Open Journal of Statistics, 2014, 4, 70-74
Published Online February 2014 (http://www.scirp.org/journal/ojs)
http://dx.doi.org/10.4236/ojs.2014.41007

Scientific Research

Tests for Two-Sample Location Problem Based on Subsample Quantiles

Parameshwar V. Pandit[1], Savitha Kumari[2], S. B. Javali[3]
[1]Department of Statistics, Bangalore University, Bangalor, India
[2]Department of Statistics, SDM College, Ujire, India
[3]Department of Public Health Dentistry, SDM College of Dental Sciences and Hospital, Dharwad, India
Email: panditpv12@gmail.com, savi_rrao@yahoo.co.in, javalimanju@rediffmail.com

Received November 6, 2013; revised December 6, 2013; accepted December 13, 2013

ABSTRACT

This paper presents a new class of test procedures for two-sample location problem based on subsample quantiles. The class includes Mann-Whitney test as a special case. The asymptotic normality of the class of tests proposed is established. The asymptotic relative performance of the proposed class of test with respect to the optimal member of Xie and Priebe (2000) is studied in terms of Pitman efficiency for various underlying distributions.

KEYWORDS

U-Statistic; Class of Tests; Two-Sample Location Problem; Asymptotic Normality; Pitman ARE; Subsample Quantiles

1. Introduction

Two-sample location problem is one of the extensively studied problems in the literature. There are many non-parametric tests available in literature for the above problem, their relative efficiency and suitability depending on the nature of the (unknown) underlying distribution F. For this problem, for example there is a whole class of locality asymptotically most powerful (distribution free) linear rank tests for each specified distribution F, which included the well known Mann-Whitney, normal score and median tests among many others ([1], Ch-III, 1.1). While the median test is particularly effective for heavy tailed symmetric distributions, the normal score and Mann-Whitney tests are relatively even handed and reasonably effective for moderately heavy tailed bell shaped distributions. During the last decade or so, new classes of tests based on the so called subsample approach have been proposed for the above problem, notable among them being Deshpande and Kochar [2], Stephenson and Gosh [3], Shetty and Govindarajulu [6], Shetty and Bhat [7] and Ahmad [8]. While Shetty and Govindarajulu [6] and Shetty and Bhat [7] based their tests on subsample medians which tend to emphasize the centre of the underlying distributions, the other two are based on statistics involving subsample extrema with the object of gaining more information from the tails of sampled distributions. The results of these papers demonstrate that the subsample approach, applied selectively, does help to improve upon the efficiency performance of the tests mentioned in the preceding paragraph in an overall sense. For example, in consonance with the efficiency results noted in the last paragraph, Shetty and Govindarajulu [6] test performs on one hand better than the Mann-Whitney test for heavy-tailed distributions, while performing better than the median test for light-tailed distributions on the other. Deshpande and Kochar [2] test, on the other hand, being sensitive to light-tailed distributions, performs substantially better than Mann-Whitney test for such underlying distributions and some what better for normal, while maintaining reasonable level of efficiency under heavy-tailed distributions. Stephenson and Ghosh [3] and Ahmad [8] tests are also relatively more sensitive than the Mann-Whitney test but less than the Deshpande and Kochar [2] test to the light-tailed distributions. Recently, Xie and Priebe [9] proposed a genera-

lization of Mann-Whitney test which includes many existing tests as its members. They studied the asymptotic relative efficiencies of the members of the proposed class of tests and investigated the optimum member (having maximum efficacy) in the class of tests.

Let $X_1, X_2, \cdots, X_m$ and $Y_1, Y_2, \cdots, Y_n$ be two independent random samples of i.i.d. observations from two populations with continuous distribution functions (d.f's), from $F(x)$ and $G(y) = F(x-\Delta)$ respectively, and consider the standard two-sample (non parametric) location problem of testing $H_0 : G = F$ versus the alternatives $H_1 : G(x) = F(x-\Delta)$ with $\Delta \neq 0$, for some unknown continuous d.f. F and a real (shift) parameter $\Delta, -\infty < \Delta < \infty$. In the above testing problem we may also consider one sided say, right sided alternatives $H_a = \Delta > 0$. In this paper, we proposed a class of tests for two-sample location problem based subsample quantiles and investigated the properties such as asymptotic normality, asymptotic efficiencies etc.

2. Materials and Methods

The study here is intended to propose new class of tests for two-sample location problem and study the properties of the test statistics. The performance of few members of the class of test statistics is also studied.

The Proposed Class of Tests

Let $X_1, X_2, \cdots, X_m$ and $Y_1, Y_2, \cdots, Y_n$ be two independent random samples of i.i.d. observations from two populations with continuous distribution functions (d.f's), from $F(x)$ and $G(y) = F(x-\Delta)$ respectively. Consider the problem, of testing is to test the null hypothesis $H_0 : \Delta = 0$ and $H_1 : \Delta > 0$. Let c and d be any fixed positive integers such that $1 \le c \le m$ and $1 \le d \le n$ with $N = m + n$ and consider all possible subsample of sizes c and d from $X_1, X_2, \cdots, X_m$ and $Y_1, Y_2, \cdots, Y_n$ respectively. Based on the subsample, define kernel

$$\phi_{(c,d)}\left(X_1, X_2, \cdots, X_c; Y_1, Y_2, \cdots, Y_d\right) = \frac{1}{c}\sum_{i=1}^{c} I\left[X_i \ge Y_{(r)d}\right],$$

where $Y_{(r)d} = r^{th}$ smallest among $Y_1, Y_2, \cdots, Y_d$ and $\phi_1(A) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$.

Then, the *U*-statistics based on h_1 is

$$V_{(c,d)}\left(X_1, X_2, \cdots, X_m, Y_1, Y_2, \cdots, Y_n\right) = \frac{1}{\binom{m}{c}\binom{n}{d}}\sum_{C}\phi\left[X_{i1}, X_{i2}, \cdots, X_{im}, Y_{j1}; Y_{j2}, \cdots, Y_{jn}\right]$$

where C is the set of all possible integers $\{i_1, i_2, \cdots, i_c\}$ taken out of the set $\{1, 2, \cdots, m\}$ without replacement and all possible integers $\{j_1, j_2, \cdots, j_d\}$ taken out of the set $\{1, 2, \cdots, n\}$ without replacement.

Asymptotic Distribution of the class of tests: Next we consider the asymptotic distribution of the test statistic $V(c,d)$. For that we consider the expectation of $V(c,d)$ is

$$E\left(V(c,d)\right) = P\left[X_i \ge Y_{(r)d}\right] = \int_{-\infty}^{\infty} P\left[Y_{(r)d} \le x\right] \mathrm{d}F(x) = \int_{-\infty}^{\infty}\sum_{j=r}^{d}\binom{d}{j} G^j(x)\left(1-G(x)\right)^{d-j}\,\mathrm{d}F(x)$$

Under H_0, $E\left(V(c,d)\right) = \dfrac{d-r+1}{d+1}$.

From the generalized U-statistics theorem due to Lehman (1951), it follows that the limiting distribution of $\sqrt{N}\left(V(c,d) - E\left(V(c,d)\right)\right)$, as $N \to \infty$ in such a way that $\dfrac{m}{N} \to \lambda, 0 < \lambda < 1$, is normal with mean zero and variance

$$\sigma_{r,s}^2 = \frac{c^2\zeta_{10}}{\lambda} + \frac{d^2\zeta_{01}}{(1-\lambda)}$$

where

$$\zeta_{10} = \mathrm{Cov}\left[\phi\left(X_1, X_2, \cdots, X_c; Y_1, Y_2, \cdots, Y_d\right), \phi\left(X_1, X_{c+1}, \cdots, X_{2c-1}; Y_{d+1}, \cdots, Y_{2d}\right)\right] = E\left[\phi_{10}^2(X)\right] - \left[\frac{d-r+1}{d+1}\right]^2$$

and

$$\zeta_{10} = \mathrm{Cov}\left[\phi\left(X_1, X_2, \cdots, X_c; Y_1, Y_2, \cdots, Y_d\right), \phi\left(X_{c+1}, \cdots, X_{2c}; Y_1, Y_{d+1}, \cdots, Y_{2d-1}\right)\right] = E\left[\phi_{01}^2\left(X\right)\right] - \left[\frac{d-r+1}{d+1}\right]^2$$

Now,

$$\phi_{10}\left(x\right) = \frac{1}{c}\left[P\left(x > Y_{(r)d}\right) + \left(c-1\right)P\left(x_2 > Y_{(r)d}\right)\right] = \frac{1}{c}\left[\sum_{j=r}^{d}\binom{d}{j}F^j\left(x\right)\left(1-F\left(x\right)\right)^{d-j} + \frac{\left(c-1\right)\left(d-r+1\right)}{d+1}\right]$$

and

$$\phi_{01}\left(x\right) = \frac{1}{c}\sum_{i=1}^{c}E\left[I\left(X_i \geq Y_{(r)d}\right)\middle|Y_1 = y\right] = \frac{1}{c}\sum_{i=1}^{c}P\left(X_i \geq Y_{(r)d}\middle|Y_1 = y\right) = \frac{1}{c}\sum_{i=1}^{c}\left(A_1 + A_2 + A_3\right)$$

where

$$A_1 = P_{H_0}\left[y \leq Y_{(r-1)(d-1)} \leq X_1\right],$$
$$A_2 = P_{H_0}\left[Y_{(r-1)(d-1)} \leq y \leq X_1\right],$$

and

$$A_3 = P_{H_0}\left[Y_{(r)(d-1)} \leq X_1 \leq y\right].$$

3. Results and Discussions

Pitman defined the asymptotic relative efficiency of one test *S* relative to another test *T* as the limiting ratio of sample sizes required to obtain the same limiting power for a sequence of alternatives converging to null hypothesis. By Noether's theorem it follows that

$$ARE\left(S,T\right) = \left[\frac{Eff\left(S\right)}{Eff\left(T\right)}\right]^2,$$

where

$$Eff\left(S\right) = \lim_{N\to\infty}\left[\frac{\mathrm{d}/\mathrm{d}\theta E\left(S\right)\big|_{\theta=\theta_0}}{\sqrt{N\mathrm{Var}_{H0}\left(S\right)}}\right].$$

For asymptotic relative efficiency comparisons, two members of the class, $V_{c,d}\left(r\right)$, namely, $V_{c,d}\left(d\right)$ and $V_{c,d}\left(1\right)$ are considered.

The mean of $V_r\left(c,d\right)$ under $H_1, \mu_{c,d}\left(\Delta\right)$ is given by

$$\mu_{c,d}\left(\Delta\right) = \int_{-\infty}^{\infty}F^d\left(x-\Delta\right)\mathrm{d}F\left(x\right)$$

Then,

$$\mu'_{c,d}\left(0\right) = \left\{\frac{\mathrm{d}}{\mathrm{d}\Delta}\mu_{c,d}\left(\Delta\right)\right\}_{\Delta=0} = -d\int_{-\infty}^{\infty}F^{d-1}\left(x\right)f^2\left(x\right)\mathrm{d}x$$

The ARE's of $V_{c,d}\left(d\right)$ and $V_{c,d}\left(1\right)$ relative to the test due to a member of Xie and Priebe [9], which has maximum efficiency are computed for various distributions, namely, Cauchy, Laplace, Logistic, Normal, Uniform and Exponential distributions.

We present ARE's of $V_{c,d}\left(d\right)$ and $V_{c,d}\left(1\right)$ for various values of *c*, *d* in the **Tables 1** and **2** respectively with respect to GMWW of Xie and Priebe [9].

4. Conclusions

From the results and discussion section, we conclude as below:

Table 1. **The ARE's $V_{c,d}(d)$ of various distributions w. r. to GMWW, when $c = d$.**

c	d	Distributions					
		Cauchy	Laplace	Logistic	Uniform	Exponential	Normal
2	2	3.4025	0.4373	2.5333	1.9680	1.6903	1.3915
3	3	5.3408	0.3574	3.0326	2.8758	2.3932	2.9298
4	4	6.9373	0.2901	3.4145	3.7471	3.0984	3.8906
5	5	7.9791	0.2384	3.7279	4.6012	3.8044	4.7998
6	6	8.5871	0.1997	3.9957	5.4482	4.5109	5.6871
7	7	8.9080	0.1703	4.2274	6.5201	5.2175	6.5546
8	8	9.0130	0.1479	4.4345	7.6281	5.9243	7.4049
9	9	8.9804	0.1302	4.6187	8.7690	6.6311	8.2411
10	10	14.7584	0.1162	4.7873	9.9401	7.3380	9.0641

Table 2. **The ARE's $V_{c,d}(d)$ of various distributions w. r. to GMWW, when $c = d$.**

c	d	Distributions					
		Cauchy	Laplace	Logistic	Normal	Uniform	Exponential
2	2	0.1791	0.0230	1.4723	0.8885	0.8885	0.0376
3	3	0.0822	0.0055	1.1407	1.0647	1.0258	0.0098
4	4	0.0459	0.0019	0.9488	1.2318	1.1426	0.0040
5	5	0.0276	0.0008	0.8175	1.3552	1.2454	0.0020
6	6	0.0175	0.0004	0.7205	1.4596	1.3396	0.0012
7	7	0.0116	0.0002	0.6445	1.5493	1.5330	0.0007
8	8	0.0079	0.0001	0.5837	1.6276	1.7272	0.0005
8	9	0.0071	0.0001	0.5948	2.0269	2.1510	0.0005
9	9	0.0034	0.0001	0.5331	1.6974	1.9218	0.0003
10	10	0.0041	0.0001	0.4910	1.7600	2.1167	0.0002

1) From the above investigation, we observe that for Cauchy, Logistic, uniform distribution and exponential distribution, the performance of the newly proposed test is better than the GMWW-test when $2 \le c = d \le 10$.

2) The performance of the new test is better as compared to GMWW-test for Normal distribution, Power distribution ($a = 2$) and Power distribution ($a = 3$) when $c = d$ and $c = 2, 3, \ldots, 10$.

3) We also observed that for Normal distribution and uniform distribution, the performance of the newly proposed test $V_{c,d}(1)$ increases with d for fixed c and is better than the GMWW-test for $c = 2$ when $3 \le d \le 10$.

4) However, the performance of new test is better than GMWW-test for Normal and uniform distribution $c \le d \le 10$, and $c = 3, 4, \ldots, 10$.

5) For Laplace distribution, the performance of the newly proposed test $V_{c,d}(1)$ is better than the GMWW-test when $c \le d \le 10$ and $c = 3, 4, 5$. However, for Cauchy distribution, Laplace distribution and exponential distribution the newly proposed class of tests is not better than the GMWW-test for all values of c and d.

Acknowledgements

Authors thank the editor and the reviewers for their useful comments. The first author thank the UGC for its support through a grant F.No.41-807/2012(SR).

REFERENCES

[1] J. Hajek and Z. Sidak, "Theory of Rank Tests," Academic Press, New York, 1967.

[2] J. V. Deshpande and S. C. Kochar, "Some Competitors of Wilcoxon-Mann Whitney Test for the Location Alternatives," *Journal of Indian Statistical Association*, Vol. 19, 1982, pp. 9-18.

[3] W. R. Stephenson and M. Gosh, "Two Sample Non-Parametric Tests Based on Subsamples," *Communications in Statistics—Theory and Methods*, Vol. 14, No. 7, 1985, pp. 1669-1684. http://dx.doi.org/10.1080/03610928508829003

[4] E. L. Lehmann, "Consistency and Unbiasedness of Certain Nonparametric Tests," *Annals of Mathematical Statistics*, Vol. 22, No. 2, 1951, pp. 165-179. http://dx.doi.org/10.1214/aoms/1177729639

[5] H. B. Mann and D. R. Whitney, "On a Test of Whether One of Two Random Variables Is Stochastically Larger than the Other," *Annals of Mathematical Statistics*, Vol. 18, No. 1, 1947, pp. 50-60. http://dx.doi.org/10.1214/aoms/1177730491

[6] I. D. Shetty and Z. Govindarajulu, "A Two-Sample Test for Location," *Communications in Statistics—Theory and Methods*, Vol. 17, No. 7, 1988, pp. 2389-2401. http://dx.doi.org/10.1080/03610928808829752

[7] I. D. Shetty and S. V. Bhat, "A Note on the Generalization of Mathisen's Median Test," *Statistics & Probability Letters*, Vol. 19, No. 3, 1994, pp. 199-204. http://dx.doi.org/10.1016/0167-7152(94)90105-8

[8] I. A. Ahmad, "A Class of Mann-Whitney-Wilcoxon Type Statistics," *The American Statistician*, Vol. 50, No. 4, 1996, pp. 324-327.

[9] J. Xie and C. E. Priere, "Generalizing the Mann-Whitney Wilcoxon Statistic," *Journal of Non-Parametric Statistics*, Vol. 12, No. 5, 2000, pp. 661-682. http://dx.doi.org/10.1080/10485250008832827