Sains Malaysiana 40(5)(2011): 521–525

A Robust Test Based on Bootstrapping for the Two-Sample Scale Problem

(Suatu Ujian Teguh Berdasarkan Kaedah Butstrap
untuk Masalah Skala bagi Dua-Sampel)

A. R. Padmanabhan, Abdul Rahman Othman & Teh Sin Yin*

ABSTRACT

For testing the homogeneity of variances, modifications of well-known tests are known which combine rigorous theory with resampling (bootstrap). We propose versions of these tests, which are computationally simpler (although asymptotically equivalent). The earlier procedures used the smooth bootstrap with two thousand bootstrap replications per sample whereas our proposals use only the classical bootstrap (or percentile method) with just one thousand bootstrap replications per sample, and also required much less computing time. Our proposals cover the Ansari-Bradley-, Mood- and Klotz-tests. We explain their superiority over the existing methodologies available in textbooks and packages.

Keywords: Ansari-Bradley; Bootstrap; Klotz; Mood; test of scale

ABSTRAK

Pada umumnya, pengubahsuaian ujian-ujian terkenal yang menggabungkan teori rapi dengan pensampelan semula (kaedah butstrap) digunakan untuk mengkaji kesamaan varians. Kami mencadangkan versi ujian yang lebih mudah daripada segi pengiraan (meskipun ia setara secara asimtotik). Prosedur yang sebelum ini menggunakan kaedah butstrap licin dengan dua ribu replikasi butstrap setiap sampel. Kami pula mengusulkan penggunaan kaedah butstrap klasik (atau kaedah persentil) dengan hanya seribu replikasi butstrap setiap sampel. Maka masa pengiraan juga jauh lebih singkat. Usul kami merangkumi ujian Ansari-Bradley, Mood dan Klotz. Kami menjelaskan keunggulan ujian-ujian tersebut berbanding dengan kaedah yang tersedia dalam buku teks dan pakej perisian di pasaran.

Kata kunci: Ansari-Bradley; kaedah Butstrap; Klotz; Mood; ujian skala

Introduction

The problem of testing for scale arises in a variety of contexts, including quality control and analysis of outer continental shelf bidding on oil and gas (Conover et al. 1981), chemistry (Bethea et al. 1975), and engineering (Hald 1967; Menden-hall et al. 1990; Nair 1984). Therefore, it is worthwhile to develop an effective methodology for this problem. In this connection, Hall and Padmanabhan (1997) have proposed some tests, which are Fligner-Killeen (F-K:med) type modifications of the Ansari-Bradley-, Mood- and Klotz-tests and estimated their quantiles by means of the smooth bootstrap.

This article proposes variants of the above tests, called refined robust tests, or refinements, for short, and shows that they are superior to the standard tests and in addition their performance improves with increasing sample sizes.

Essentially, the standard tests of scale procedures given in the text books (i.e. Ansari-Bradley1960; Mood 1954; Klotz 1960 and Siegel-Tukey 1960) are meant only for symmetric distributions. They should not be used for skewed distributions. In fact, rigorous theory (Fligner & Hettmansperger 1979) shows that the more the skewness of the underlying distribution, the higher the actual (empirical) level will be. Typically, it will be way above the nominal level.

This paper is organized as follows. The second section reviews the scale statistics. The third section describes the simulation studies by presenting extensive computer-intensive (bootstrap) method. The fourth section contains the main results and discussions. The final section presents the conclusions.

The Scale Statistics

The standard tests for the two-sample scale problem are based on linear rank statistics of the form:

$$h = \sum_{i=1}^{n_1} a_N\left(R_{1i}\right), \tag{1}$$

where $a_N(1)$, $a_N(2)$, …, $a_N(N)$ are a set of scores; n_1 and n_2 are the two sample sizes, $N=n_1+n_2$, and R_{1i} is the rank for the i^{th} observation in the first sample in the combined sample of size N adjusted (aligned) for location. The scores for Ansari-Bradley- (Ansari & Bradley 1960), Mood- (Mood 1954) and Klotz- (Klotz 1962) are given in Table 1, where Φ^{-1} denotes the inverse of the cumulative standard normal distribution function.

TABLE 1. Scale statistics for the two-sample scale problem

Score Type	Standard Statistic
Ansari-Bradley	$h_1 = \sum_{i=1}^{n_1} \left\{ \frac{N+1}{2} - \left\| R_{1i} - \frac{N+1}{2} \right\| \right\}$
Mood	$h_2 = \sum_{i=1}^{n_1} \left[R_{1i} - \frac{N+1}{2} \right]^2$
Klotz	$h_3 = \sum_{i=1}^{n_1} \left\{ \Phi^{-1} \left[R_{1i} - \frac{N+1}{2} \right] \right\}^2$
Score Type	Fligner & Killeen (1976) Modifications Statistic
Ansari-Bradley	$h_1 = \sum_{i=1}^{n_1} \left(\frac{R_{1i}}{N} + 1 \right)$
Mood	$h_2 = \sum_{i=1}^{n_1} \left(\frac{R_{1i}}{N+1} \right)^2$
Klotz	$h_3 = \sum_{i=1}^{n_1} \left\{ \Phi^{-1} \left[\frac{N+R_{1i}}{2N+1} \right] \right\}^2$

SIMULATION

Let $X_1 = (X_{1,1}, \ldots, X_{1,m_1})$ and $X_2 = (X_{2,1}, \ldots, X_{2,m_2})$ denote two independent samples and M_1 and M_2 represent medians of the first sample and second sample, respectively. The statistic h denotes any of the standard rank statistics we evaluate (Table 1). The simulation for the two-sample scale problem was performed used the Statistical Analysis Software (SAS) package based on 2,000 samples with 1,000 and 2,000 bootstrap samples drawn from each sample. Three important skewed distributions were simulated, each of size (20, 20) and (40, 40). They were 3 degree chi-square distribution ($\chi^2(3)$), exponential and standard lognormal distributions. The $\chi^2(3)$ -distribution was chosen because educational and psychological research data typically have this skewed distribution (Keselman et al. 1998a; Keselman et al. 1998b; Keselman et al. 2007; Keselman et al. 2004; Othman et al. 2002). The last two distributions were chosen because of their importance in biostatistics, industrial engineering and reliability (Steland et al. 2011).

The algorithm to conduct the simulation in Hall and Padmanabhan (1997) is as follows:

1) Simulate sample of χ^2_3 -distribution with sample size $(n_1, n_2) = (20, 20)$.
2) Compute test statistic with sample aligned (adjusted) for location, that is $\tau_1 = h(X_1 - X_2 - M_2)$.
   Compute the samples adjusted for both location and scale, $\xi_1 = (\xi_{1,1}, \ldots, \xi_{1,n_1})$ and $\xi_2 = (\xi_{2,1}, \ldots, \xi_{2,n_2})$, where MAD_i denote the median of the absolute deviations from the median of the X_i sample. Therefore, for the general formula is defined as:
   $$\xi_{i,j} = \frac{X_{i,j} - M_i}{\text{MAD}_i}, \quad i = 1,2 \quad j = 1,2,\ldots,n_i.$$
3) Then, combine the two samples to obtain the pooled sample (ξ_1, ξ_2).
4) Denote by $U_{(1)} \le U_{(2)} \le \ldots \le U_{(n_1+n_2)}$ the ordered statistics of the pooled sample (ξ_1, ξ_2).
5) Define $U_{(n_1+n_2+1)} = 2U_{(n_1+n_2)} - U_{(n_1+n_2-1)}$.
6) Let F^* denote the continuous distribution function that assign uniformly the probability $\frac{1}{n_1+n_2}$ to the interval $\left(U_{(k)}, U_{(k+1)}\right)$, where $k = 1, 2, \ldots, n_1 + n_2$.
7) A bootstrap sample of size $n_1 + n_2$, say $Z^*_1, \ldots, Z^*_{n_1}$, $Z^*_{n_1+1}, \ldots, Z^*_{n_1+n_2}$ is drawn from the interval. The value $Z^*_k = R_k U_{(k)} + (1 - R_k) U_{(k+1)}$, where R_k is a random number between 0 and 1.
8) Write $\xi^*_{1,1} = Z^*_1, \ldots, \xi^*_{1,n_1} = Z^*_{n_1}$ and $\xi^*_{2,1} = Z^*_{n_1+1}, \ldots, \xi^*_{2,n_2} = Z^*_{n_1+n_2}$, such as bootstrap sample of size n_1 for the first sample is $\xi^*_1 = \left(\xi^*_{1,1}, \ldots, \xi^*_{1,n_1}\right)$ and bootstrap sample of size n_2 for the second sample is $\xi^*_2 = \left(\xi^*_{2,1}, \ldots, \xi^*_{2,n_2}\right)$.
9) Let M^*_1 and M^*_2 be the medians of the first and second bootstrap samples respectively and $\xi^*_1 - \left(\frac{n_1 - 1}{n_1}\right) M^*_1$ and $\xi^*_2 - \left(\frac{n_2 - 1}{n_2}\right) M^*_2$ be the corresponding samples aligned (adjusted) for location.
10) Let the value of the (test) statistic based on these aligned samples be $V^*_{1,1}$.
11) Repeat this process 1,999 more times to get 2,000 values; say, $V^*_{1,1}, V^*_{1,2}, \ldots, V^*_{1,2000}$.
12) Then, arrange $V^*_{1,1}, V^*_{1,2}, \ldots, V^*_{1,2000}$ in the increasing order such as $V^*_{(1,1)} \le V^*_{(1.2)} \le \ldots \le V^*_{(1,2000)}$ be the ordered statistics of the pooled samples.
13) Perform one-sided test with nominal level. If $T_1 < V^*_{(1,100)}$, then reject null hypothesis (H_0). Otherwise, accept H_0.
14) Repeat Step 1-13 for 2,000 replications.
15) Calculate the empirical level of the standard rank statistic (proportion of empirical rejections) based on 2,000 bootstrap samples (for each sample), i.e. $\frac{\text{total number of rejections}}{\text{2000 replications}} \cong 0.05$.
16) Repeat Step 2 to Step 15 by using exponential distribution and lognormal distribution, respectively.
17) Repeat Step 1 to Step 16 for sample size (40, 40).

We propose a computationally simpler modification of these tests by using classical bootstrap instead of the smooth bootstrap. This is done by replacing steps 4 to 8 in the algorithm above with a single step below:

Next, bootstrapping the pooled sample (ξ_1, ξ_2) to obtain bootstrap sample of size n_1 for the first sample and bootstrap sample of size n_2 for the second sample, such as $\xi^*_1 = \left(\xi^*_{1,1}, \ldots, \xi^*_{1,n_1}\right)$ and $\xi^*_2 = \left(\xi^*_{2,1}, \ldots, \xi^*_{2,n_2}\right)$.

The samples aligned (adjusted) for location in Step 9 used in our method is $\xi_1^* - M_1^*$ and $\xi_2^* - M_2^*$. Note that in Step 13, the continuity correction was applied for our method, i.e. $T_1 < V^*_{(1,90)}$, since the frequent occurrence of ties (resolved by the mid-rank method) made the bootstrap distribution very discrete and condensed. Thus we applied the continuity correction for improving the approximation to the discrete distribution (by a continuous distribution). This was inspired by the case of approximating binomial by normal distribution, using the continuity correction (i.e. P(Stat ≥ K) = P(Stat ≥ K +0.5) and P(Stat ≤ K) = P(Stat ≤ K -0.5), K = integer). For our method, we also repeated the whole process for 1,000 bootstrap samples, instead of 2,000 bootstrap samples.

Results and Discussion

Table 2 shows that for $\chi^2(3)$ -distribution with sample size (20,20), the empirical levels of our tests, based on 1,000 bootstrap samples, are .05 (Ansari-Bradley), .056 (Mood) and .0725 (Klotz), whereas the corresponding levels based on the existing methodologies are .089, .1035 and .104, way above the nominal levels, showing their non-robustness.

In fact, the higher the skewness, the more pronounced the superiority of our methodology (over the existing ones) becomes, as shown by the results for the exponential distribution (Table 3) which is more skewed than $\chi^2(3)$ and the lognormal (Table 4), which is the most skewed of the three distributions. For the exponential distribution with sample size (20,20), the empirical levels of our tests (based on 1,000 bootstrap samples) are .0555 (Ansari-Bradley), .066 (Mood) and .0825 (Klotz), whereas the corresponding levels based on the existing methodologies are around .14. For the lognormal, the empirical levels are .063 (Ansari-Bradley), .0705 (Mood) and .081 (Klotz), whereas the corresponding levels based on the existing methodologies are .1425 (Ansari-Bradley), .1685 (Mood) and .1725 (Klotz).

Tables 2 to 4 indicate that with sample size (20,20), the results for the robust tests based on 2,000 bootstrap samples, are slightly inferior to those based on 1,000 bootstrap samples. But, the opposite is observed with the sample size (40,40). This means that when the sample size increases, the number of bootstrap need to be amplify to obtain robust empirical values. The results in Table 2 to 4 also show that the empirical levels of our tests with sample size (40,40) closer to .050 compare with sample size (20,20). While, the empirical levels based on the existing methodologies with sample size (40,40) are far way above the nominal levels compare with sample size (20,20).

The empirical values of our method are comparable to the Hall and Padmanabhan (1997) method. In fact, our methodology requires much less computing time than that of Hall and Padmanabhan (1997), as explained below:

1. Hall and Padmanabhan (1997) require the smooth bootstrap, whereas we use only the classical bootstrap.
2. Hall and Padmanabhan (1997) work with Fligner-Killeen (1976) modifications of the statistics, whereas our procedure required no such modifications.
3. Hall and Padmanabhan (1997) applied smooth bootstrap to overcome the problem of tie which is time consuming, whereas we resolve ties by the mid-rank method which is build in function in most of the statistical packages (e.g. SAS, PASW (formerly SPSS), Minitab, etc.).
4. No continuity correction was used on Hall and Padmanabhan (1997) procedure, whereas continuity correction was applied in our method for improving the approximation to the discrete distribution.

TABLE 2. Simulate sample of $\chi^2(3)$-distribution with sample size (20,20) and (40,40)

Score Type	Empirical value for sample size (20,20)			
	Standard Test of Scale	Hall and Padmanabhan (1997) Procedure	Refinement of Hall and Padmanabhan (1997) Procedure	
		Number of Bootstrap	Number of Bootstrap	
		2,000	1,000	2,000
Ansari-Bradley	.0890	.0530	.0500	.0520
Mood	.1035	.0410	.0560	.0585
Klotz	.1040	.0460	.0725	.0725
Score Type	Empirical value for sample size (40,40)			
	Standard Test of Scale	Hall and Padmanabhan (1997) Procedure	Refinement of Hall and Padmanabhan (1997) Procedure	
		Number of Bootstrap	Number of Bootstrap	
		2,000	1,000	2,000
Ansari-Bradley	.0955	.0545	.0515	.0500
Mood	.1240	.0485	.0515	.0495
Klotz	.1325	.0475	.0635	.0620

TABLE 3. Simulate sample of exponential distribution with sample size (20,20) and (40,40)

Score Type	Empirical value for sample size (20,20)			
	Standard Test of Scale	Hall and Padmanabhan (1997) Procedure	Refinement of Hall and Padmanabhan (1997) Procedure	
		Number of Bootstrap	Number of Bootstrap	
		2,000	1,000	2,000
Ansari-Bradley	.1375	.0635	.0555	.0595
Mood	.1465	.0355	.0660	.0715
Klotz	.1490	.0405	.0825	.0835
Score Type	Empirical value for sample size (40,40)			
	Standard Test of Scale	Hall and Padmanabhan (1997) Procedure	Refinement of Hall and Padmanabhan (1997) Procedure	
		Number of Bootstrap	Number of Bootstrap	
		2,000	1,000	2,000
Ansari-Bradley	.1595	.0640	.0565	.0560
Mood	.1855	.0385	.0570	.0555
Klotz	.2095	.0400	.0740	.0745

TABLE 4. Simulate sample of lognormal distribution with sample size (20,20) and (40,40)

Score Type	Empirical value for sample size (20,20)			
	Standard Test of Scale	Hall and Padmanabhan (1997) Procedure	Refinement of Hall and Padmanabhan (1997) Procedure	
		Number of Bootstrap	Number of Bootstrap	
		2,000	1,000	2,000
Ansari-Bradley	.1425	.0700	.0630	.0650
Mood	.1685	.0370	.0705	.0715
Klotz	.1725	.0365	.0810	.0800
Score Type	Empirical value for sample size (40,40)			
	Standard Test of Scale	Hall and Padmanabhan (1997) Procedure	Refinement of Hall and Padmanabhan (1997) Procedure	
		Number of Bootstrap	Number of Bootstrap	
		2,000	1,000	2,000
Ansari-Bradley	.1550	.0585	.0535	.0520
Mood	.1885	.0490	.0590	.0550
Klotz	.1990	.0490	.0695	.0685

CONCLUSIONS

This study showed that our proposed robust tests for the two-sample scale problem work well when the underlying distribution are skewed. This new method has empirical values closer to the nominal level compared to the standard tests of scale. All the standard tests of scale have liberal empirical values ranging from .0890 to .2095, the liberalism increasing with the skewness; that is, the higher the skewness of the distribution, the more liberal the test becomes. Moreover, rigorous theory (Fligner & Hettmansperger 1979) tells us that while the performance of our robust tests give better results with increasing of sample sizes, exactly the opposite happens with the methodologies now available. This study could be extended to the multi-sample case in an obvious fashion.

In the light of these findings, we make the following recommendations: when robustness is the main criterion, choose the Ansari-Bradley statistic. Suppose, some information about the tail-weight of the underlying distribution is available or the tail-weight can be estimated as in Hall and Padmanabhan (1997). Based on this additional information, we can choose one of the above three tests (i.e. Ansari-Bradley, Mood, or Klotz). This way we can achieve both robustness of level and efficiency of power.

ACKNOWLEDGMENTS

We greatly appreciate the helpful comments of the anonymous referee and editor. Their comments have contributed in the improvement of this article. The work that led to the publication of this paper was funded by the Short Term Grant Scheme of Universiti Sains Malaysia (USM) and supported by the USM Fellowship.

REFERENCES

Ansari, A.R. & Bradley, R.A. 1960. Rank-sum tests for dispersions. *Ann. Math. Stat.* 31: 1174-1189.

Bethea, R.M., Duran, B.S. & Boullion, T.L. 1975. *Statistical Methods for Engineers and Scientists*. New York: Marcel-Dekker.

Conover, W.J., Johnson, M.E. & Johnson, M.M. 1981. A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics* 23: 351-361.

Fligner, M.A. & Hettmansperger, T.P. 1979. On the use of conditional asymptotic normality. *J. Roy. Statistical Society* Ser. B 41: 178-183.

Fligner, M.A. & Killen, T.J. 1976. Distribution-free two-sample tests for scale. *J. Amer. Statistical Assoc*. 71: 210-213.

Hald, A. 1967. *Theory with Engineering Applications*. (7th ed.). New York: Wiley.

Hall, P. & Padmanabhan, A.R. 1997. Adaptive inference for the two-sample scale problem. *Technometrics* 39(4): 412-422.

Keselman, H.J., Kowalchuk, R.K. & Lix, L.M. 1998a. Robust nonorthogonal analyses revisited: An update based on trimmed means. *Psychometrika* 63: 145–163.

Keselman, H.J., Lix, L.M. & Kowalchuk, R.K. 1998b. Multiple comparison procedures for trimmed means. *Psychol. Methods* 3(1): 123-141.

Keselman, H.J., Othman, A.R., Wilcox, R.R. & Fradette, K. 2004. The new and improved two-sample *t*-test. *Psychol. Science* 15: 47-51.

Keselman, H.J., Wilcox, R.R., Lix, L.M., Algina, J. & Fradette, K. 2007. Adaptive robust estimation and testing. *British J. Math. Statist. Psych*. 60: 267-293.

Klotz, J. 1962. Nonparametric tests for scale. *Ann. Math. Stat.* 32: 498-512.

Mendenhall, W., Wackerly, D.D. & Schaeffer, R.L. 1990. *Mathematical Statistics with Applications*. (4th ed.). Boston: PWS-Kent.

Mood, A. 1954. On the asymptotic efficiency of certain non-parametric tests. *Ann. Math. Stat*. 25: 514-522.

Nair, V.N. 1984. On the behavior of some estimators from probability plots. *J. Amer. Statistical Assoc*. 79: 823-830.

Othman, A.R., Keselman, H.J., Wilcox, R.R., Fradette, K. & Padmanabhan, A.R. 2002. A test of symmetry. *J. Mod. Appl. Stat. Meth.* 2: 310-315.

Siegel, S. & Tukey, J.W. 1960. A nonparametric sum of ranks procedure for relative spread in unpaired samples. *Journal of American Statistical Association* 55(291): 429-445.

Steland, A., Padmanabhan, P. & Akram, M. 2007. *Resampling methods for the nonparametric and generalized behrens-fisher problem.* Sankhya Ser. A (accepted).

A.R. Padmanabhan, Abdul Rahman Othman & Teh Sin Yin*
Robust Statistics Computational Laboratory
School of Distance Education
Universiti Sains Malaysia
11800 Minden, Penang
Malaysia

A.R. Padmanabhan
Monash University
Clayton, Victoria 3800
Australia

Abdul Rahman Othman
Institute of Postgraduate Studies
Universiti Sains Malaysia
11800 Minden, Penang
Malaysia

Teh SinYin*
School of Mathematical Sciences
Universiti Sains Malaysia
11800 Minden
Penang, Malaysia

*Corresponding author; email: syin.teh@gmail.com

Received: 22 March 2010
Accepted: 1 October 2010