assumptions of normality and homogeneity of variance need to be fulfilled. In real life applications, these conditions are rarely achieved and any violation will lead to inaccuracy in decision based on the testing procedure. When these two problems simultaneously arise, rates of type I error are usually inflated resulting in spurious rejection of null hypotheses and the power of the test statistics is reduced.

The usual group means and variances, which are, respectively, the location and scale measures for the classical methods, are greatly influenced by the presence of outliers in the score distribution. The existence of outliers in a sample data will cause the probability of type I error to be less than the nominal level and concurrently lower the power of the test statistic. In the application of *t*-test, outliers can inflate the sample variance and simultaneously lower the value of the test (Wilcox and Keselman [26]). Even when sampling from a perfectly symmetrical distribution, outliers can still cause the *t*-test to loose power when compared against modern methods. Modern methods here are methods that are based on robust measures of location (Wilcox and Keselman [26]).

According to Keselman et al. [9], reduction in the power to detect differences between groups occurs because of the usual population standard deviation is greatly influenced by the presence of the extreme observations in a distribution of scores. Furthermore, the standard error for the usual mean can become seriously inflated when the underlying distribution is heavy tailed (Lix and Keselman [13]). In addition, the classical least squares estimators can be highly inefficient when assumptions of normality are not fulfilled. Hence, 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 non-normality and variance heterogeneity can be obtained (Lix and Keselman [13]). Wilcox et al. [25] stated that one is able to obtain test statistics that do not suffer losses in power due to non-normality by using trimmed means and variances based on Winsorized sum of squares.

The sample mean is the most common estimator used in most statistical analyses. However, this estimator is also very sensitive to the presence of outliers and skewness. Under these conditions, any test that used the sample mean as the estimator will produce low power and distorted rates of type I error. These include the *t*-test and ANOVA. To address this problem, Wilcox and Keselman [26] suggested using estimators of robust measures of location and rank-based methods. Some of these robust estimators are the *M* measure of location and trimmed mean.

The sample trimmed mean (will be referred to as "trimmed mean" throughout this article) is one of the estimators which is able to handle the problems of outliers and non-normal data. When using this estimator, the smallest and the largest observations in the distribution will be trimmed, this will automatically discard the outliers. By using trimmed mean, high power, accurate probability coverage, relatively low standard errors, a negligible amount of bias and a good control over the probability of a type I error can be achieved (Wilcox and Keselman [26]).

There are two possibilities of estimating the trimmed mean, i.e., equal amount of trimming or symmetric trimming and unequal amount of trimming or asymmetric trimming. In symmetric trimming, the trimming is done equally on both sides of the distribution. While for asymmetric trimming, the trimming is done on only one side or unequally on both sides of the distribution. Othman et al. [17] recommended that when the data are said to be skewed to the right, then in order to achieve robustness to non-normality and greater sensitivity to detect effects, one should trim data just from the upper tail of the data distribution. Hogg [7], Hertsgaard [5] and Tiku [22, 23] suggested that the data should have different amounts of trimming percentages from the right and left tails of the distribution. Keselman et al. [11] proposed a method called *adaptive robust estimators* to determine the number of observations to be trimmed from each tail of the distribution. By using this method, the total amount of trimming is determined a priori before making the decision whether to trim the data symmetrically, asymmetrically or not to trim at all (Keselman et al. [11]).

These two strategies stand on the trimming percentage that has to be stated in advance. It needs the fix amount of trimming percentage. The strategies are tight down with this amount of trimming. In our proposed method, the fixed trimming percentage problem can be avoided, since the trimming is done based on the shape of the distribution. Thus, we do not have to determine the total amount of data that need to be trimmed in advance because the determination of the total amount of trimming is done automatically.

Method

This paper focuses on the F_t method with new trimming strategies using robust scale estimators MAD_n, T_n and LMS_n and also with 15% symmetric trimming $(\hat{\nu})$. The F_t statistic with $\hat{\nu}$ is the default procedure for the F_t. The $\hat{\nu}$ is included in this study for comparison purposes. These methods were compared in terms of type I error and power under conditions of normality and non-normality which will be represented by skewed *g*- and *h*-distributions.

F_t statistic

Lee and Fung [12] 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*-trimmed mean of group *j* is calculated by:

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

where

g_{1j} = number of observations $X_{(i)j}$ such that $(X_{(i)j} - \hat{M}_j) < -2.24$ (scale estimator),

g_{2j} = number of observations $X_{(i)j}$ such that $(X_{(i)j} - \hat{M}_j) > 2.24$ (scale estimator),

$\hat{M}_j$ = median of group *j*.

Scale estimator can be MAD_n, T_n or LMS_n.

The constant 2.24 is the scaling factor to improve the distribution of robust scales computed on non-normal data. This constant was chosen due to the good efficiency of the estimator under normality. Note that 2.24 is approximately equal to the square root of the 0.976 quartile of a chi-square distribution with one degree of freedom.

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

$$SSD_{tj} = (g_{1j}+1)(X_{(g_{1j}+1)j} - \overline{X}_{tj})^2 + (X_{(g_{1j}+2)j} - \overline{X}_{tj})^2$$

$$+ \cdots + (X_{(n_j - g_{2j} - 1)j} - \overline{X}_{tj})^2 + (g_{2j}+1)(X_{(n_j - g_{2j})j} - \overline{X}_{tj})^2$$

$$- \{(g_{1j})[X_{(g_{1j}+1)j} - \overline{X}_{tj}] + (g_{2j})[X_{(n_j - g_{2j})j} - \overline{X}_{tj}]\}^2 / n_j.$$

Hence the trimmed F statistic is defined as

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

where J is the 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.

Scale estimators

The value of a breakdown point is a main factor to be considered when looking for a scale estimator (Wilcox [24]). Rousseeuw and Croux [19] have introduced several scale estimators with highest breakdown point such as MAD_n, T_n and LMS_n. Due to their good performances in Huber [8], Rousseeuw and Croux [19], Yahaya et al. [21] and Md. Yusof et al. [14], these scale estimators were chosen for this study. All these scale estimators have 0.5 breakdown value and also exhibit bounded influence functions. These estimators are also chosen because of their simplicity and computational ease.

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 [19]).

This robust scale estimator is given by

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

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

T_n

Suitable for asymmetric distribution, Rousseeuw and Croux [19] 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

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

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

LMS_n

LMS_n is another scale estimator with a 50% breakdown point. The computation is based on the length of the shortest half sample as shown below:

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

given $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ are the ordered data. The default value of c' is 0.7413 which achieves consistency at Gaussian distributions. LMS_n also has influence function the same as MAD (Rousseeuw and Leroy [18]) and its efficiency equals that of the MAD as well (Grubel [3]).

Empirical Investigation

This paper only focused on unequal sample sizes and homogeneous variances for four groups with small samples. Two cases of groups of size $N = 60$ and $N = 80$ were chosen. For $N = 60$, the sample, were set at $n_1 = 12$, $n_2 = 14$, $n_3 = 16$ and $n_4 = 18$ and for $N = 80$, they were set at $n_1 = 10$, $n_2 = 20$, $n_3 = 20$ and $n_4 = 30$. For both sizes, we used homogeneous variances of 1.

Each method will be 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). The *g*- and *h*-distributions were first proposed by Hoaglin [6]. 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 this study, 5000 datasets were simulated for each of the procedure. The random samples were drawn using SAS generator RANNOR (SAS Institute Inc. [20]).

For type I error, the group means were set as $(0, 0, 0, 0)$. However, in the case of power, one of the group means will be non-zero. Three separation patterns of the means were identified based on the effect size stated in Cohen [2]. The pattern was classified as minimum, intermediate and maximum separation. In minimum, intermediate and maximum case, the group means were set as $(-1, 0, 0, 1)$, $(-1, -0.5, 0.5, 1)$ and $(-1, -1, 1, 1)$, respectively.

Results and Conclusions

The performance of the four F_t procedures under unequal sample sizes and homogeneous variances are shown in the tables and figures below.

Conventionally, a procedure can be considered robust if its type I error is between 0.5α to 1.5α (Bradley [1]). Thus, when the nominal level is set at $\alpha = 0.05$, the type I error rate should be in between 0.025 and 0.075. Type I error rates are considered liberal when they are above the 0.075 limit while those below the 0.025 limit are considered conservative.

Based on Table 1, only the $\hat{\nu}$ column produced *p*-values which are within the Bradley's interval for both total sample sizes. When the total sample size increased from $N = 60$ to $N = 80$, the *p*-values for $\hat{\nu}$ also improved, producing *p*-values which are nearer to the nominal level $(\alpha = 0.05)$. F_t produces liberal type I error, especially for both total sample sizes with new trimming strategies. The only exceptions are under extremely skewed distribution. F_t with the three new trimming strategies produce good type I error rates for both sample sizes.

Table 1. Type I error rates

$N = 60$ (12, 14, 16, 18)					$N = 80$ (10, 20, 20, 30)			
Distribution	F_t with scale estimator Variances = (1:1:1:1)				F_t with scale estimator Variances = (1:1:1:1)			
	MAD_n	T_n	LMS_n	$\hat{\nu}$	MAD_n	T_n	LMS_n	$\hat{\nu}$
$g = 0.0$ and $h = 0.0$	0.1194	0.1030	0.1594	**0.0476**	0.1162	0.1006	0.1460	**0.0532**
$g = 0.5$ and $h = 0.0$	0.1632	0.1546	0.1890	**0.0458**	0.1508	0.1468	0.1686	**0.0530**
$g = 0.5$ and $h = 0.5$	**0.0572**	**0.0502**	**0.0600**	**0.0342**	**0.0542**	**0.0528**	**0.0552**	**0.0474**
Average	0.1133	0.1026	0.1361	**0.0425**	0.1071	0.1000	0.1233	**0.0512**

Looking across Table 2, the average power values show that LMS_n is the best performer for both total sample sizes followed by MAD_n, T_n and $\hat{\nu}$ in the second, third and fourth ranked, respectively. However, the power values corresponding to each particular distribution and setting do not differ much. As the number of observations increased from $N = 60$ to $N = 80$, the power values also increased.

Table 2. Power rates for $J = 4$

Distribution	F_t with scale estimator, $N = 60$				F_t with scale estimator, $N = 80$			
	MAD_n	T_n	LMS_n	$\hat{\nu}$	MAD_n	T_n	LMS_n	$\hat{\nu}$
MINIMUM								
$g = 0.0$ and $h = 0.0$	0.9954	0.9954	0.9924	0.9946	0.9984	0.9988	0.9980	0.9982
$g = 0.5$ and $h = 0.0$	0.9850	0.9842	0.9884	0.9664	0.9960	0.9946	0.9972	0.9902
$g = 0.5$ and $h = 0.5$	0.8764	0.8496	0.9000	0.7030	0.9278	0.9160	0.9476	0.8318
Average	0.9523	0.9431	0.9603	0.8880	0.9741	0.9698	0.9809	0.9401
	MAD_n	T_n	LMS_n	$\hat{\nu}$	MAD_n	T_n	LMS_n	$\hat{\nu}$
INTERMEDIATE								
$g = 0.0$ and $h = 0.0$	0.9998	0.9996	0.9986	0.9992	1.0000	1.0000	0.9998	1.0000
$g = 0.5$ and $h = 0.0$	0.9966	0.9960	0.9972	0.9880	0.9988	0.9984	0.9994	0.9974
$g = 0.5$ and $h = 0.5$	0.9324	0.9108	0.9462	0.7766	0.9690	0.9616	0.9792	0.8920
Average	0.9763	0.9688	0.9807	0.9213	0.9893	0.9867	0.9928	0.9631
	MAD_n	T_n	LMS_n	$\hat{\nu}$	MAD_n	T_n	LMS_n	$\hat{\nu}$
MAXIMUM								
$g = 0.0$ and $h = 0.0$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$g = 0.5$ and $h = 0.0$	1.0000	1.0000	1.0000	0.9996	1.0000	0.9998	1.0000	1.0000
$g = 0.5$ and $h = 0.5$	0.9872	0.9826	0.9914	0.8928	0.9956	0.9944	0.9986	0.9644
Average	0.9957	0.9942	0.9971	0.9641	0.9985	0.9981	0.9995	0.9881

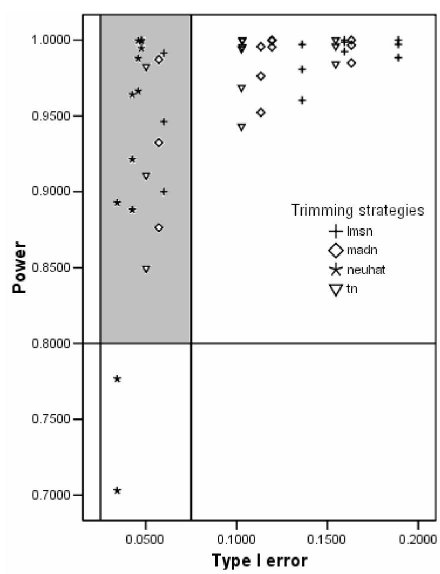

Figure 1. Type I error vs. power for $F_t\ (N = 60)$.

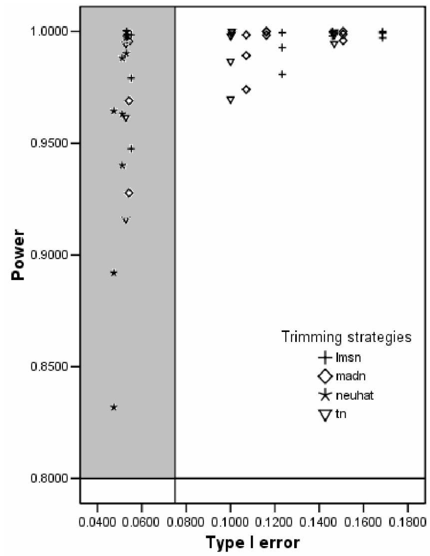

Figure 2. Type I error vs. power for $F_t\ (N = 80)$.

According to Figures 1 and 2, we can observe that the F_t statistic with some of the new trimming strategies proposed in this study show better performance in terms of type I error than $\hat{\nu}$. These type I error rates are close to the nominal level. With regard to the power values, except for a few $\hat{\nu}$ procedures for $N = 60$, all the other procedures generate power values higher than 0.8. As noted by Murphy and Myors [16], the power of a test is judged to be adequate if the value is 0.8 and above.

The shaded area in the figures consist of the trimming strategies that concurrently meet the Bradley's criterion and also have power values higher than 0.8. Based on Figure 2, all the power values of the trimming strategies are located in the area above the 0.8 benchmark.

Conclusions

To evaluate the robustness of a test, several other standards have been used in the past. Procedures that were considered not robust for some researchers could be deemed as robust for others. Some researchers would consider that the procedures with conservative type I error rates fail to perform. However, Mehta and Srinivasan [15] and Hayes [4] stated that conservative procedures in which the true type I error rate is less than or equal to the nominal level can still be considered as robust. Yet a conservative test will be lower in power than a less conservative test because a more conservative test is less likely to reject any null hypothesis (Hayes [4]). As for the liberal test, Hayes [4] suggested avoiding using this test. He defined the liberal test as a test that tends to underestimate the true *p*-value. Using a liberal test for testing hypothesis will increase the probability of type I error to a value greater than the nominal level, which implies that there is a bigger risk of making a type I error. Nonetheless, Keselman et al. [10] pointed out that there is no one universal standard by which tests can be judged to be robust, so different interpretations of these results are possible. This study also identified some promising procedures that performed well in terms of type I error and produced reasonable power.

The best procedure that should be taken into consideration is the F_t with trimming strategy LMS_n. By using this trimming strategy, reasonable type I error and high power rates were achieved simultaneously especially for skewed leptokurtic distribution. This procedure works best under the condition of unequal sample sizes and homogeneous variances.

Acknowledgement

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

References

[1] J. V. Bradley, Robustness? British J. of Mathematical and Statistical Psychology 31 (1978), 144-152.

[2] J. Cohen, Statistical Power Analysis for the Behavioral Sciences, Academic Press, New York, 1988.

[3] R. Grubel, The length of the shorth, Ann. Statist. 16(2) (1988), 619-628.

[4] A. F. Hayes, Statistical Methods for Communication Science, Mahwah, Erlbaum, NJ, 2005.

[5] D. Hertsgaard, Distribution of asymmetric trimmed means, Comm. Statist. Simulation Comput. 8 (1979), 359-367.

[6] D. C. Hoaglin, Summarizing shape numerically: the *g*-and *h*-distributions, Exploring Data Tables, Trends, and Shapes, D. Hoaglin, F. Mosteller and J. Tukey, eds., Wiley, New York, 1985, pp. 461-513.

[7] R. V. Hogg, Adaptive robust procedures: a partial review and some suggestions for future applications and theory, J. Amer. Statist. Assoc. 69 (1974), 909-927.

[8] P. J. Huber, Robust Statistics, Wiley, New York, 1981.

[9] H. J. Keselman, L. M. Lix and R. K. Kowalchuk, Multiple comparison procedures for trimmed means, Psychological Methods 3(1) (1998), 123-141.

[10] H. J. Keselman, R. K. Kowalchuk, J. Algina, L. M. Lix and R. R. Wilcox, Testing treatment effects in repeated measure designs: trimmed means and bootstrapping, British J. of Mathematical and Statistical Psychology 53 (2000), 175-191.

[11] H. J. Keselman, R. R. Wilcox, L. M. Lix, J. Algina and K. H. Fradette, Adaptive robust estimation and testing, British J. of Mathematical and Statistical Psychology 60 (2007), 267-293.

[12] H. Lee and K. Y. Fung, Behaviour of trimmed *F* and sine-wave *F* statistics in one-way ANOVA, Sankhya 47 (1985), 186-201.

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

[14] Z. Md. Yusof, A. R. Othman and S. S. Syed Yahaya, Comparison of type I error rates between T_1 and F_t statistics for unequal population variance using variable trimming, Malaysian J. of Mathematical Sciences 4(2) (2010), 203-215.

[15] J. S. Mehta and R. Srinivasan, On the Behrens-Fisher problem, Biometrika 57 (1970), 649-655.

[16] K. R. Murphy and B. Myors, Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests, Mahwah, Lawrence Erlbaum, NJ, 1998.

[17] A. R. Othman, H. J. Keselman, R. R. Wilcox, K. Fradette and A. R. Padmanabhan, A test of symmetry, J. of Modern Applied Statistical Methods 1 (2002), 310-315.

[18] P. J. Rousseeuw and A. M. Leroy, Robust Regression and Outlier Detection, John Wiley & Sons, New York, 1987.

[19] P. J. Rousseeuw and C. Croux, Alternatives to the median absolute deviation, J. Amer. Statist. Assoc. 88 (1993), 1273-1283.

[20] SAS Institute Inc., SAS Online Doc® 9.1.2. Cary, NC: SAS Institute Inc., 2004.

[21] S. S. Syed Yahaya, A. R. Othman and H. J. Keselman, Testing the equality of location parameters for skewed distributions using S_1 with high breakdown robust scale estimators, Theory and Applications of Recent Robust Methods, Series: Statistics for Industry and Technology, M. Hubert, G. Pison, A. Struyf and S. Van Aelst, eds., Birkhäuser, Basel, 2004, pp. 319-328.

[22] M. L. Tiku, Robustness of MML estimators based on censored samples and robust test statistics, J. Statist. Plann. Inference 4 (1980), 123-143.

[23] M. L. Tiku, Robust statistics for testing equality of means and variances, Comm. Statist. Theory Methods 11 (1982), 2543-2558.

[24] R. R. Wilcox, Introduction to Robust Estimation and Hypothesis Testing, 2nd ed., San Diego, Academic Press, CA, 2005.

[25] R. R. Wilcox, H. J. Keselman and R. K. Kowalchuk, Can tests for treatment group equality be improved? The bootstrap and trimmed means conjecture, British J. of Mathematical and Statistical Psychology 51(1) (1998), 123-134.

[26] R. R. Wilcox and H. J. Keselman, Modern robust data analysis methods: measures of central tendency, Psychological Methods 8 (2003), 254-274.