Journal of Statistics Applications & Probability
An International Journal

http://dx.doi.org/10.18576/jsap/050309

A Class of Nonparametric Tests for the Two-Sample Location Problem

Parameshwar V Pandit [1,*] *and Deepa R. Acharya* [2]

[1] Department of Statistics, Bangalore University, Bengaluru-560056, India
[2] Department of Statistics, Govt.Science College, Bengaluru-560001, India

Received: 11 July. 2016, Revised: 17 Aug. 2016, Accepted: 19 Aug. 2016
Published online: 1 Nov. 2016

Abstract: The two-sample location problem is one of the fundamental problems encountered in Statistics. In many applications of Statistics, two-sample problems arise in such a way as to lead naturally to the formulations of the null hypothesis to the effect that the two samples come from identical populations. A class of nonparametric test statistics is proposed for two-sample location problem based on U-statistic with the kernel depending on a constant 'a' when the underlying distribution is symmetric. The optimal choice of 'a' for different underlying distributions is determined. An alternative expression for the class of test statistics is established. Pitman asymptotic relative efficiencies indicate that the proposed class of test statistics does well in comparison with many of the test statistics available in the literature. The small sample performance is also studied through Monte-Carlo Simulation technique.

Keywords: Asymptotic relative efficiency, two-sample location problem, U-statistics Optimal test.

1 Introduction

Let $X_{11}, X_{12}, ..., X_{1n_1}$ and $X_{21}, X_{22}, ..., X_{2n_2}$ be two independent random samples from absolutely continuous distributions with c.d.f's $F(x)$ and $F(x-\Delta)$ respectively, where $F(x)+F(-x)=1$ for all $-\infty < x < \infty$. Here Δ is the location parameter. A popular nonparametric test for testing $H_0 : \Delta = 0$ versus $H_1 : \Delta \neq 0$ is the Wilcoxon-Mann-Whitney (W) [8] test. Besides, W-test, a number of distribution-free tests are available in the literature. Mathinsen [9] proposed a test for this problem based on the number of observations in X-sample not exceeding the median of Y-sample. Moods median (M) [10] test is particularly effective in detecting shift in location in distributions which are symmetric and heavy tailed. The Normal scores (NS)(refer Randles and Wolfe [12]) test, Gastwirth's L and H [3] tests and the RS test due to Hogg, Fisher and Randles [4]are effective in detecting shift in normal distribution, shifts in moderately heavy tailed distributions and shifts in skewed distributions respectively . The SG test proposed by Shetty and Govindarajulu [13] takes care of two suspected outliers at the extremes of both the samples. Deshpande and Kochar [2], Stephenson and Ghosh [15] Shetty and Bhat [14] are few other test procedures for this problem among others. The generalization of the test due to Deshpande and Kochar [2] is considered by Kumar, Singh and Ozturk [6]. Ahmad [1] proposed a generalization of Mann-Whitney test for this problem based on subsample extremes. Recently Pandit and Savitha kumari [11] proposed a class of tests for two sample location problem based on subsample quantiles. In this paper, we propose a class of distribution-free tests which are effective in detecting the shift in distributions that are symmetric.

The class of test statistics is proposed in section 2. An alternative expression for the proposed class is also given in section 2. Section 3 contains the distributional properties of the proposed class of test statistics. Section 4 is devoted to study the performance of the proposed class of tests in terms of Pitman asymptotic relative efficiencies(ARE) and empirical power. Section 5 contains some remarks and conclusions.

* Corresponding author e-mail: panditpv12@gmail.com

2 The proposed class of statistics

We propose a test based on the following U-statistic which is given by

$$U_a = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} h(x_{1i}, x_{2j}) \; ; n = n_1 + n_2$$

where

$$h(X_{1i}, X_{2j}) = \begin{cases} 1 & if \quad min(X_{1i}, X_{2j}) > 0 \\ a(-a) & if \quad X_{1i}, X_{2j} < 0 \quad and \quad X_{1i} + X_{2j} > 0(<)0 \\ -1 & if \quad max(X_{1i}, X_{2j}) < 0 \\ 0 & otherwise \end{cases}$$

The test based on U_a rejects $H_0 : \Delta = 0$ against $H_1 : \Delta \neq 0$ when $|U_a|$ is too large. The test is distribution-free for all n, with null distribution depending on the choice of $'a'$.

Alternative expression for U_a

Let $m = \sum_{i=1}^{2} \sum_{j=1}^{n_i} I[X_{ij} > 0]$ and $l = \sum_{i=1}^{2} \sum_{j=1}^{n_i} I[X_{ij} < 0]$ so that $n = m + l$ and let $n^* = m - 1$.

Further,

$$W^+ = \sum_{i=1}^{2} \sum_{j=1}^{n_i} R_{ij}^+ I[X_{ij} > 0]$$
$$\text{and } W^- = \sum_{i=1}^{2} \sum_{j=1}^{n_i} R_{ij}^+ I[X_{ij} < 0]$$

where R_{ij}^+ is the the rank of X_{ij} in $|X_{11}|, |X_{12}|, ..., |X_{1n_1}|$ and $|X_{21}|, |X_{22}|, ..., |X_{2n_2}|$ and set $W = W^+ - W^-$. Note that $W^+ + W^- = \frac{n(n+1)}{2}$.
Similarly , we can set

$$U_a^{\pm} = \sum_{k=1}^{n_1} \sum_{j=1}^{n_2} h_{\pm}(x_{1k}, x_{2j})$$

where

$$h_+(X_{1k}, X_{2j}) = \begin{cases} 1 & if \quad min(X_{1k}, X_{2j}) > 0 \\ a & if \quad X_{1k}, X_{2j} < 0 \quad and \quad X_{1k} + X_{2j} > 0 \\ 0 & otherwise \end{cases}$$

and $h_-(X_{1k}, X_{2j}) = h_+(X_{1k}, X_{2j}) - h(X_{1k}, X_{2j})$

Here $W^+ (W^-)$ represent the signed-rank statistic corresponding to the number $m(l)$ of positive (negative) X_{kj} 's. Then, we can establish the following relation between $U_a' = n_1 n_2 U_a$, W^+ *and* n^* *as*

$$U_a^{+'} = aW^+ + \binom{m+1}{2}(1-a)$$
$$\text{and } U_a^{-'} = aW^- + \binom{l+1}{2}(1-a)$$

so that

$$U_a' = U_a^{+'} - U_a^{-'}$$
$$= aW + \frac{1}{2} n^*(n+1)(1-a) \tag{1}$$

Exact null distribution

The exact null distribution of U_a can be enumerated in general by simply noting that , under H_0 , each combination of signed ranks $\pm 1, \pm 2, ..., \pm n$ yielding a value of $U_a^{'}$ has probability $\frac{1}{2^n}$, the total number of such combinations being 2^n. The c.d.f. of $U_a^{'}$ can then be conveniently written using (1) for any given or predetermined value of 'a'. Thus the range of values of $U_a^{'}$ is random and depends on the selected value of 'a'.

3 Distributional properties of U_a

The mean of U_a is given by

$$\begin{aligned}\mu(\Delta) &= E(U_a)\\ &= P[Min(X_{1k}, X_{2j}) > 0] + aP[X_{1k}, X_{2j} < 0, X_{1k} + X_{2j} > 0]\\ &- aP[X_{1k}, X_{2j} < 0, X_{1k} + X_{2j} < 0] - P[Max(X_{1k}, X_{2j}) < 0]\\ &= A_1 + aA_2 - A_3\end{aligned}$$

where

$$A_1 = \frac{1 - F(-\Delta)}{2}$$

$$A_2 = \int_{-\infty}^{0} [1 - 2F(-x-\Delta)]dF(x) + \int_{-\infty}^{-\Delta} [1 - 2F(-x-\Delta)]dF(x) + F(-\Delta)$$

$$A_3 = \frac{F(-\Delta)}{2}.$$

Under H_0 , $E[U_a] = 0$ and $Var(U_a) = \frac{1}{n_1 n_2} \sum_{c=0}^{l} \sum_{d=0}^{l} \binom{n_1 - l}{l - c} \binom{n_2 - l}{l - d} \zeta_{c,d}$

where $\zeta_{0,0} = 0,\ \zeta_{1,0} = \zeta_{0,1} = 1 + \frac{a^2}{3}$ *and* $\zeta_{1,1} = \frac{1}{2}(1 + a^2)$

Since U_a is a U-statistic, its asymptotic distribution of $\sqrt{n}U_a$, under H_0 is $N(0, \sigma_a^2)$ where $\sigma_a^2 = \frac{\zeta_{1,0}}{\lambda} + \frac{\zeta_{0,1}}{1-\lambda} = \frac{1}{4}(1 + \frac{a^2}{3})$, which is the direct consequence of Lehmann [7].

4 Asymptotic Relative Efficiency and optimal value of 'a'

The asymptotic relative efficiency of U_a with respect to two-sample t-test, T is given by

$$ARE(U_a, T) = \frac{4}{1 + \frac{a^2}{3}} \left[(1-a) f(0) + 2a \int_{-\infty}^{\infty} f^2(x)dx \right]^2,$$

assuming $\sigma^2 = Var(F)$ is one. The optimal value a^* of $'a'$ is obtained by solving $\frac{d}{da} ARE(U_a, T) = 0$ and verifying $\frac{d^2}{da^2} ARE(U_a, T) < 0$ for the solution. The value of $'a'$ thus obtained is $a^* = \frac{6}{f(0)} \int_{-\infty}^{\infty} f^2(x)dx - 3$. Hence, the ARE of 'optimal' statistic U_a^* is

$$\begin{aligned}ARE(U_{a^*}, T) &= 4\left[f^2(0) + 12\left\{ \int_{-\infty}^{\infty} f^2(x)dx - \frac{1}{2} f(0) \right\}^2 \right]\\ &\geq 12 \left\{ \int_{-\infty}^{\infty} f^2(x)dx \right\}^2 \qquad (2)\end{aligned}$$

The asymptotic relative efficiency of the proposed test with respect to Wilcoxon's(W), Mood's median test(M), Gastwirth L and H tests [3], Normal Scores(NS)test(refer Randles and wolfe [12]), Hogg, Fisher and Randles(RS)test [4], Shetty and Govindarajulu(SG) [13] test, Shetty and Bhat [14] test $T(b,d)$, Deshpande and Kochar [2] test $L(c,d)$ and two-sample test T are given in the following tables 1-3.

Table 1: Asymptotic relative efficiency of U_{a^*} with respect to T, W, $T(1,3)$, $T(1,5)$, $T(2,3)$, $T(2,5)$

		Asymptotic relative efficiency of U_{a^*} relative to					
Distribution	a^*	T	W	$T(1,3)$	$T(1,5)$	$T(2,3)$	$T(2,5)$
Cauchy	0	0.4052	1.1323	1.1430	1.0623	1.1833	1.0865
Laplace	0	2.0000	1.3333	1.2432	1.1998	1.2872	1.2270
Logistic	1	1.0966	1.0000	1.0013	1.0288	1.0475	1.0525
Normal	$3(\sqrt{2}-1)$	0.9643	1.0098	1.0645	1.1048	1.1047	1.1323
Triangular	1	0.8889	1.0000	1.0833	1.2005	1.1266	1.1582
Uniform	3	1.3333	1.0000	1.4571	1.7921	1.5085	1.7400

Table 2: Asymptotic relative efficiency of U_{a^*} relative to RS, M, H, L, NS, SG

	Asymptotic relative efficiency of U_{a^*} relative to					
Distribution	RS	M	H	L	NS	SG
Cauchy	1.6656	0.9996	0.9953	5.0502	1.8834	1.6023
Laplace	1.6664	0.994	1.1842	2.6658	1.5740	1.1998
Logistic	1.2374	1.3199	1.0479	1.2720	1.0326	1.0184
Normal	1.2613	1.5132	1.1608	1.0891	.09642	1.1045
Triangular	1.2500	1.334	1.1965	1.0000	0.7883	1.1325
Uniform	1.2505	3.0045	2.0007	0.5002	∞	1.7013

Table 3: Asymptotic relative efficiency of U_{a^*} with respect to $L(c,d)$

Distribution	$d=1$	$d=2$	$d=3$
Laplace	1.5238	1.7143	1.9730
Logistic	1.1428	1.2857	1.3987
Normal	1.1539	1.2982	1.3745
Uniform	1.5237	1.8357	1.5107

Empirical Powers

Monte Carlo simulation is carried out for finding the empirical powers of our test statistic U_{a^*} for three distributions namely, Normal, Uniform and Cauchy when $n_1 = n_2 (= 8)$ and $\alpha(= 0.01, 0.05, 0.10)$. Empirical power is the proportion of 10000 trials for which the test based on U_{a^*} rejects $H_0 : \Delta = 0$
versus $H_1 : \Delta > 0$. In table 4 and 5, the empirical powers of U_{a^*} are presented.

Table 4: Empirical powers of $U_{a^*} n_1 = n_2 (= 8)$

	Normal Distribution			Cauchy Distribution		
$\Delta \downarrow \alpha \rightarrow$	0.01	0.05	0.10	0.01	0.05	0.10
1	0.0466	0.1763	0.2582	0.0418	0.1562	0.2311
2	0.1297	0.6438	0.7008	0.0999	0.3119	0.4265
4	0.2830	0.6812	0.7578	0.2087	0.5351	0.6633
5	0.3259	0.7071	0.8297	0.2550	0.6008	0.7307
6	0.3619	0.7699	0.8724	0.2906	0.6482	0.7707
8	0.4165	0.8391	0.9149	0.3272	0.7173	0.8239
10	0.4428	0.8661	0.9379	0.3753	0.7621	0.8681

Table 5: Empirical powers of $U_{a^*} n_1 = n_2 (= 8)$ for Uniform Distribution

$\Delta \downarrow \alpha \rightarrow$	0.01	0.05	0.10
0.1	0.0195	0.1005	0.1594
0.2	0.0622	0.2327	0.3428
0.3	0.1405	0.0.4263	0.5583
0.4	0.2320	0.6021	0.7530
0.5	0.3235	0.7598	0.8778

5 Remarks and Conclusions

1. The class of tests proposed in this paper, U_{a^*} is consistent for testing $H_0 : \Delta = 0$ against $H_1 : \Delta > 0$.
2. U_{a^*} is more efficient than RS, M, H, L, NS, $T(b,d)$ and SG tests for light and medium tailed distributions.
3. The test based on U_{a^*} is better than $L(c,d)$ for $c = 1$ for all symmetric distributions.
4. The gain in efficiency of U_{a^*} with respect to W test is more for heavy tailed distributions. However, the gain is moderate for medium and light tailed distributions.

Acknowledgment

The second author would like to thank University Grants Commission for its support under FDP scheme.Also the authors are grateful to the anonymous referee for a careful checking of the details and for helpful comments that improved this paper.

References

[1] Ahmad, I. A. A class of Mann-Whitney- Wilcoxon type statistics. *The American Statistician*, Vol.50, No.4, 324–327,(1996).
[2] Deshpande, J. V. and Kochar, S. C. Some Competitors of Wilcoxon-mann-Whitney Test for the Location Alternatives,*Journal of Indian Statist. Assoc.*, 19, 9–18, (1982).
[3] Gastwirth, J. L. Percentile modi?cations of two sample rank tests,*Journal of Amer. Statist. Assoc.*, 60, 1127–1141,(1965).
[4] Hogg,R.V., Fisher,D.M., and Randles,R.H. A two-sample adaptive distribution-free test. *Journal of Amer. Statist. Assoc.*70, 656-61,(1975).
[5] Kumar, N. A class of two-sample tests for location based on sub-sample medians. *Communications in Statistics (Theory and Methods)* , 26 (4), 943 –951, (1997).
[6] Kumar, N.,Singh, R. S.and Ozturk O. A New Class of Distribution-Free Tests for Location Parameters. *Communications in Statistics (Theory and Methods)* , 22 (1 and 2), 107–128, (2003).
[7] Lehmann, E. L. Consistency and unbiasedness of certain nonparametric tests, *Ann.Math.Statist.*, 22, 165–179,(1951).
[8] Mann, H. B. and Whitney, D. R. On a test of whether one of two random variables is stochastically larger than other,*Ann. Math. Statist.*, 18, 50–60, (1947).
[9] Mathisen, H. C. A method of testing the hypothesis that two-samples are from the same population, *Ann. Math. Statist.*, 14, 188–194,(1943).
[10] Mood, A. M. On asymptotic efficiency of certain nonparametric two-sample tests,*Ann. Math. Statist.*, 25, 514–533,(1954).

[11] Pandit, P. V., Savitha Kumari and Javali, S. B. Tests for Two-Sample Location Problem Based on Subsample Quantiles,*Open Journal of Statistics* , 4, 70–74, (2014).
[12] Randles, R. H. and Wolfe, D. A. Introduction to the Theory of Nonparametric Statistics, *John Wiley and Sons*, *New York*, (1979).
[13] Shetty, I. D. and Z. Govindarajulu A two-sample test for location, *Comm. Statist- Theory and Meth.*17, 2389–2401,(1988).
[14] Shetty, I. D. and Bhat , S. V. A note on the generalization of Mathisen ′s median test, *Statistics and Probability letters*, 19, 199–204,(1994).
[15] Stephenson, W. R. and Ghosh , M. Two-sample nonparametric tests based on subsamples, *Commun. Statist-Theor, Meth.*, 14, No.1, 1669–1684,(1985).

Parameshwar V. Pandit received the PhD degree in Statistics from Karnatak University ,Dharwad. His research interests are in the areas of Statistics including Parametric and Nonparametric Inference,Inference on Reliability,Survival analysis,Nonparametric Process Control. He has published more than sixty research articles in the journals of international repute in the area of Statistics and applied sciences. He is serving as regional editor,technical editor of statistical journals and reviewed articles for more than twenty five international journals.

Deepa R. Acharya is Assistant Professor of Statistics at Government Science College,Bengaluru. She recieved M.Phil.degree in Statistics from Karnatak University , Dharwad. Her area of research includes Nonparametric inference and published research articles in reputed international journals of statistics.