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 $\Delta$ 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: panditp12@gmail.com
2 The proposed class of statistics

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

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

where

\[ h(X_{1i}, X_{2j}) = \begin{cases} 
1 & \text{if } \min(X_{1i}, X_{2j}) > 0 \\
-1 & \text{if } \max(X_{1i}, X_{2j}) < 0 \\
0 & \text{otherwise}
\end{cases} \]

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 \( \alpha' \).

**Alternative expression for \( U_a \)**

Let \( m = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I[X_{ij} > 0] \) and \( l = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I[X_{ij} < 0] \) so that \( n = m + l \) and let \( n^* = m - 1 \).

Further,

\[ W^+ = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} R_{ij}^+ I[X_{ij} > 0] \]

and \( W^- = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} R_{ij}^- I[X_{ij} < 0] \)

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

Similarly, we can set

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

where

\[ h_+(X_{1k}, X_{2j}) = \begin{cases} 
1 & \text{if } \min(X_{1k}, X_{2j}) > 0 \\
0 & \text{otherwise}
\end{cases} \]

\[ h_-(X_{1k}, X_{2j}) = h_+(X_{1k}, X_{2j}) - h(X_{1k}, X_{2j}) \]

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 \( U_a = n_1 n_2 U_a \) and \( n^* \) as

\[ U_a^{\alpha'} = aW^+ + \left(\frac{m^*+1}{2}\right)(1-a) \]

and \( U_a^{\beta'} = aW^- + \left(\frac{l^*+1}{2}\right)(1-a) \)

so that

\[ U_a^{\alpha'} = U_a^{\alpha'} - U_a^{\beta'} = aW + \frac{1}{2} n^*(n+1)(1-a) \] (1)
3 Distributional properties of $U_a$

The mean of $U_a$ is given by

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

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 $\text{Var}(U_a) = \frac{1}{n \sigma^2} \sum_{c=0}^{l} \sum_{d=0}^{l} \binom{n-1-c}{l-c} \binom{n-1-d}{l-d} \zeta_{c,d}$

where $\zeta_{0,0} = 0$, $\zeta_{0,1} = 1 + a^2$ and $\zeta_{1,1} = \frac{1}{4}(1 + a^2)$

Since $U_a$ is a U-statistic, its asymptotic distribution of $\sqrt{n}U_a$, under $H_0$ is $N(0, \sigma_n^2)$ where

$$
\sigma_n^2 = \frac{\zeta_{0,0}}{\lambda^2} + \frac{\zeta_{0,1}}{1 - \lambda^2} = \frac{1}{4}(1 + \frac{a^2}{1^2}),
$$

which is the direct consequence of Lehmann(1951).

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

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

assuming $\sigma^2 = \text{Var}(F)$ is one. The optimal value $a^*$ of 'a' is obtained by solving $\frac{d}{da} \text{ARE}(U_a, T) = 0$ and verifying $\frac{d^2}{da^2} \text{ARE}(U_a, T) < 0$ 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

$$
\text{ARE}(U_a^*, T) = 4 \left[ f^2(0) + 12 \left( \int_{-\infty}^{\infty} f^2(x)dx - \frac{1}{2} f(0) \right) \right]^2 \geq 12 \left( \int_{-\infty}^{\infty} f^2(x)dx \right)^2 \tag{2}
$$
The asymptotic relative efficiency of the proposed test with respect to Wilcoxon’s (W), Mood’s median test (M), Gastwirth L and H tests (1965), Normal Scores (NS) test (refer Randles and Wolfe 1979), Hogg, Fisher and Randles (RS) test (1975), Shetty and Govindarajulu (SG) (1988) test, Shetty and Bhat (1994) test $T(b,d)$, Deshpande and Kochar (1982) 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 $T, W, T(1,3), T(1,5), T(2,3), T(2,5)$

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$a^*$</th>
<th>$T$</th>
<th>$W$</th>
<th>$T(1,3)$</th>
<th>$T(1,5)$</th>
<th>$T(2,3)$</th>
<th>$T(2,5)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cauchy</td>
<td>0</td>
<td>0.4052</td>
<td>1.1323</td>
<td>1.1430</td>
<td>1.0623</td>
<td>1.1833</td>
<td>1.0865</td>
</tr>
<tr>
<td>Laplace</td>
<td>0</td>
<td>2.0000</td>
<td>1.3333</td>
<td>1.2432</td>
<td>1.1998</td>
<td>1.2872</td>
<td>1.2270</td>
</tr>
<tr>
<td>Logistic</td>
<td>1</td>
<td>1.0966</td>
<td>1.0000</td>
<td>1.0013</td>
<td>1.0288</td>
<td>1.0475</td>
<td>1.0525</td>
</tr>
<tr>
<td>Normal</td>
<td>$3(\sqrt{2}-1)$</td>
<td>0.9643</td>
<td>1.0098</td>
<td>1.0645</td>
<td>1.1048</td>
<td>1.1047</td>
<td>1.1323</td>
</tr>
<tr>
<td>Triangular</td>
<td>1</td>
<td>0.8889</td>
<td>1.0000</td>
<td>1.0833</td>
<td>1.2005</td>
<td>1.1266</td>
<td>1.1582</td>
</tr>
<tr>
<td>Uniform</td>
<td>3</td>
<td>1.3333</td>
<td>1.0000</td>
<td>1.4571</td>
<td>1.7921</td>
<td>1.5085</td>
<td>1.7400</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$RS$</th>
<th>$M$</th>
<th>$H$</th>
<th>$L$</th>
<th>$NS$</th>
<th>$SG$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cauchy</td>
<td>1.6656</td>
<td>0.9996</td>
<td>0.9953</td>
<td>5.0502</td>
<td>1.8834</td>
<td>1.6023</td>
</tr>
<tr>
<td>Laplace</td>
<td>1.6664</td>
<td>0.9944</td>
<td>1.1842</td>
<td>2.6658</td>
<td>1.5740</td>
<td>1.1998</td>
</tr>
<tr>
<td>Logistic</td>
<td>1.2374</td>
<td>1.3199</td>
<td>1.0479</td>
<td>1.2720</td>
<td>1.0326</td>
<td>1.0184</td>
</tr>
<tr>
<td>Normal</td>
<td>1.2613</td>
<td>1.5132</td>
<td>1.1608</td>
<td>1.0891</td>
<td>0.9642</td>
<td>1.1045</td>
</tr>
<tr>
<td>Triangular</td>
<td>1.2500</td>
<td>1.3340</td>
<td>1.1965</td>
<td>1.0000</td>
<td>0.7883</td>
<td>1.1325</td>
</tr>
<tr>
<td>Uniform</td>
<td>1.2505</td>
<td>3.0045</td>
<td>2.0007</td>
<td>0.5002</td>
<td>$\infty$</td>
<td>1.7013</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$d = 1$</th>
<th>$d = 2$</th>
<th>$d = 3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laplace</td>
<td>1.5238</td>
<td>1.7143</td>
<td>1.9730</td>
</tr>
<tr>
<td>Logistic</td>
<td>1.1428</td>
<td>1.2857</td>
<td>1.3987</td>
</tr>
<tr>
<td>Normal</td>
<td>1.1539</td>
<td>1.2982</td>
<td>1.3745</td>
</tr>
<tr>
<td>Uniform</td>
<td>1.5237</td>
<td>1.8357</td>
<td>1.5107</td>
</tr>
</tbody>
</table>

**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 $U_\alpha \ast n_1 = n_2 (= 8)$

<table>
<thead>
<tr>
<th>$\Delta \downarrow \alpha \rightarrow$</th>
<th>Normal Distribution</th>
<th>Cauchy Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0466 0.1763 0.2582</td>
<td>0.0418 0.1562 0.2311</td>
</tr>
<tr>
<td>2</td>
<td>0.1297 0.6438 0.7008</td>
<td>0.0999 0.3119 0.4265</td>
</tr>
<tr>
<td>4</td>
<td>0.2830 0.6812 0.7578</td>
<td>0.2087 0.5351 0.6633</td>
</tr>
<tr>
<td>5</td>
<td>0.3259 0.7071 0.8297</td>
<td>0.2550 0.6008 0.7307</td>
</tr>
<tr>
<td>6</td>
<td>0.3619 0.7699 0.8724</td>
<td>0.2906 0.6482 0.7707</td>
</tr>
<tr>
<td>8</td>
<td>0.4165 0.8391 0.9379</td>
<td>0.3753 0.7621 0.8681</td>
</tr>
</tbody>
</table>

Table 5: Empirical powers of $U_\alpha \ast n_1 = n_2 (= 8)$ for Uniform Distribution

<table>
<thead>
<tr>
<th>$\Delta \downarrow \alpha \rightarrow$</th>
<th>0.01</th>
<th>0.05</th>
<th>0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.0195</td>
<td>0.1005</td>
<td>0.1594</td>
</tr>
<tr>
<td>0.2</td>
<td>0.0622</td>
<td>0.2327</td>
<td>0.3428</td>
</tr>
<tr>
<td>0.3</td>
<td>0.1405</td>
<td>0.4263</td>
<td>0.5583</td>
</tr>
<tr>
<td>0.4</td>
<td>0.2320</td>
<td>0.6021</td>
<td>0.7530</td>
</tr>
<tr>
<td>0.5</td>
<td>0.3235</td>
<td>0.7598</td>
<td>0.8778</td>
</tr>
</tbody>
</table>

5 Remarks and Conclusions

1. The class of tests proposed in this paper, $U_\alpha$ is consistent for testing $H_0 : \Delta = 0$ against $H_1 : \Delta > 0$.
2. $U_\alpha$ 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 $U_\alpha$ is better than $L(c,d)$ for $c = 1$ for all symmetric distributions.
4. The gain in efficiency of $U_\alpha$ 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


**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 received 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.