

Testing the Equality of Location Parameters for Skewed Distributions Using S_1 with High Breakdown Robust Scale Estimators

S.S. Syed Yahaya, A.R. Othman and H.J. Keselman

Abstract. A simulation study had been carried out to compare the Type I error and power of S_1 , a statistic recommended by Babu et al. (1999) for testing the equality of location parameters for skewed distributions. Othman et al. (in press) showed that this statistic is robust to the underlying populations and is also powerful. In our work, we modified this statistic by replacing the standard errors of the sample medians with four alternative robust scale estimators; the median absolute deviation (MAD) and three of the scale estimators proposed by Rousseeuw and Croux (1993); Q_n , S_n , and T_n . These estimators were chosen based on their high breakdown value and bounded influence function, and in addition, they are simple and easy to compute. Even though MAD is more appropriate for symmetric distributions (Rousseeuw and Croux, 1993), due to its popularity and for the purpose of comparison, we decided to include it in our study. The comparison of these methods was based on their Type I error and the power for $J = 4$ groups in an unbalanced design having heterogeneous variances. Data from the Chi-square distribution with 3 degrees of freedom were considered. Since the null distribution of S_1 is intractable, and its asymptotic null distribution may not be of much use for practical sample sizes, bootstrap methods were used to give a better approximation. The S_1 statistic combined with each of the scale estimators was shown to have good control of Type I errors.

Mathematics Subject Classification (2000). 62G10.

Keywords. Type I error, power, bootstrap, skewed distributions, breakdown value.

1. Introduction

Progress has been made in terms of finding better methods for controlling Type I error and power to detect treatment effects in the one-way independent group designs. Through a combination of impressive theoretical developments, more flexible statistical methods, and faster computers, serious practical problems that seemed

insurmountable only a few years ago can now be addressed. These developments are important to applied researchers because they greatly enhance the ability to discover true differences between groups and improved their effort in trying to guard against seeing benefits that do not exist.

Analysis of variance (ANOVA) is one of the most commonly used statistical methods for locating treatment effects in the one-way independent groups design. Generally, violating the assumptions associated with the standard ANOVA method can seriously hamper its ability to detect true differences. Non-normality and heteroscedasticity are the two general problems in ANOVA. In particular, when these problems occur at the same time, rates of Type I error usually are inflated or depressed, resulting in spurious rejections of null hypotheses. They can also substantially reduce the power of a test, resulting in treatment effects going undetected. Reductions in the power to detect differences between groups occur because the usual 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 underlying distribution has heavy tails (Wilcox et al., 1998). Therefore, the standard error of the F statistics in ANOVA is larger than it should be and power accordingly will be depressed.

To achieve a good test, one needs to be able to control Type I errors and to increase the power. We do not want to lose power, and at the same time we do not want to inflate the Type I error. In recent years, numerous methods for locating treatment effects simultaneously controlling Type I error and power to detect treatment effects have been studied. The classical least squares estimators can be highly inefficient in non-normal models. In their effort to control the Type I error and power rate, investigators were looking into numerous robust methods. Robust methods generally are insensitive to assumptions about the overall nature of the data. Robust measures of location such as trimmed means, medians or M-estimators were considered as the alternatives for the usual least squares estimator, that is, the usual mean. These measures of location tendency have been shown to have better control over Type I error and power to detect treatment effects (Othman, Keselman, Padmanabhan, Wilcox, & Fradette, in press). Using trimmed means and variances based on Winsorized sum of squares will enable one to obtain test statistics which do not suffer losses in power due to non-normality (Wilcox, Keselman, & Kowalchuk, 1998).

Babu, Padmanabhan, and Puri (1999) proposed a more flexible statistical method that can deal with asymmetric distributions and heteroscedastic settings. Known as the S_1 statistics, this method is one of the latest procedures in assessing the effects of a treatment variable across groups. Othman et al. (in press) replaced the standard errors of the sample medians in S_1 with asymptotic variances but this modification did not result in better Type I error control compared to the former.

Unlike methods using trimmed means, when using S_1 one can work with the original data without having to transform or to trim the data in achieving symmetry. Simple transformations may fail to deal effectively with outliers and

heavy tails
the observ:
1997).

2. Methc

This paper
 S_1 method
tors propos
and T_n we
tion, the b
compared i
non-norma
 S_1 is more
central ten
the sample

2.1. S_1 Me

When deali
population
sample me
1999). S_1 is
To underst
skewed dist
skewed dist
For testing
(i, j), the ξ

ξ

u

M_j is
 ω_j is 1
 n_j is 1

S_1 is the su
divided by
distribution
sampling d
et al. (in 1
According
approximat
method is

heavy tailed distributions. Even the popular strategy of taking logarithms of all the observations does not necessarily reduce problems due to outliers (Wilcox, 1997).

2. Methods

This paper focuses on the S_1 method and the modified S_1 methods. The modified S_1 methods are the S_1 statistics combined with each of the four scale estimators proposed by Rousseeuw and Croux (1993). These estimators, MAD_n , S_n , Q_n , and T_n were chosen for their high breakdown value and bounded influence function, the basic tools for judging robustness (Wilcox, 1997). The S_1 methods were compared in terms of Type I error and power under conditions of normality and non-normality. Non-normality will be represented by skewed distributions since S_1 is more appropriate for skewed data. S_1 methods use sample medians as the central tendency. Being simple and having the highest possible breakdown value, the sample median is still a popular robust estimator of location.

2.1. S_1 Method

When dealing with a skewed distribution, the parameter of interest is therefore the population median. For this particular case, the S_1 statistics which are based on sample medians will be more appropriate to compare distributions (Babu et al., 1999). S_1 is a solution to the problem when the assumption of symmetry is suspect. To understand S_1 , consider the problem of comparing location parameters for skewed distributions. Let $Y_j = (Y_{1j}, Y_{2j}, \dots, Y_{n_jj})$ be a sample from an unknown skewed distribution F_j and let M_j be the population median of F_j ; $j = 1, 2, \dots, J$. For testing $H_0 : M_1 = M_2 = \dots = M_J$ versus $H_1 : M_i \neq M_j$ for at least one pair (i, j) , the S_1 statistic is defined as

$$S_1 = \sum_{1 \leq i < j \leq J} |s_{ij}| \quad \text{where } s_{ij} = \frac{(\hat{M}_i - \hat{M}_j)}{\sqrt{(\hat{w}_i + \hat{w}_j)}}$$

$$w_j = \left(\frac{1}{n_j} \sum |Y_{ij} - \hat{M}_j| \right)^2 \quad \hat{w}_j = \frac{w_j}{n_j}, \text{ and}$$

\hat{M}_j is the sample median from the j th group,

w_j is the squared mean absolute deviation from sample median $M \cap j$, and

n_j is the sample size for group j .

S_1 is the sum of all possible differences of sample medians from the J distributions divided by their respective sample standard errors, \hat{w} . Therefore, if there are J distributions, then the number of possible differences equals $J(J-1)/2$. Since the sampling distribution of S_1 is unknown, Babu et al. (1999), followed by Othman et al. (in press), used the bootstrap percentile method for obtaining p -values. According to Babu et al. (1999), the bootstrap method is known to give a better approximation than the one based on the normal approximation theory and this method is attractive, especially when the samples are of moderate size. Taking

into consideration the intractability of the sampling distribution of S_1 and the reliability of the bootstrap method, the p-values in our study were obtained by using the percentile bootstrap method (see, e.g., Efron and Tibshirani, 1993). To obtain the p-value, the percentile bootstrap method is used as follows:

1. Calculate S_1 based on the available data.
2. Generate bootstrap samples by randomly sampling with replacement n_j observations from the j th group yielding $Y_{1j}^*, Y_{2j}^*, \dots, Y_{n_jj}^*$
3. Each of the sample points in the bootstrapped groups must be centered at their respective estimated medians.
4. Use the bootstrap sample to compute the S_1 statistic, denoted by S_{1b}^* .
5. Repeat Step 2 to Step 4 B times yielding $S_{11}^*, S_{12}^*, \dots, S_{1B}^*$. $B = 599$ appears sufficient in most situations when $n \geq 12$ (Wilcox, 1997).
6. Calculate the p-value as $(\# \text{ of } S_{1b}^* > S_1) / B$

The amount of computer time depends mainly on how long it takes to evaluate the bootstrap replications and increases linearly with B . The number B varies according to approximations. For estimating the standard error, $B = 50$ is often enough to give a good estimate, while larger B is needed for estimating the percentiles. 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. Hypothesis testing will adopt the same range of B as the percentile to achieve acceptable accuracy.

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

2.2. Scale Estimators

When searching for measures of scale, the breakdown value turns out to have considerable practical importance (Wilcox, 1997). The four scale estimators proposed by Rousseeuw and Croux (1993) have the optimum breakdown value of 0.5. These scale estimators have explicit formulas, which guarantee uniqueness of the estimates. They also have bounded influence functions, which is one of the most important properties for robust estimators. Another advantage of using these estimators is their simplicity, which make them easy to compute.

Let $X = (x_1, x_2, \dots, x_n)$ be a random sample from any distribution and let the sample median be denoted by $\text{med}_i x_i$.

2.2.1. MAD_n . A very robust scale estimator is the median absolute deviation about the median, given by

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

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

The MAD_n has the best possible breakdown value, and its influence function is bounded, with the sharpest possible bound among all scale estimators

(Rousseeuw useful and

Despi efficiency (on dispersi attaches ec not seem t

2.2.2. S_n . be used as efficient an One s

S_n is tion is mov Instead of r a typical d which mea study by F ceeded in r the highest more effic

2.2.3. Q_n . they have c proposed a

where d is

and $h = \lfloor n$
The e formula, su its breakdc function an with small

2.2.4. T_n . (1993) whi T_n defined

(Rousseeuw and Croux, 1993). Huber (1981) identified MAD_n as the single most useful ancillary estimate of scale due to its high breakdown property.

Despite all these advantages, MAD_n has some drawbacks. It has very low efficiency (only 37%) with Gaussian distributions. MAD_n takes a symmetric view on dispersion, because one first estimates a central value (the median) and then attaches equal importance to positive and negative deviations from it, which does not seem to be a natural approach for asymmetric distributions.

2.2.2. S_n . Rousseeuw and Croux (1993) suggested alternatives to MAD_n that can be used as initial or ancillary scale estimates in the same way but that are more efficient and not slanted towards symmetric distributions.

One such estimator is S_n , defined as

$$S_n = c \operatorname{med}_i \{ \operatorname{med}_j |x_i - x_j| \}$$

S_n is very similar to MAD_n . The only difference being that the med_j operation is moved outside the absolute value. This makes S_n a location free estimator. Instead of measuring the deviation of observations from a central value, S_n looks at a typical distance between observations. Another advantage is its explicit formula which means that this estimator is always uniquely defined. A modest simulation study by Rousseeuw and Croux found that the correction factor $c = 1.1926$ succeeded in making S_n unbiased for finite samples. They also proved that S_n has the highest possible breakdown value. In terms of efficiency, S_n was proven to be more efficient (58.23 %) than MAD_n .

2.2.3. Q_n . Even though the influence functions for MAD_n and S_n are bounded, they have discontinuities. For a smooth influence function, Rousseeuw and Croux proposed an estimator Q_n defined as

$$Q_n = d \{ |x_i - x_j|; i < j \}_{(k)}$$

where d is a constant factor,

$$k = \binom{h}{2} \approx \binom{n}{2} / 4$$

and $h = \lceil n/2 \rceil + 1$.

The estimator Q_n shares the attractive properties of S_n ; a simple and explicit formula, suitable for asymmetric distributions, and attains the optimal value for its breakdown value (50 %). Other added advantages are the smooth influence function and the high efficiency (82 %) with Gaussian distributions. However, with small samples, S_n performs better than Q_n .

2.2.4. T_n . Another promising scale estimator proposed by Rousseeuw and Croux (1993) which possesses the attractive properties of the robust scale estimator is T_n defined as

$$T_n = 1.3800 \frac{1}{h} \sum_{k=1}^h \left\{ \operatorname{med}_{j \neq i} |x_i - x_j| \right\}_{(k)}$$

It was proven that T_n has a 50 % breakdown value, a continuous influence function, and an efficiency of 52 %, which makes it more efficient than MAD_n . Like S_n and Q_n , this estimator has a simple and explicit formula which guarantees uniqueness and it is suitable for asymmetric distributions.

Taking into consideration all the attractive properties attached to the scale estimators, such as the breakdown value, continuous influence function, and their efficiency, we substituted the standard errors derived from them in place of $\hat{\omega}$ in S_1 .

3. Procedures

The procedures investigated were:

1. S_1 with MAD_n
2. S_1 with Q_n
3. S_1 with S_n
4. S_1 with T_n
5. S_1 with $\hat{\omega}$

Each of these five methods was tested for treatment group equality under two types of distributions, the normal and skewed distributions. Note that for the rest of this paper, each of these methods will be referred to by its scale estimator, MAD_n , Q_n , S_n , T_n , and $\hat{\omega}$. We compared MAD_n , Q_n , S_n , and T_n with the existing procedure, $\hat{\omega}$ in terms of their Type I error and power rate.

4. Empirical Investigation

For comparison with the work done by Othman et al. (in press), this paper focused on an unbalanced completely randomized design containing four groups with small samples. Since S_1 is appropriate for skewed distributions, we chose the χ_3^2 distribution for simulating the non-normality condition. The skewness and kurtosis values for the χ_3^2 distribution are 1.63 and 4.00 respectively. This distributional shape was chosen for reasons of comparability to the work. Type I error rates had been found to be distorted when the underlying distribution is skewed, e.g., the case of the two sample t-test in Sawilowsky and Blair (1992). Other conditions which are known to highlight the strengths and weaknesses of test for equality of location are heteroscedasticity, and the pairing of variances and group sizes.

For this reason, only unbalanced designs and unequal variance of 36:1 ratio will be considered (see Table 1). Variances and group sizes are both positively and negatively paired. For positive pairings, the group having the fewest number of observations was associated with the population having the smallest variance, while the group having the greatest number of observations was associated with the population having the largest variance, whereas for the negative pairings, the group with largest observations was paired with smallest variance and the group with smallest observations was paired with population having largest variance. These conditions were chosen since they typically produce conservative results for

D

the positive (in press). heterogeneous 1,1,1 respec

Our (unbalanced extreme co by research

The r (tute, 1989) having me shown in T

For T group mea uniquely d Three patt ability. Ou equally spa 1).

For e samples we

5. Result

The result outlined in

TABLE 1. Design Specification for the Four Groups.

PAIRING	GROUP SIZES				GROUP VAR			
	1	2	3	4	1	2	3	4
POSITIVE	10	15	25	30	1	1	1	36
NEGATIVE	10	15	25	30	36	1	1	1

TABLE 2. Type I Error rates.

Distribution	Pairing	SI with corresponding scale estimators				
		MAD _n	Q _n	S _n	T _n	$\hat{\omega}$
χ_3^2	pos	0.027	0.017	0.029	0.032	0.004
	neg	0.034	0.019	0.029	0.037	0.007
Normal	pos	0.025	0.012	0.021	0.023	0.017
	neg	0.029	0.023	0.028	0.030	0.013
	Average	0.029	0.018	0.027	0.031	0.010

the positive pairings and liberal results for the negative pairings (Othman et al., in press). We set the samples at $n_1 = 10, n_2 = 15, n_3 = 25$ and $n_4 = 30$ and heterogeneous variances at 1,1,1, and 36 respectively for positive pairings and 36, 1,1,1 respectively for the negative pairings.

Our choices of these extreme conditions (skewness, heteroscedasticity, and unbalanced designs) were based on the premise that if a procedure works under extreme conditions, it is likely to work under most conditions to be encountered by researchers.

The random samples were drawn using SAS generator RANNOR (SAS institute, 1989). The variates were standardized, and then transformed to χ_3^2 variates having mean μ_j and variance σ_j^2 . The design specification for the four groups is shown in Table 1.

For Type I error, the group means were (0, 0, 0, 0). For power, one of the group means will be non-zero. Cohen (1977) stated that for the effect size to be uniquely determined, the pattern separation of the means should be specified. Three patterns were identified, the minimum, intermediate, and maximum variability. Our study focused on the intermediate variability, where the J means were equally spaced over the range. In this case, the group means were (-1, -0.5, 0.5, 1).

For each of the designs, 1000 datasets were simulated, and 599 bootstrap samples were generated.

5. Results

The results for Type I error and power rates for the methods investigated were outlined in Table 2 and Table 3, respectively.

TABLE 3. Power rates.

Distribution	Pairing	S_1 with corresponding scale estimators				
		MAD_n	Q_n	S_n	T_n	$\hat{\omega}$
χ_3^2	pos	0.078	0.059	0.075	0.091	0.100
	neg	0.088	0.064	0.082	0.102	0.131
Normal	pos	0.403	0.365	0.396	0.408	0.588
	neg	0.227	0.260	0.221	0.278	0.715
	Average	0.199	0.187	0.194	0.220	0.384

Based on the liberal criterion of robustness (Bradley, 1978), a test can be considered robust if its empirical rate of Type I error, $\hat{\alpha}$, is within the interval $0.5\alpha \leq \hat{\alpha} \leq 1.5\alpha$. For the nominal level $\alpha = 0.05$, the Type I error rate should be between 0.025 and 0.075. The empirical Type I error rates in Table 2 indicate robustness in three of the methods investigated. MAD_n , S_n and T_n produced average values ranging from 0.028 to 0.031, all within the Bradley's liberal criterion. These methods produced higher Type I error rates for skewed distribution compared to the normal distributions. Even though the method using Q_n estimator did not satisfy Bradley's liberal criterion, the average error rate for both distributions were higher than the default S_1 method (with $\hat{\omega}$). The average value for Q_n was 0.018 while for $\hat{\omega}$ the average value was only 0.010. However, both methods produced average Type I error rates which were considered to be too conservative, meaning that the estimated rates of Type I error were below 0.025. $\hat{\omega}$ produced a more conservative Type I error rate for skewed distribution.

For both distributions, the empirical Type I error rates for the positive pairings were smaller than for negative pairings, except for the pairings for S_n which showed no variability when the data were skewed.

Our new methods, combining the Babu et al. (1999) S_1 and Rousseeuw and Croux (1993) scale estimators were able to show some improvement over the default S_1 using $\hat{\omega}$ in terms of Type I error rate. The T_n method resulted in the best average error rate of 0.031, which was nearest to the nominal level. All the methods studied (excluding $\hat{\omega}$) produced better average error rate for the skewed distribution compared to normal distribution. The average error rate across the three methods, MAD_n , S_n , and T_n , exhibited small variability for both distributions.

The average power values outlined in Table 3 show two sets of results. The low values belong to the four methods when data were skewed, whereas the larger values were obtained when data were normal. Ranging from 0.075 to 0.131, the average power rate for the new methods under skewed distribution were low. The average values when data were normal ranged from a low of 0.221 and to a high of 0.408. The default S_1 under normal distribution resulted in an average value rate of 0.7. Even though the default S_1 under skewed distributions produced the highest average power rate compared to the rest of the methods, the value of

0.131 was little varia
The
much vari
was obvio

6. Concl

This pape
research, v
tistics whi
of a good s
et al. (199
(1993). Th
cating tree
this paper
When syn
tions by u
better Ty
methods, l
methods, c

The r
with value
The f
research de
on exponen
which were

Referenc

- [1] G.J. B
Possibi
- [2] J.V. B:
- [3] J. Coh
New Y
- [4] B. Efr
1-26.
- [5] B. Efr
Inc, Ne
- [6] A.R. C
Compa
Math.
- [7] P.J. R
Amer.

0.131 was still very low. The mean values across the four methods showed very little variability for both distributions.

The negative and positive pairings for the skewed distributions did not show much variability for each method, but for the normal distribution, the variability was obvious for different pairings.

6. Conclusion

This paper focused on the situation, common in psychological and educational research, where the observations are from skewed distributions. One requires statistics which are robust especially in locating treatment effects. Realizing the need of a good statistic in addressing this problem, we integrate the S_1 statistic by Babu et al. (1999) with the high breakdown scale estimators of Rousseeuw and Croux (1993). This paper has shown some improvement in the statistical solution of locating treatment effects. In controlling the Type I error rate, the study reported in this paper leads us to formulate the following conclusions and recommendations. When symmetry is suspect, we can avoid trimming or transforming the observations by using one of the methods in our paper. These new methods produced better Type I error rates than the default S_1 using $\hat{\omega}$. Three of the investigated methods, MAD_n , S_n , and T_n , reasonably controlled Type I errors; the remaining methods, Q_n and $\hat{\omega}$, were conservative at a significance level of 0.05.

The methods are considered robust when they meet the criteria for robustness with values in between 0.025 and 0.075 for 0.05 level.

The findings on power rate did not show any improvement from the previous research done by Othman et al. (in press). Babu et al. (1999) in their investigation on exponential and log-normal distributions with S_1 also produced low power rates which were less than 0.10.

References

- [1] G.J. Babu, A.R. Padmanabhan, and M.L. Puri, *Robust One-Way ANOVA Under Possibly Non-Regular Conditions*. *Biom. J.* **41** (1999), 321-339.
- [2] J.V. Bradley, *Robustness?* *British J. Math. Statist. Psych.* **31** (1978), 144-152.
- [3] J. Cohen, *Statistical Power Analysis For the Behavioral Sciences*. Academic Press, New York, 1977.
- [4] B. Efron, *Bootstrap Methods: Another Look at the Jackknife*. *Ann. Statist.* **7** (1979), 1-26.
- [5] B. Efron and R.J. Tibshirani, *An Introduction to the Bootstrap*. Chapman and Hall Inc, New York, 1993.
- [6] A.R. Othman, H.J. Keselman, A.R. Padmanabhan, R.R. Wilcox, and K. Fradette, *Comparing Measures of the "Typical" Score Across Treatment Groups*. *British J. Math. Statist. Psych.* (in press).
- [7] P.J. Rousseeuw and C. Croux, *Alternatives to the Median Absolute Deviation*. *J. Amer. Statist. Assoc.* **88** (1993), 1273-1283.

- [8] P.J. Rousseeuw and S. Verboven, *Robust Estimation in Very Small Samples*. Comput. Statist. Data Anal. **40** (2002), 741-758.
- [9] SAS Institute Inc., *IML Software: Usage and Reference, Version 6*. 1st Edition, Cary, New York, 1989.
- [10] S.S. Sawilowsky and R.C. Blair, *A More Realistic Look at the Robustness and Type II Error Probabilities of the t Test to Departures From Population Normality*. Psychol. Bullet. **111** (1992), 352-360.
- [11] R.G. Staudte and S.J. Sheather, *Robust Estimation and Testing*. John Wiley and Sons, Inc., New York, 1990.
- [12] R.R. Wilcox, *Introduction to Robust Estimation and Hypothesis Testing*. Academic Press, New York, 1997.
- [13] R.R. Wilcox and H.J. Keselman, *Power Analyses When Comparing Trimmed Means*. J. Modern Appl. Statist. Methods **1** (2002), 24-31.
- [14] 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. Math. and Statist. Psych. **51** (1998), 123-134.

S.S. Syed Yahaya
School of Quantitative Sciences, Universiti Utara Malaysia
06010 Sintok, Kedah, Malaysia
e-mail: syedys8@yahoo.com

A.R. Othman
School of Distance Education, Universiti Sains Malaysia
11800 Minden, Penang, Malaysia
e-mail: oarahman@usm.my

H.J. Keselman
Department of Psychology, University of Manitoba, 190 Dysart Road
Winnipeg, Manitoba R3T 2N2, Canada
e-mail: kesel@cc.umanitoba.ca