Distribution-Free Tests for Two-Sample Location Problems Based on Subsamples

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

[1] Department of Statistics, Govt. Science College, Bangalore -560001
[2] Department of Statistics, Bangalore University, Bangalore-560056
Email: panditpv12@gmail.com

Abstract. Nonparametric tests for location problems have received much attention in the literature. Many nonparametric tests have been proposed for one, two and several samples location problems. In this paper a class of test statistics is proposed for two sample location problem when the underlying distributions of the samples are symmetric. The class of test statistics proposed is linear combination of U-statistics whose kernel is based on subsamples extrema. The members of the new class are shown to be asymptotically normal. The performance of the proposed class of tests is evaluated using Pitman Asymptotic Relative Efficiency. It is observed that the members of the proposed class of tests are better than the existing tests in the literature.

Keywords: Asymptotic relative efficiency, two-sample location problems, Nonparametric Tests, Symmetric distributions, U-statistics.

1 Introduction

Let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ be two independent random samples from two populations with continuous distribution functions $F(x)$ and $G(y) = F(x-\theta)$ respectively. Consider two sample location problem of testing $H_0 : G(x) = F(x)$ for all x against the alternative $H_1 : G(x) = F(x-\theta)$, with $\theta \neq 0$ for some unknown continuous distribution function F and a real (shift) parameter θ, $-\infty < \theta < \infty$. In the above testing problem, we may also consider one sided say, right sided alternative $H_1 : \theta > 0$. There are many nonparametric tests available in the literature for the above problem. Mann-Whitney's test [7] is a popular nonparametric procedure for this problem. Mood's median (M) test [3] is effective in detecting shift in location in populations whose distributions are symmetric and heavy tailed. Gastwirth's H and L tests [2] are effective in detecting shifts in moderately heavy tailed distributions. The Normal Scores (NS) test [3] is effective in detecting a shift in the normal distribution. The RS test due to Hogg, Fisher and Randles [4] is effective in detecting shifts in distributions that are skewed. The SG test proposed by Shetty and Govindarajulu [10] based on subsample medians takes care of two suspected outliers at the extremes of both the samples. A generalization of Mathisen [8] is considered by Shetty and Bhat [11]. Their relative efficiency and suitability depends on the nature of the (unknown) underlying distribution F. Ahmad [1] proposed a generalization of Mann-Whitney test for this problem.

Section 2 contains the new proposed class of tests for two sample location problem. The distribution theory of the class of test statistics is presented in Section 3. The asymptotic relative efficiency comparisons are discussed in Section 4. In section 5 some remarks and conclusions are presented.

2 Proposed Class of Test Statistics

Let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ be two independent random samples from two populations with continuous distribution functions $F(x)$ and $G(y) = F(x-\theta)$ respectively. Consider two sample location problem of testing $H_o : F(x) = G(x)$ for all x against the alternative $H_1 : G(x) = F(x-\theta)$ with $\theta > 0$ where $-\infty < \theta < \infty$.

A test procedure is proposed based on the statistic which is given by

$$\begin{aligned} U_{mn} &= U_2 - U_1 \\ &= \frac{1}{\binom{n}{k+3}} \sum_C h\left(Y_{i_1}, Y_{i_2}, \ldots, Y_{i_{(k+3)}}\right) - \frac{1}{\binom{m}{k+3}} \sum_D h\left(X_{i_1}, X_{i_2}, \ldots, X_{i_{(k+3)}}\right) \end{aligned}$$

where $h\left(X_1, X_2, \ldots, X_{(k+3)}\right) = \frac{1}{\binom{k+3}{3}} \sum_A I\left(\mathrm{MIN}(X_{j1}, X_{j2}, X_{j3}) > -\mathrm{MAX}(X_{j4}, \ldots, X_{j(k+3)})\right)$ and C is the set of $\binom{n}{k+3}$ combination of integers $\{1,2,\ldots,n\}$, D is $\binom{m}{k+3}$ combinations of integers $\{1,2,\ldots,m\}$ and A is the set of $\binom{k+3}{3}$ combinations of integers $\{j_1, j_2, \ldots, j_{(k+3)}\}$.

The test criterion is to reject H_0 in favour of H_1 for large values of $U_{m,n}$. That is, reject H_0 if $U_{m,n} > c$. If the alternative is two sided, that is, $H_1 : \theta \neq 0$, then reject H_0 when $U_{m,n}$ is too small or $U_{m,n}$ is too large.

3 Distributional Properties of $U_{m,n}$

The mean of $U_{m,n}$ is given by

$$\begin{aligned} \gamma(F,G) &= E\left(U_{m,n}\right) \\ &= \gamma(G) - \gamma(F) \\ &= 0, \qquad \text{under } H_0 \end{aligned}$$

The following theorem gives the asymptotic distribution of $U_{m,n}$.

Theorem 1: *Under H_0, the limiting distribution of $\sqrt{N}\left(U_{m,n} - \gamma(F,G)\right)$, $N = m+n$, as $n \to \infty$ is normal with mean zero and variance $\sigma^2 = \frac{\sigma_0^2}{\lambda(1-\lambda)}$, where $\lambda = \lim_{N\to\infty} \frac{m}{N}$.*

Proof: The proof of the theorem follows from Lehmann [5] by noting that $U_{m,n}$ is a two sample U-statistics and the asymptotic variance of $\sqrt{N}\left(U_{m,n} - \gamma(F,G)\right)$ is $\sigma^2 = \frac{\sigma_0^2}{\lambda(1-\lambda)}$, where $\lambda = \lim_{N\to\infty} \frac{m}{N}$, and under H_0,

$$\begin{aligned}\sigma_0^2 &= \left(k+3\right)^2 \xi_{10} \\ &= Cov\left(h\left(X_1, X_2, \ldots, X_{k+3}; Y_1, Y_2, \ldots, Y_{k+3}\right), h\left(X_1, X_{k+4}, \ldots, X_{2k+5}; Y_1, Y_{k+4}, \ldots, Y_{2k+5}\right)\right) \\ &= \left(k+3\right)^2 \int_{-\infty}^{\infty} \left(\binom{k+2}{2} A + \binom{k+2}{3} B\right)^2 dF\left(x\right) - \gamma^2\left(F, G\right)\end{aligned}$$

with $A = \frac{k}{k+2}\left[1 - \bar{F}^{k+2}\left(x\right)\right]$ and $B = \frac{k-1}{k+2} + \frac{3}{k+2} F^{k+2}\left(x\right)$.

In the following table 1, we tabulate asymptotic variance of $\sqrt{N}\left(U_{m,n}\right)$ under H_0 for different values of k.

Table 1. Asymptotic null variance of $\sqrt{N}\left(U_{m,n}\right)$

k	2	3	4	5	6	7	8
Variance	0.492857	0.586753	0.614718	0.612066	0.595542	0.573087	0.548568

4 Asymptotic Relative Efficiency (ARE)

In this section we first obtain the Pitman asymptotic relative efficiency of $U_{m,n}$, with respect to the classical t-test. For this we compute the efficacy of $U_{m,n}$, given by

$$\begin{aligned}eff\left(U_{m,n}\right) &= \left[\lim_{n\to\infty} \frac{\gamma'\left(F, G\right)}{\sqrt{N} Var_o\left(U_{m,n}\right)}\right]^2 \\ &= \frac{\left[6k\int_{-\infty}^{\infty} F^{k+1}\left(y\right) f^2\left(y\right) dy\right]^2}{\left(k+3\right)^2 \xi_0}\end{aligned}$$

The asymptotic relative efficiencies of the proposed test $U_{m,n}$ with respect to Wilcoxon's test(W), Fisher and Randles (RS) test, Mood's Median test (M-test), Gastwirth H and L tests, Normal Scores(NS) test, Shetty and Govindarajulu test (SG), Shetty and Bhat (T(1,3), T(1,5), T(2,3), T(2,5)) tests are evaluated. Table 2 to Table 12 give the AREs eff($U_{m,n}$) for various values of k.

Table 2. ARE's of $U_{m,n}$ relative to Wilcoxon's $\left(W\right)$ test

$ARE\left(U_{m,n}, W\right)$	k =2	k =3	k =4	k =5	k =6
Laplace	1.5808	1.463733	1.334533	1.211067	1.100467
Logistic	0.973919	0.939084	0.896407	0.850903	0.805946
Uniform	1.5217	1.8406	2.169	2.5007	2.8335
Triangular	2.161998	2.214985	2.248509	2.269659	2.283271

Table 3. ARE 's of $U_{m,n}$ relative to Fisher and Randles $\left(RS\right)$ test

$ARE\left(U_{m,n}, RS\right)$	k =2	k =3	k =4	k =5	k =6
Laplace	1.975723	1.829373	1.667879	1.513622	1.375313
Logistic	1.217881	1.174381	1.120994	1.064024	1.007795
Uniform	1.902801	2.301607	2.712255	3.127023	3.543165
Triangular	2.702311	2.767437	2.810509	2.836927	2.853926

Table 4. ARE's of $U_{m,n}$ relative to Mood's Median $\left(M-test\right)$ test

$ARE\left(U_{m,n},M\right)$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$
Laplace	1.184962	1.097187	1.000329	0.907812	0.82486
Logistic	1.299086	1.252685	1.195739	1.13497	1.074992
Uniform	4.571866	5.530078	6.516745	7.513309	8.513173
Triangular	2.882528	2.951997	2.997942	3.026121	3.044254

Table 5. ARE's of $U_{m,n}$ relative to Gastwirth $\left(H\right)$ test

$ARE\left(U_{m,n},H\right)$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$
Laplace	1.404028	1.300026	1.185262	1.075641	0.977353
Logistic	1.031424	0.994584	0.949371	0.901123	0.853502
Uniform	3.044437	3.682517	4.339545	5.003164	5.668981
Triangular	2.586558	2.648894	2.690122	2.715408	2.731679

Table 6. ARE's of $U_{m,n}$ relative to Gastwirth $\left(L\right)$ test

$ARE\left(U_{m,n},L\right)$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$
Laplace	3.160685	2.92656	2.668208	2.421434	2.200173
Logistic	1.251979	1.207261	1.15238	1.093814	1.036011
Uniform	0.761165	0.920696	1.084965	1.250882	1.417348
Triangular	2.161896	2.213998	2.248456	2.269591	2.283191

Table 7. ARE's of $U_{m,n}$ relative to Normal Scores $\left(NS\right)$ test

$ARE\left(U_{m,n},NS\right)$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$
Laplace	1.86619	1.727954	1.575413	1.429707	1.299067
Logistic	1.016346	0.980044	0.935492	0.887949	0.841025
Uniform	∞	∞	∞	∞	∞
Triangular	1.704061	1.745128	1.77229	1.788949	1.799668

Table 8. ARE's of $U_{m,n}$ relative to Shetty and Govindarajulu $\left(SG\right)$ test

$ARE\left(U_{m,n},SG\right)$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$
Laplace	1.422456	1.317089	1.200818	1.089758	0.99018
Logistic	1.002352	0.96655	0.922611	0.875723	0.829445
Uniform	2.588803	3.131386	3.690083	4.254384	4.820553
Triangular	2.448455	2.507463	2.546489	2.570425	2.585827

Table 9. ARE's of $U_{m,n}$ relative to Shetty and Bhat $\left(T\left(2,3\right)\right)$ test

$ARE\left(U_{m,n},T\left(2,3\right)\right)$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$
Laplace	1.526163	1.413143	1.288408	1.169209	1.062432
Logistic	1.020238	0.983747	0.939039	0.891371	0.844276
Uniform	2.29552	2.776588	3.271987	3.772364	4.2744
Triangular	2.43578	2.495476	2.533246	2.557074	2.57241

Table 10. ARE 's of $U_{m,n}$ relative to Shetty and Bhat $\left(T\left(1,3\right)\right)$ test

$ARE\left(U_{m,n},T\left(1,3\right)\right)$	k =2	k =3	k =4	k =5	k =6
Laplace	1.4742	1.365	1.2445	1.1294	1.0262
Logistic	0.9855	0.9503	0.9071	0.861	0.8155
Uniform	2.2172	2.6819	3.1604	3.6437	4.1286
Triangular	2.3527	2.4094	2.4469	2.4699	2.4847

Table 11. ARE 's of $U_{m,n}$ relative to Shetty and Bhat ($T\left(1,5\right)$)? test

$ARE\left(U_{m,n},T\left(1,5\right)\right)$	k =2	k =3	k =4	k =5	k =6
Laplace	1.422478	1.317136	1.200876	1.089775	0.990252
Logistic	1.001975	0.966136	0.92223	0.875414	0.829162
Uniform	2.727061	3.298566	3.887097	4.481541	5.077957
Triangular	2.595436	2.659045	2.699291	2.724681	2.741022

Table 12. ARE 's of $U_{m,n}$ relative to Shetty and Bhat ($T\left(2,5\right)$) test

$ARE\left(U_{m,n},T\left(2,5\right)\right)$	k =2	k =3	k =4	k =5	k =6
Laplace	1.454813	1.347077	1.228174	1.114547	1.012762
Logistic	1.02507	0.988406	0.943487	0.895593	0.848275
Uniform	2.647816	3.202714	3.774143	4.351314	4.930398
Triangular	2.504051	2.565421	2.60425	2.628746	2.644512

5 Some Remarks and Conclusions

1. A class of test statistics for two-sample location problem is considered in the paper assuming that the underlying distribution of the sample drawn is symmetric.
2. The asymptotic variance of the few members, $U_{m,n}$ (for k =2, 3, 4, 5, 6) of the class of test statistics are computed as a ready reference.
3. The performance of the members of the proposed class is evaluated in terms of asymptotic relative efficiencies (AREs).
4. From table 2 to table 12, it is observed that the performance of the proposed test is better than the tests existing in the literature for this problem if the distributions of the samples drawn are Laplace, Logistic, Triangular, or Uniform.
5. For heavy tailed distributions such as Laplace and logistic distributions, the performance in terms of ARE decreases with k (that is with subsample size).
6. For light tailed distributions such as Triangular and Uniform the performance in terms of ARE increases with k (that is with subsample size).

References

1. Ahmad, I.A.(1996). A class of Mann-Whitney-Wilcoxon type statistics. The American Statistician, Vol.50, No.4, 324-327.
2. Gastwirth, J. L.(1965). Percentile modifications of two sample rank tests, Journal of Amer. Statist. Assoc, 0, 1127-1141.
3. Hajek,J. and Sidak, Z.(1967).Theory of rank tests, Academic Press, New York.
4. Hogg,R.V., Fisher,D.M., and Randles,R.H.(1975). A two-sample adaptive distribution-free test. Journal of Amer. Statist. Assoc.70, 656-61.

5. Lehmann,E.L.(1951).Consistency and unbiasedness of certain nonparametric tests, Ann. Math. Statist., 22,165-179.
6. Lehmann, E.L.and D' Abrera, H.J.M. (2006). Nonparametrics: Statistical Methods based on Ranks, Sringer-Verlag, New York.
7. Mann,H.B. and Whitney,D.R.(1947).On a test of whether one of two random variables is stochastically larger than other, Ann.Math.Statist.,18,50-60.
8. Mathisen, H.C.(1943). A method of testing the hypothesis that two-samples are from the same population, Ann. Math. Statist.14, 188-194.
9. Randles, R. H. and Wolfe, D.A. (1979). Introduction to the Theory of Nonparametric Statistics, John Wiley and Sons, New York.
10. Shetty, I.D. and Z. Govindarajulu(1988). A two-sample test for location, Comm. Statist-Theory and Meth.17, 2389-2401.
11. Shetty, I.D. and Bhat, S.V.(1994). A note on the generalization of Mathisen's median test, Statistics and Probability letters, 19,199-204.

