World Applied Sciences Journal 17 (Special Issue of Applied Math): 66-71, 2012
ISSN 1818-4952

Testing on the Difference of Student's Performance using Robust Methods

Zahayu Md. Yusof, Suhaida Abdullah and Sharipah Soaad Syed Yahaya

School of Quantitative Sciences, UUM College of Arts and Sciences,
Universiti Utara Malaysia, 06010 UUM, Sintok Kedah

Abstract: ANOVA is known to be adversely affected by non-normality and unbalanced design. Type I error and power rates are substantially affected when these problems occur simultaneously. Continuously using ANOVA under the influence of these problems eventually will result in unreliable findings. This study proposed a robust procedure known as modified S_1 and F_t methods. This procedure combines the S_1 and F_t statistics with a popular robust scale estimator, MAD_n. A simulation study was conducted to compare the robustness (Type I error) of the method with respect to its counterpart from the parametric and non parametric aspects namely ANOVA and Kruskal Wallis respectively. Since the null distribution of S_1 is intractable, bootstrap methods were used to give better approximation. The F_t used the approximation method. The findings were in favor of the S_1 and F_t methods especially when the data were skewed. The performance of the methods was further demonstrated on real education data.

Key words: Robust statistics . type i error . robust scale estimators . skewed distributions

INTRODUCTION

Classical statistical methods such as t-test and ANOVA which are frequently used by researchers to test their work are confined to certain assumptions. One of the assumptions is that every classical method has to abide by the assumption that the population under study is normally distributed. Apart from this assumption, researchers also must ensure that the group variances are equal or homogenous. These two assumptions, if occur simultaneously will inflate the Type I error and cause spurious rejection of the null hypothesis. The uninformed usage of these methods under violations of their assumptions eventually will result in unreliable findings. Can we imagine the degree of the damage done to the research due to this mistake? However, most researchers are not aware of the seriousness of the error because they are only the users of the statistical methods. Most quantitative researchers, especially in the field of business, economics and social sciences, rely heavily on the classical methods to solve their problems. Continuous practice of the classical methods without considering the assumptions will most probably generate erroneous results.

In view of all the aforementioned violations, an estimator that is stable and insensitive to all these violations is needed. In other words, the estimator has to be robust. In 1960's, [1, 2] developed the theory of robustness that paved the way for finding practical solutions in statistics. The theory of robustness developed was basically centered on parametric models. That is whilst their methods recognized that the parametric model might not be the "true" model, but nevertheless made inferences about its parameters with robust and efficient methods. Robustness signifies insensitivity to small deviations from the assumptions [3].

As mentioned by [4] small departures from normality can substantially lower the power when comparing the means of two or more groups. Let us look at the example of analysis of variance (ANOVA) and the drawbacks of this method when assumptions are not met. ANOVA is one of the most commonly used statistical methods for locating treatment effects in one-way independent group design. Generally, violating the assumptions associated with standard ANOVA method can seriously hamper the ability to detect true differences. Non-normality and heteroscedasticity are the two usual assumption violations detected in ANOVA. In particular, when these problems occur at the same time, rates of Type I error are usually inflated, thus causing false rejections of the null hypothesis. They can also substantially reduce the power of a test, resulting in treatment effects going undetected. Reduction in the power to detect differences between groups occurs because the usual population standard deviation (σ) is very sensitive to outliers and will be greatly influenced by their presence. Consequently, the standard error of the mean (σ^2/n) can become seriously inflated when the

Corresponding Author: Zahayu Md. Yusof, School of Quantitative Sciences, UUM College of Arts and Sciences, Universiti Utara Malaysia, 06010 UUM, Sintok Kedah

underlying distribution has heavy tails [5]. Therefore, the standard error of the F statistics is larger than it should be and power accordingly will be depressed. In order to achieve a good test, one needs to be able to control Type I error and power of test. In other words, neither should power be lost nor Type I error be inflated.

In our study, we would like to suggest two statistical procedures that are known to be able to handle the problems of non normality and variance heterogeneity simultaneously. These procedures, the modified S_1 and modified F_t statistics are categorized under robust statistics.

The proposed procedures to be adopted in this study are among the latest procedures in robust statistics. Modified S_1 and modified F_t were proposed by [6, 7] respectively. These two procedures are for testing the equality of the central tendency measures for J groups with $H_0 : \theta_1 = \theta_2 = ... = \theta_J$, where θ_j is the central tendency parameter corresponding to distribution F_j: j = 1,2,...,J. S_1 uses median while F_t uses trimmed mean as the central tendency measures.

METHODS

This paper focuses on the modified S_1 and F_t methods, which combines S_1 and F_t statistics with one of the scale estimators suggested by [8].

These methods were compared in terms of Type I error under conditions of normality and nonnormality which will be represented by skewed g-and h-distributions.

S_1 statistic: In the quest for a good robust statistics for testing location parameters for skewed distributions [9] discovered the S_1 statistics which uses the median as the central measures. It is the sum of all possible differences of sample medians from the J distributions divided by their respective sample standard errors.

Let $Y_{ij} = (Y_{1j}, Y_{2j}, ..., Y_{n_j j})$ be a sample from an unknown distribution F_j and let M_i be the population median of F_j: j = 1, 2,..., J. For testing $H_0 : M_1 = M_2 = ... = M_J$ versus H_1: $M_i \neq M_j$ for at least one pair (i,j), the S_1 statistic is defined as

$$S_1 = \sum_{1 \le i < j \le J} |s_{ij}|$$

where

$$s_{ij} = \frac{(\hat{M}_i - \hat{M}_j)}{\sqrt{(\hat{w}_i + \hat{w}_j)}}$$

$$\omega_j = \left(\frac{1}{n_i} \sum |Y_{ij} - \hat{M}_j| \right)^2$$

$$\hat{\omega}_j = \frac{\omega_j}{n_j}$$

$\hat{M}_j$ is the sample median from the jth group, of group j

ω_j is the squared mean absolute deviation from sample median $\hat{M}_j$ and

n_j is the sample size for group j.

Modification on S_1 was done by substituting the default scale estimator, $\hat{\omega}_j$ with the well known robust scale estimator, MAD_n. This scale estimator was chosen based on its robustness properties such as highest breakdown point and bounded influence function. Breakdown point is a measure of an estimator's resistance to contamination. The higher the breakdown point, the more robust is the estimator, for example sample the mean has a low breakdown point, that is 0 and the median has the highest breakdown point that is 0.5. Influence function is the derivative of a statistical functional T(F) that measures the relative extent a small perturbation in F has on T(F). To minimize the influence, the influence function must be bounded. Another advantage of using this estimator is its simplicity, which makes it easy to compute.

F_t statistic: [10] introduced a statistical procedure that is able to handle problems with sample locations when nonnormality occurs but the homogeneity of variances assumption still applies. This statistic is known as trimmed F statistic. We denote it as F_t. They also suggested that this new statistic is used as an alternative to the classical F method involving one-way independent group design. Furthermore, this procedure is easy to compute.

To further understand the F_t method, let

$X_{(1)j}, X_{(2)j}, ..., X_{(n_j)j}$ be an ordered sample of group j with size n_j and let

$k_j = [gn_j]+1$ where [x] is the largest integer ≤x.

We calculated the g-trimmed mean of group j by using:

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

where

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

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

$\hat{M}_j$ = median of group j and the scale estimator MAD_n.

For the equal amounts of trimming in each tail of the distribution, the Winsorized sum of squared deviations is defined as

$$SSD_{tj} = (g_j + 1)\left(X_{(gj+1)j} - \overline{X}_{tj}\right)^2 + \left(X_{(gj+2)j} - \overline{X}_{tj}\right)^2 + ... + \left(X_{(n_j - g_j - 1)j} - \overline{X}_{tj}\right)^2 + (g_j + 1)\left(X_{(n_j - g_j)j} - \overline{X}_{tj}\right)^2$$

When allowing different amounts of trimming in each tail of the distribution, the Winsorized sum of squared deviations is then defined as,

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

Note that we used trimmed means in the SSD_{tj} formula instead of Winsorized means.
Hence the g-trimmed F is defined as

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

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) degree of freedom.
Scale Estimator, MAD_n

Let $X = (x_1, x_2, \ldots, x_n)$ be a random sample from any distribution and let the sample median be denoted by $med_i\ x_i$

MAD_n is median absolute deviation about the median. Given by MADn = b med $|x_i$-med $x_i|$ with b as a constant,this scale estimator is very robust with best possible breakdown point and bounded influence function. [3] identified MAD_n as the single most useful ancillary estimate of scale due to its high breakdown property. MAD_n is simple and easy to compute.

The constant b is needed to make the estimator consistent for the parameter of interest. For example if the observations are randomly sampled from a normal distribution, by including b = 1.4826, the MAD_n will estimate σ, the standard deviation. With constant b = 1, MAD_n will estimate 0.75σ and this is known as MAD.

Bootstrap method: Since the sampling distribution of S_1 is intractable and its asymptotic null distribution may not be of much use for practical sample sizes, the bootstrap method is considered to give a better approximation. Therefore, to assess statistical significance in this study, percentile bootstrap method [11] was used. According to [9], the bootstrap method is known to give a better approximation than the one on the normal approximation theory and this method is attractive, especially when the samples are of moderate size.

Bootstrap was introduced by [12] as a computer-based method for estimating the standard error of an estimator, $\hat{\theta}$. This method has gained a great deal of popularity in empirical research. The word bootstrap is used to indicate that the observed data are used not only to obtain an estimate of the parameter but also to generate new samples from which many more estimates may be obtained and hence an idea of the variability of the estimate [13]. The basic idea is that in the absence of any other information about a population, the values in a random sample are the best guide to the distribution and resampling the sample is the best guide to what can be expected from resampling the population. To obtain the p-value, the percentile bootstrap method is used as follows,

- Calculate S_1 based on the available data.
- Generate bootstrap samples by randomly sampling with replacement n_j observations from the jth group yielding $Y^*_{1j}, Y^*_{2j}, \ldots, Y^*_{n_j j}$.
- Each if the sample points in the bootstrapped groups must be centered at their respective estimated medians.
- Use the bootstrap sample to compute the S_1 statistic denoted by S_1^*.
- Repeat Step 2 to Step 4 B times yielding $S^*_{11}, S^*_{12}, \ldots, S^*_{1B}$. B = 599 appears sufficient in most situations when n≥12 [14].
- Calculate the p-value as (# of $S^*_{1B} > S_1$) / B

Type I error and power of test corresponding to each method will be determined and compared.

EMPIRICAL INVESTIGATION

Since this paper deals with robust method where sensitivity to small changes is of the main concern, manipulating variables could help in identifying the robustness of each method. Four variables were manipulated to create conditions which are known to highlight the strengths and weaknesses of tests for the equality of location parameters.

Number of Groups: Investigations were done on four unbalanced completely randomized groups designs since previous research has looked on these designs [15-17].

Distributional Shape: In investigating the effects of distributional shape on Type I error and power, two types of distribution representing different level of skewness were being considered. The standard normal distribution and the g-and-h distribution with g = 0.5 and h = 0.5. Each of these distributions represents zero and extreme skewness respectively. The skewness for the g-and-h distribution with g = 0.5 and h = 0.5, γ_1 and γ_2 are undefined.

Variance heterogeneity: Variance heterogeneity is one of the general problems in testing the equality of location measures. Therefore, in looking at the effect of this condition to the test, the variances with ratio 1:1:1:36 were assigned to the groups. Though this ratio may seem extreme, ratios similar to this case and larger, have been reported in the literature [18].

Pairings of unequal variances and group sizes: Variances and group sizes were positively and negatively paired for comparison. For positive pairings, the group having the largest group observations was paired with the population having largest group variance and while the group having the smallest number of observations was paired with the population having smallest variance. For the negative pairings, the group with largest number of observations was paired with smallest variance and the group with smallest number of observations was paired with largest group variance. These conditions were chosen since they typically produce conservative results for the positive pairings and liberal results for the negative pairings [19].

The random samples were generated using SAS generator [20]. The variates were standardized and transformed to g-and-h variates having mean μ_j and σ_i^2. The design specification for four groups is shown in Table 1.

To test the Type I error, the group means were (0, 0, 0 and 0). For each design, 5000 datasets were simulated. For S_1 statistic 599 bootstrap samples were generated.

Table 1: Design specification

	Group sizes				Population variances			
	1	2	3	4	1	2	3	4
+ve	10	15	20	25	1	1	1	36
-ve	10	15	20	25	36	1	1	1

SIMULATION RESULTS

The robustness of a method is determined by its ability in controlling the Type I error. By adopting Bradley's liberal criterion of robustness [21], a test can be considered robust if its empirical rate of Type I error α, is within the interval 0.5α and 1.5α. If the nominal level is $\alpha = 0.05$, the empirical Type I error rate should be in between 0.025 and 0.075. Correspondingly, a test is considered to be non-robust if, for any particular condition, its Type I error rate is not within this interval. We chose this criterion since it was widely used by most robust statistic researchers [16, 22-24] to judge robustness. Nevertheless, for [25], if the empirical Type I error rate do not exceed the 0.075 level, it is considered robust. The best procedures are those procedures that can produce Type I error rates closest to the nominal (significance) level.

The Type I error rates presented in Table 2 were obtained from the tests performed on the four groups case.

As can be observed in Table 2, under normal distribution, the average Type I error rates for F_t with MAD_n and ANOVA inflate above the 0.1 level. This is due to the large values of Type I error rates when the pairings are negative. Nevertheless, the modified S_1 method is still in control of its Type I error and the results are consistent for both pairings. Under extremely skewed distribution, again, the average results for F_t with MAD_n and ANOVA show inflated average Type I error rates which are caused by the results of the negative pairings. However, the Type I error for F_t with MAD_n and Kruskall Wallis under positive pairing are robust, but not in the case of ANOVA, which produced the worst result with Type I error for both pairings are above the 0.1 level. In contrast, even though the Type I error rates for S_1 are out of the Bradley's robustness constraint (between 0.025 and 0.075), the Type I error rates are consistent and small. Nevertheless, according to [25], if the empirical Type I error rate do not exceed the 0.075 level, the procedure is considered robust.

ANALYSIS ON REAL DATA

The performance of the modified S_1 and modified F_t methods were then demonstrated on real data. Four

Table 2: Type I error rates

		Methods			
Distribution	Pairing	S_1 with MAD_n	F_t with MAD_n	ANOVA	Kruskall Wallis
Normal	+ve	0.0244	0.0774	0.0336	0.0448
	-ve	0.0260	0.3542	0.2850	0.1158
	Ave	0.0252	0.2158	0.1593	0.0803
g = 0.5	+ve	0.0174	0.0370	0.1492	0.0498
h = 0.5	-ve	0.0194	0.2814	0.3554	0.1022
	Ave	0.0184	0.1592	0.2523	0.0760

Table 3: Descriptive statistics for each group

					95% confidence interval for mean			
Group	n	Mean of the marks	Std. deviation	Std error	Lower bound	Upper bound	Min	Max
1	33	72.07	15.65	2.72	66.53	77.62	7	94
2	19	70.13	9.13	2.10	65.73	74.53	56	90
3	24	73.38	10.75	2.20	68.84	77.91	60	96
4	20	79.21	6.11	1.37	76.35	82.06	68	93

Table 4: Results of the test using different methods

Methods	p-value
S_1 with MAD_n	0.0400
F_t with MAD_n	0.0021
ANOVA	0.0870
Kruskall Wallis	0.0160

classes (groups) of Decision Analysis (2^{nd} Semester 2010/2011) conducted by 4 different lecturers were chosen at random. The final marks were recorded and tested for the equality between the classes. The sample sizes for Class 1, 2, 3 and 4 were 33, 19, 24 and 20 respectively. The result for the descriptive statistics for each of the groups and the results of the test in the form of p-values are given in Table 3 and 4 respectively.

For comparison, the data were tested using all the four procedures mentioned in this study namely ANOVA, Kruskall Wallis, S_1 with MAD_n and F_t with MAD_n. As can be observed in Table 4, when testing using ANOVA, the result fails to reject the null hypothesis such that the performance for all groups is equal which indicates that the test fails to detect the difference which exists between the groups. On contrary, when using Kruskall Wallis, S_1 with MAD_n and F_t with MAD_n, the tests show significant results (reject the null hypothesis). Both the non parametric (Kruskall Wallis) and robust methods (S_1 with MAD_n and F_t with MAD_n) show better detection compared to ANOVA. F_t with MAD_n shows the strongest significance (p = 0.0021) as compared to the other methods followed by Kruskall Wallis and S_1 with MAD_n. However, for F_t with MAD_n and Kruskall Wallis, we have to interpret the result with caution due to the inflated Type I error rates shown in the simulation result.

CONCLUSIONS

The goal of this paper is to find the alternative procedures in testing location parameter for skewed distribution by simultaneously controlling the Type I error and power rates. Classical method such as t-test and ANOVA is not robust to nonnormality and heteroscedasticity. When these problem occur at the same time, the Type I error will increase causing wary rejection of the null hypothesis and power of test can be substantially reduced from theoretical values, which will result in differences going undetected. Realizing the need of a good statistic in addressing these problems, we integrate the S_1 statistic by [9] and F_t statistic introduced by [10] with the high breakdown scale estimators of [8] and these new methods are known as the modified S_1 and F_t methods. This study has shown some improvement in the statistical solution of detecting differences between location parameters.

The result indicates that ANOVA fails to detect the difference which exists between the groups. Both the non parametric (Kruskall Wallis) and robust methods (S_1 with MAD_n and F_t with MAD_n) show better detection. Even though F_t with MAD_n shows stronger significance (p = 0.0021) as compared to the Kruskall Wallis (p = 0.0160), but as shown in the simulation results, F_t with MAD_n in general produced inflated

Type I error rates, which usually results in spurious rejection of the null hypothesis. Thus, misrepresentation of the result could occur.

To improve the performance of the modified S_1 and F_1 methods, we should consider using different types of robust scale estimators. There are plenty of robust scale estimators proposed by [8] we can choose from.

ACKNOWLEDGEMENT

The authors would like to acknowledge the work that led to this paper is partially funded by the University Research Grant Scheme of the Universiti Utara Malaysia.

REFERENCES

1. Huber, P.J., 1964. Robust estimation of a location parameter. Annals of Mathematical Statistics, 35: 73-101.
2. Hampel, F., 1968. Contribution to the theory of robust estimation. Ph.D Thesis, University of California, Berkeley.
3. Huber, P.J., 1981. Robust statistics. New York: Wiley.
4. Wilcox, R.R. and H.J. Keselman, 2003. Modern robust data analysis methods: Measures of central tendency. Psychological Methods, 8: 254-274.
5. Wilcox, R.R. and H.J. Keselman, 2002. Power analysis when comparing trimmed means. Journal of Modern Applied Statistical Methods, 1 (1): 24-31.
6. Syed Yahaya, S.S., 2005. Robust statistical procedures for testing the equality of central tendency parameters under skewed distributions. Unpublished Ph.D. Thesis, Universiti Sains Malaysia.
7. Yusof, Z.Md., A.R. Othman and S.S. Syed Yahaya, 2007. Type I error rates of trimmed F statistic. In proceeding of the 56th Session of the International Statistical Institute (ISI 2007), 22-29 August 2007. Lisbon Portugal. (In CD).
8. Rousseeuw, P.J. and C. Croux, 1993. Alternatives to the median absolute deviation. Journal of the American Statistical Association, 88: 1273-1283.
9. Babu, J.G., A.R. Padmanabhan and M.P. Puri, 1999. Robust one-way ANOVA under possibly non-regular conditions. Biometrical Journal, 41 (3): 321-339.
10. Lee, H. And K.Y. Fung, 1985. Behaviour of trimmed F and sine-wave F statistics in one-way ANOVA. Sankhya: The Indian Journal of Statistics, 47 (Series B): 186-201.
11. Efron, B. and R.J. Tibshirani, 1993. An introduction to the bootstrap. New York: Chapman & Hall.
12. Efron, 1979. Bootstrap methods: Another look at the Jacknife. Annals of Statistics, 7: 1-26.
13. Staudte, R.G. and S.J. Sheather, 1990. Robust estimation and testing. New York: Wiley.
14. Wilcox, R.R., 2005. Introduction to robust estimation and hypothesis testing (2nd Edn). San Diego, CA: Academic Press.
15. Lix, L.M. and H.J. Keselman, 1998. To trim or not to trim: Tests of location equality under heteroscedasticity and nonnormality. Educational and Psychological Measurement, 58 (3): 409-429.
16. Othman, A.R., H.J. Keselman, A.R. Padmanabhan, R.R. Wilcox and K. Fradette, 2004. Comparing measures of the 'typical' score across treatment groups. British Journal of Mathematical and Statistical Psychology, pp: 215-234.
17. Yuen, K.K., 1974. The two-sample trimmed t for unequal population variances. Biometrika, 61: 165-170.
18. Keselman, H.J., R.R. Wilcox, J. Algina, K. Fradette, amd A.R. Othman, 2002. A power comparison of robust test statistics based on adaptive estimators. Journal of Modern Applied Statistical Methods.
19. Othman, A.R., H.J. Keselman, A.R. Padmanabhan, R.R. Wilcox and K. Fradette, 2003. An improved Welch-James test statistic. In proceeding of the Regional Conference on Integrating Technology in the Mathematical Sciences 2003. Universiti Sains Malaysia, Pulau Pinang, Malaysia.
20. SAS Institute Inc., 1999. SAS/IML User's Guide version 8. Cary, NC: SAS Institute Inc.
21. Bradley, J.V., 1978. Robustness?. British Journal of Mathematical and Statistical Psychology, 31: 144-152.
22. Keselman, H.J., R.K. Kowalchuk, J. Algina, L.M. Lix and R.R. Wilcox, 2000. Testing treatment effects in repeated measure designs: Trimmed means and bootstrapping. British Journal of Mathematical and Statistical Psychology, 53: 175-191.
23. Syed Yahaya, S.S., A.R. Othman and H.J. Keselman, 2004. Testing the equality of location parameters for skewed distributions using S_1 with high breakdown robust scale estimators. In Hubert, M., G. Pison, A. Struyf and S. Van Aelst (Eds.). Theory and Applications of Recent Robust Methods, Series: Statistics for Industry and Technology, Birkhauser, Basel, pp: 319-328.
24. Wilcox, R.R., H.J. Keselman, J. Muska and R. Cribbie, 2000. Repeated measures ANOVA: Some new results on comparing trimmed means and means. British Journal of Mathematical and Statistical Psychology, 53: 69-82.
25. Guo, J.-H. and W.-M. Luh, 2000. An invertible transformation two-sample trimmed t-statistic under heterogeneity and nonnormality. Statistic & Probability letters, 49: 1-7.