Available online at www.sciencedirect.com
ScienceDirect
Journal of Statistical Planning and Inference 137 (2007) 3261 – 3267
journal of statistical planning and inference
www.elsevier.com/locate/jspi

Re-sampling methods for testing for location against unrestricted and ordered alternatives

Gutti Jogesh Babu[a,*,1], A.R. Padmanabhan[b]

[a]*Department of Statistics, Pennsylvania State University, PA 16802, USA*
[b]*Department of Mathematics and Statistics, Monash University, Clayton 3800, Australia*

Available online 30 March 2007

Abstract

Robust tests are proposed for unrestricted and ordered alternatives in the multi-sample problem without assuming homogeneity of scales and/or symmetry of the underlying distributions. The methodology consists of bootstrapping appropriately centered Mann–Whitney statistics. Data sets from Physics and Psychology illustrate the methodology.

MSC: 62G09; 62G10

Keywords: Bootstrap; Mann–Whitney statistic; Multi-sample F–K:med statistic

1. Introduction

The two-sample and multi-sample location problems assume normality and homogeneity of error variances. However, violations of these assumptions are common (cf. Chow and Liu, 1992; Cressie, 1981, 1997; Micceri, 1989; Wilcox, 1987, 1995). The more realistic nonparametric Behrens–Fisher problem of testing for the equality of the medians of the distributions with possibly unequal variances was investigated under the additional assumption of symmetry, in the two-sample case by Fligner and Policello (1981) and in the multi-sample case by Rust and Fligner (1984). In presence of heteroscedasticity, Cressie (1997) uses optimal weighted linear combination of sample means. However, for skewed distributions, his jackknife procedure seems to require fairly large samples (Cressie, 1997, p. 50).

Nevertheless most of the distributions arising in Bioavailability Studies and Psychology are skewed (cf. Chow and Liu, 1992; Micceri, 1989; Wilcox, 1995). Therefore it is worthwhile to develop a methodology without the symmetry assumption. This was investigated in Babu and Padmanabhan (2002) in the two-sample case. The present paper extends it to the multi-sample case and also proposes a methodology for testing for ordered alternatives, thereby generalizing the results of Jonckheere (1954).

Section 2 describes the preliminaries. Section 3 provides the simulation studies and a discussion of the results. Section 4 contains some illustrations. The justification for bootstrap asymptotics follows from the theory in Babu and Padmanabhan (2002).

* Corresponding author. Tel.: +1 814 8651348; fax: +1 814 8637114.
E-mail address: babu@stat.psu.edu (G.J. Babu).
[1] Research supported in part by NSF Grants AST-0326524 and AST-0434234.

0378-3758/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2007.03.010

2. Bootstrap procedures

For $c > 2$, let $X_1 = (X_{1,1}, \ldots, X_{1,n_1})$, $X_2 = (X_{2,1}, \ldots, X_{2,n_2})$, ..., and $X_c = (X_{c,1}, X_{c,2}, \ldots, X_{c,n_c})$ be independent samples from respective continuous distributions $F_1, F_2, \ldots, F_c$, where

(i) $F_i(x) = F((x - M_i)/\sigma_i)$, $i = 1, 2, \ldots, c$,
(ii) F is an arbitrary distribution with median zero and the standard deviations $\sigma_1, \sigma_2, \ldots, \sigma_c$ are possibly unequal.

Let $X_{i;\mathrm{med}}$ and S_i be, respectively, the median and standard deviation of the ith sample, $i = 1, \ldots, c$. For $1 \leqslant j < k \leqslant c$, let $W_{j,k}$ be the Mann–Whitney statistic defined by

$$W_{j,k} = \sum_{g=1}^{n_j} \sum_{h=1}^{n_k} I(X_{j,g} \leqslant X_{k,h})$$

and

$$U_{j,k} = \frac{W_{j,k}}{n_j n_k}.$$

As in Babu and Padmanabhan (2002), the samples are centered at their medians and then divided by their standard deviations to yield the location-cum-scale-aligned observations. Let

$$Z_{i,l} = (X_{i,l} - X_{i;\mathrm{med}})/S_i, \quad l = 1, \ldots, n_i; \quad i = 1, \ldots, c. \tag{1}$$

Let H_0: $M_1 = M_2 = \cdots = M_c$, H_{U}: $M_i \neq M_j$ for at least one pair (i, j) and H_{A}: $M_1 \leqslant M_2 \leqslant \cdots \leqslant M_c$ with at least one strict inequality.

For symmetric F, $p_{jk} = P(X_{j,1} \leqslant X_{k,1}) = 0.5$ under H_0. Utilizing this, Rust and Fligner (1984) proposed a modification of the Kruskal–Wallis statistic for testing H_0 against H_{U}. However, for skewed F, p_{jk} will be unknown and typically different from 0.5 even under H_0. Therefore, we begin with the following estimate $\tilde{p}_{jk}$ (of p_{jk}). Write for fixed j, k,

$$\begin{aligned} Q_{jk} &= n_j + n_k, \\ \xi_i &= Z_{j,i}, \quad i = 1, \ldots, n_j, \\ \xi_{n_j+i} &= Z_{k,i}, \quad i = 1, \ldots, n_k, \end{aligned} \tag{2}$$

where dependence of ξ's on j, k is suppressed for ease of notation.

Define

$$\tilde{p}_{jk} = \frac{1}{Q_{j,k}^2} \sum_{g=1}^{Q_{j,k}} \sum_{l=1}^{Q_{j,k}} I(\xi_g S_j \leqslant \xi_l S_k). \tag{3}$$

Let

$$T_{j,k} = \sqrt{n_k}(U_{j,k} - \tilde{p}_{jk}), \quad T_U = \sum_{1 \leqslant j < k \leqslant c} |T_{j,k}| \quad \text{and} \quad T_A = \sum_{1 \leqslant j < k < \leqslant c} T_{j,k}. \tag{4}$$

Let $Z^*_{1,1}, \ldots, Z^*_{1,m}, \ldots, Z^*_{c,1}, \ldots, Z^*_{c,n_c}$ be a bootstrap sample from the $Z_{i,l}$'s defined in (1). In estimating the quantiles of T_U and T_A, the following bootstrap procedures were studied.

Bootstrap I: Let $X^*_{j,g} = Z^*_{j,g} S_j$ and $X^*_{k,l} = Z^*_{k,l} S_k$, $g = 1, \ldots, n_j$, $l = 1, \ldots, n_k$,

$$U^*_{j,k} = \frac{1}{n_j n_k} \sum_{g=1}^{n_j} \sum_{l=1}^{n_k} I(X^*_{j,g} \leqslant X^*_{k,l}).$$

Next let S_j^* and S_k^* denote the standard deviations of the samples $X_j^* = (X_{j,1}^*, \ldots, X_{j,n_j}^*)$ and $X_k^* = (X_{k,1}^*, \ldots, X_{k,n_k}^*)$. Let $Q_{j,k}$ be as in (2),

$$\xi_i^* = Z_{j,i}^*, \quad i = 1, \ldots, n_j,$$

$$\xi_{n_j+i}^* = Z_{k,i}^*, \quad i = 1, \ldots, n_k,$$

$$p_{j,k}^* = \frac{1}{Q_{j,k}^2} \sum_{g=1}^{Q_{j,k}} \sum_{l=1}^{Q_{j,k}} I(\xi_g^* S_j^* \leqslant \xi_l^* S_k^*)$$

and $T_{j,k}^* = \sqrt{n_k}(U_{j,k}^* - p_{j,k}^*)$. The dependence of ξ^*'s on j, k is suppressed here for notational convenience.

Bootstrap II: Recall that $X_{j;\text{med}}$ and $X_{k;\text{med}}$ are the medians of the *j*th and *k*th samples, respectively. Now let $X_{j,g}^*$ and $X_{k,l}^*$ be as in Bootstrap I. Write

$$X_{j,g}^{**} = X_{j,g}^* + X_{j;\text{med}}, \quad g = 1, \ldots, n_j \quad \text{and} \quad X_{k,l}^{**} = X_{k,l}^* + X_{k;\text{med}}, \quad l = 1, \ldots, n_k.$$

Let $X_{j;\text{med}}^{**}$ and $X_{k;\text{med}}^{**}$ be the medians of the samples $X_j^{**} = X_{j,1}^{**}, \ldots, X_{j,n_j}^{**}$ and $X_k^{**} = X_{k,1}^{**}, \ldots, X_{k,n_k}^{*}$. Note that the standard deviations of X_j^{**} and X_k^{**} are simply S_j^* and S_k^* defined in Bootstrap I. Write

$$v_i^* = (X_{j,i}^{**} - X_{i;\text{med}}^{**})/S_j^*, \quad i = 1, \ldots, n_j,$$

$$v_{n_j+1}^* = (X_{k,l}^{**} - X_{k;\text{med}}^{**})/S_k^*, \quad l = 1, \ldots, n_k,$$

$$p_{j,k}^{**} = \frac{1}{Q_{j,k}^2} \sum_{r=1}^{Q_{j,k}} \sum_{t=1}^{Q_{j,k}} I(v_r^* S_j^* \leqslant v_t^* S_k^*),$$

and

$$T_{j,k}^{**} = \sqrt{n_k}(U_{j,k}^{**} - p_{j,k}^{**}),$$

again as before the dependence of v^*'s on j, k is suppressed. For T_U defined by (4), let T_U^* and T_U^{**} be its counterparts based on Bootstrap I and II, respectively. T_U and T_U^* have the same asymptotic null distribution. T_U and T_U^{**} have the same asymptotic distribution both under the null hypothesis and under contiguous alternatives. Similar results hold for T_A,T_A^* and T_A^{**}. The proofs follow from a multi-sample extension of the asymptotics for the two-sample case in Babu and Padmanabhan (2002). The details are omitted.

3. Monte Carlo simulations

Now to the finite sample behavior. As Bootstrap II turned out to be extremely conservative for small samples, we shall be giving the results for only Bootstrap I.

All the results are based on 2000 simulations, involving pseudo-random samples generated by **IMSL** subroutines, performed on an Alpha 1 computer at Monash University. Let N(μ, σ) denote the Normal distribution with mean μ and standard deviation σ.

The distributions studied were (standard) normal = N(0, 1), 25% 4N = 0.75N(0, 1) + 0.25N(0, 4) and the standard versions of the exponential and lognormal (centered at their medians). At nominal levels $\alpha = 0.025$ and 0.05, two-sided and one-sided tests were performed for general and ordered alternatives, respectively.

Let T denotes T_U or T_A. Both Bootstrap I and II were based on the following percentile method. From any given sample, 500 bootstrap samples were drawn.

Let $t_1^*, t_2^*, \ldots, t_{500}^*$ be the corresponding values of T and $t_{(1)}^* \leqslant t_{(2)}^* \leqslant \cdots \leqslant t_{(500)}^*$ be the corresponding ordered values. Then $t_{(12)}^*, t_{(25)}^*, t_{(475)}^*, t_{(488)}^*$ are, respectively, the bootstrap estimates of the null 2.5%, 5%, 95% and 97.5% quantiles of T.

Let $C_1 = (-\infty, t_{12}^*)$, $C_2 = (-\infty, t_{25}^*)$, $C_3 = (t_{488}^*, \infty)$ and $C_4 = (t_{476}, \infty)$. To find the empirical level, the procedure is repeated 2000 times. Left 2.5%, Left 5%, Right 2.5%, and Right 5% denote the percentages of times T falls in C_1,

Table 1
Empirical levels (unrestricted alternatives): $n_1 = n_2 = n_3 = 10$

	Normal	25% 4N	Exponential	Lognormal
Equal scales				
Right 2.5%	3.00	2.81	3.20	3.32
Right 5%	5.89	5.90	6.03	6.32
Unequal scales (1, 2, 4)				
Right 2.5%	3.45	2.94	3.36	3.41
Right 5%	6.00	6.13	6.22	6.50

Table 2
Empirical levels (unrestricted alternatives): $n_1 = n_2 = n_3 = 20$

	Normal	25% 4N	Exponential	Lognormal
Equal scales				
Right 2.5%	2.81	2.37	2.98	2.85
Right 5%	5.30	4.12	4.61	5.70
Unequal scales (1, 2, 4)				
Right 2.5%	2.95	2.60	3.00	3.02
Right 5%	5.20	4.82	5.63	5.82

Table 3
Empirical levels (ordered alternatives): $n_1 = n_2 = n_3 = 20$

	Normal	25% 4N	Exponential	Lognormal
Equal scales				
Left 2.5%	2.20	2.50	2.70	3.01
Left 5%	5.10	5.20	5.83	6.12
Right 2.5%	2.30	2.42	2.28	2.39
Right 5%	5.20	5.35	4.14	4.75
Unequal scales (1, 2, 4)				
Left 2.5%	2.50	2.90	2.89	3.16
Left 5%	5.40	5.50	5.72	6.35
Right 2.5%	2.80	2.93	2.22	2.55
Right 5%	5.24	5.66	4.92	5.14

C_2, C_3 and C_4 respectively. For $T = T_U$ the test is one-sided and only Right 2.5% and Right 5% are relevant. For $T = T_A$, the test is two-sided and hence Left 2.5%, Left 5%, Right 2.5% and Right 5% are all relevant.

Remark 3.1. In what follows, unequal scales (a, b, c) will mean that samples 1, 2 and 3 were drawn from distributions with scale parameters drawn from distributions with scale parameters a, b and c, respectively. Shift$(0, b, c)$ will mean that samples 2 and 3 have been shifted to the right by b and c, respectively. Positive pairing and negative pairing will correspond to unequal scales (1, 2, 4) and (4, 2, 1), respectively.

Tables 1–5 cover the balanced case, while Tables 6 and 7 cover the unbalanced case $n_1 = 10$, $n_2 = 20$ and $n_3 = 30$ for testing for unrestricted and ordered alternatives, respectively. As already mentioned, only the results of Bootstrap I are reported.

Remark 3.2. Due to restrictions on computer time, testing for ordered alternatives could not be undertaken for $n_1 = n_2 = n_3 = 10$. Simulations were run only for testing for unrestricted alternatives, in view of its relevance to the illustrations in Section 4.

Table 4
Empirical powers (unrestricted alternatives): Shift (0, 0.5, 1.0), $n_1 = n_2 = n_3 = 20$

	Normal	25% 4N	Exponential	Lognormal
Equal scales				
Right 2.5%	33.20	15.62	74.10	40.35
Right 5%	45.12	22.15	81.94	51.18
Unequal scales (1, 2, 4)				
Right 2.5%	21.40	10.80	62.40	29.80
Right 5%	29.60	16.20	72.18	40.56

Table 5
Empirical powers (ordered alternatives): Shift (0, 0.5, 1.0), $n_1 = n_2 = n_3 = 20$

	Normal	25% 4N	Exponential	Lognormal
Equal scales				
Left 2.5%	0.0	0.0	0.0	0.0
Left 5.0%	0.0	0.0	0.0	0.0
Right 2.5%	57.64	38.30	90.0	67.62
Right 5.0%	70.10	47.26	96.0	76.48
Unequal scales (1, 2, 4)				
Left 2.5%	0.0	0.0	0.0	0.0
Left 5.0%	0.0	0.0	0.0	0.0
Right 2.5%	41.0	26.20	70.0	51.72
Right 5.0%	58.0	34.10	82.32	62.83

Table 6
Empirical levels (unrestricted alternatives): $n_1 = 10$, $n_2 = 20$ and $n_3 = 30$

	Equal scales	Positive pairing	Negative pairing
(a) *Normal*			
Right 2.5%	2.48	3.02	3.10
Right 5.0%	5.0	6.10	6.14
25% 4N			
Right 2.5%	2.51	2.86	2.90
Right 5.0%	5.09	5.90	6.33
(b) *Exponential*			
Right 2.5%	3.0	3.17	3.48
Right 5%	5.80	6.0	6.21
Lognormal			
Right 2.5%	3.30	3.39	3.41
Right 5%	6.10	6.18	6.43

4. Illustrations

Two data sets one each from Physics and Psychology are selected for illustrations.
1. The following data:

$$X_1 = (87, 95, 98, 100, 109, 100, 81, 75, 68, 67),$$

$$X_2 = (78, 78, 78, 86, 87, 81, 73, 67, 75, 82, 83),$$

$$X_3 = (84, 86, 85, 82, 77, 76, 80, 83, 81, 78, 78, 78),$$

Table 7
Empirical levels (ordered alternatives): $n_1 = 10$, $n_2 = 20$ and $n_3 = 30$

	Equal scales	Positive pairing	Negative pairing
(a) *Normal*			
Left 2.5%	2.88	3.12	3.31
Left 5%	5.40	5.61	5.85
Right 2.5%	2.90	2.94	3.16
Right 5%	5.59	5.76	6.10
(b) *25% 4N*			
Left 2.5%	2.81	2.93	3.12
Left 5.0%	5.67	5.81	5.91
Right 2.5%	2.92	2.94	3.08
Right 5%	5.71	5.92	6.20
(c) *Exponential*			
Left 2.5%	2.70	2.82	2.92
Left 5.0%	5.10	5.26	5.96
Right 2.5%	2.66	2.97	3.54
Right 5%	5.58	5.87	6.22
(d) *Lognormal*			
Left 2.5%	2.69	2.90	3.25
Left 5.0%	5.13	5.63	5.87
Right 2.5%	2.91	3.12	3.70
Right 5%	5.88	5.93	6.50

taken from Cressie (1997, p. 46), pertain to the Heyl and Cook measurements of the Acceleration of Gravity, expressed as deviations from $980,060 \times 10^{-3}\,\mathrm{cm/s^2}$, in units of $10^{-3}\,\mathrm{cm/s^2}$.

The two-sample F–K:med test was proposed by Fligner and Killen (1976) for symmetric distributions. It was later generalized to cover both symmetric and skewed distributions in Hall and Padmanabhan (1997). As there is no danger of confusion, we shall call this generalized version also the F–K:med test. The hypothesis of homogeneity of scales was tested using a multi-sample extension of this test. The values of this statistic and its bootstrap quantiles were 168 and 103, resulting in the rejection of the null hypothesis of equal medians. Next, the value of T_U was 3.25, while its 95% bootstrap quantile was 3.49, leading to the acceptance of the null hypothesis of equal locations. Both these findings are consistent with those of Cressie (1997, p. 45).

2. The data,

$$X_1 = (0.998, 0.469, 0.53, 0.558, 0, 0, 0, 0, 0.282, 2.680),$$

$$X_2 = (0.250, 0, 0, 0.390, 0.348, 0, 0.207, 0.444, 0, 0.318),$$

$$X_3 = (0.250, 0, 0, 0, 0, 0.115, 0.795, 0.177, 0, 0.158, 0),$$

taken from Wilcox (1987) on the skin resistance of three groups arose in a study dealing with schizophrenia.

The values of the multi-sample F–K:med statistic and its 95% bootstrap quantiles were 116 and 94, respectively, thereby warranting the rejection of the null hypothesis of equal scales. Next T_U and its 95% bootstrap quantile were found to be 1.015 and 0.791, respectively, leading to the rejection of the null hypothesis of equal locations.

Acknowledgments

We thank Professor Rand Wilcox of the University of Southern California for the Psychology data sets. We thank the anonymous referee, whose constructive suggestions have helped in improving the paper.

References

Babu, G.J., Padmanabhan, A.R., 2002. Re-sampling methods for the non-parametric Behrens–Fisher problem. Sankhyā A 64, 678–692.
Chow, S.C., Liu, J.P., 1992. Design and Analysis of Bioavailability and Bioequivalence Studies. Marcel Dekker, New York.
Cressie, N., 1981. M-estimation in presence of unequal scales. Statist. Neerlandica 34, 19–32.
Cressie, N., 1997. Jackknifing in the presence of heteroscedasticity. Technometrics 39, 49–51.
Fligner, M.A., Killen, T.J., 1976. Distribution-free two-sample tests for scale. J. Amer. Statist. Assoc. 71, 210–213.
Fligner, M.A., Policello, G.E., 1981. Robust rank procedures for the Behrens–Fisher problem. J. Amer. Statist. Assoc. 76, 162–178.
Hall, P., Padmanabhan, A.R., 1997. Adaptive inference for the two-sample scale problem. Technometrics 39, 412–422.
Jonckheere, A.R., 1954. A distribution-free k-sample test against ordered alternatives. Biometrika 39, 133–145.
Micceri, T., 1989. The unicorn, the normal curve and other improbable creatures. Psychological Bull. 105, 156–166.
Rust, S., Fligner, M.A., 1984. A modification of the Kruskal–Wallis statistic for the generalized Behrens–Fisher problem. Comm. Statist. Theory Methods 13, 2013–2027.
Wilcox, R.R., 1987. New designs in analysis of variances. Annual Rev. Psych. 38, 29–60.
Wilcox, R.R., 1995. ANOVA: the practical importance of heteroscedastic methods, using trimmed means versus means, and designing simulation studies. British J. Math. Statist. Psych. 49, 163–170.