International J. of Math. Sci. & Engg. Appls. (IJMSEA)
ISSN 0973-9424, Vol. 10 No. III (December, 2016), pp. 151-158

ON DISTRIBUTION FREE TESTS FOR THE TWO-SAMPLE LOCATION PROBLEM BASED ON SIGNS OF EXTREME OBSERVATIONS

PARAMESHWAR V. PANDIT[1] **AND SAVITHA KUMARI**[2]
[1] Department of Statistics, Bangalore University,
Bangalore -560056, India
[2] Department of Statistics, SDM College, Ujire-574240, India

Abstract

A class of distribution-free tests for two-sample location problem is based on the signs of most extreme observations in the sub-samples of sizes c and d from X and Y samples respectively. The test statistics have been expressed in terms of linear rank statistics. The asymptotic normality of the test statistics is established. Asymptotic efficiencies indicate that members of our class do well in comparison with some already existing test statistics for light and medium tailed distributions.

1. Introduction

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 formulation of null hypothesis to the effect that the two samples come from identical populations. There are many non-parametric tests available in

---

Key Words : *Two-sample location problem, Linear rank statistics sub-samples, Pitman efficiency.*

AMS Subject Classification : 62G10.

literature for the two sample location problem, their relative efficiency and suitability depending on the nature of the (unknown) underlying distribution. Wilcoxon-Mann-Whitney W-test is a popular nonparametric test for this problem. Besides W-test a number of distribution-free tests are available in the literature. Mathisen [4] proposed a test for this problem based on the number of observations in the X-sample not exceeding the median of Y-sample. Moods median test (M) is particularly effective in detecting shift in location in the normal distribution. Gastwirths [3] L and H tests are effective in detecting shifts in moderately heavy tailed distributions. The RS test due to Hogg, Fisher and Randles [5] is effective in detecting shifts in distributions that are skewed. 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 Ghosh [12], Shetty and Govindarajulu [10] and Shetty and Bhat [11] and Ahmed [1]. While Shetty and Govindarajulu [10] and Shetty and Bhat [12] based their tests on sub sample medians which tend to emphasize the centre of the underlying distributions, the other two are based on statistics involving sub sample extreme with the object of gaining more information from the tails of sampled distributions. The results of these papers demonstrate that the sub sample approach, applied selectively, does help to improve upon the efficiency performance of the tests in an overall sense. For example, Shetty and Govindarajulu [10] 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 [12] and Ahmed [1] tests are also relatively more sensitive than the Mann-Whitney test but less sensitive than the Deshpande and Kochar [2] test to the light tailed distributions.

In this paper, we propose a new class of the distribution-free test statistics which is the convex combination of two U-statistics. Among the U-statistics involved in the combination, one has the kernel based on subsample maxima and the other has the kernel based on sub sample minima. The distributional properties of the proposed class is studied and the asymptotic relative efficiencies of few members of the class

are investigated relative to few other statistics exist in the literature, particularly the optimal member of the class proposed by Xie and Priebe [13].

The SG test proposed by Shetty and Govindarajulu [10] based on subsample medians takes care of two suspected outliners at extremes of both the samples. Deshpande and Kochar [2] test is effective in detecting shift in distributions that are light tailed. Stephenson and Ghosh [12], $U(c,d)$ and Shetty and Bhat [11] $T(c,d)$ tests are few other test procedures for this problem. In this section, we propose a class of distribution free tests which are effective in detecting the shifts in distributions that are symmetric, medium and light tailed. The test statistics is proposed in section 2. An alternative expression for the class of test statistics is given in section 3. The distributional properties of the proposed class of tests are presented in section 4. Section 5 is devoted to the study of Pitman asymptotic relative efficiency and section 6 for some comments.

2. Methods

2.1. The Proposed Class of Tests

Suppose $X_1, X_2, \cdots, X_m$ and $Y_1, Y_2, \cdots, Y_n$ are independent random samples form continuous distribution with c.d.f.'s $F(x)$ and $F(x-\theta)$ respectively. We wish to test $H_0 : \theta = 0$ against $H_1 : \theta > 0$ with $F(x) + F(-x) = 1$.

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

$$V_{(c,d)}(X_1, X_2, \cdots, X_m, Y_1, Y_2, \cdots, Y_n) = \frac{1}{\binom{m}{c}\binom{n}{d}} \sum_A h[X_{i1}, X_{i2}, \cdots, X_{ic}; Y_{j1}, Y_{j2}, \cdots, Y_{jd}]$$

where A denotes the sum over all $\binom{m}{c}\binom{n}{d}$ combinations of X and Y sample observations and $h(x_1, \cdots, x_c; y_1, \cdots, y_d) = h_2(y_1, \cdots, y_d) - h_1(x_1, \cdots, x_c)$.

Here

$$h_1(x_1, \cdots, x_c) = \begin{cases} 1 & \text{if the largest in absolute value among} \\ & (x_1, \cdots, x_c) \text{ is positive} \\ 0 & \text{otherwise} \end{cases}$$

and

$$h_2(y_1, \cdots, y_d) = \begin{cases} 1 & \text{if the largest in absolute value among} \\ & (y_1, \cdots, y_d) \text{ is positive} \\ \\ 0 & \text{otherwise} \end{cases}$$

2.2. An Alternative Expression for $V(c,d)$

The computational effort associated with the evaluation of $V(c,d)$ is enormous when the sample sizes are large. However, it is possible to derive an alternative expression for $V(c,d)$ as linear rank statistics.

Suppose $X_{|1|}, \cdots, X_{|m|}$ are the observations arranged in the order of increasing absolute value. Consider distinct sub-samples of size c for which $X_{\setminus|k|}$ is the largest in absolute value. For $k \geq c$, we have $(c-1)$ places in the sub samples that can be filled with $(k-1)$ objects. $X_{|1|}, \cdots, X_{|k-1|}$, each of which may appear at most once. Therefore, there are $\begin{pmatrix} k-1 \\ c-1 \end{pmatrix} = b_c(k)$, distinct sub-samples for which

$$h_1(x_{i1}, x_{i2}, \cdots, x_{ic}) = \begin{cases} 1, & \text{if } X_{|k|} > 0 \\ \\ 0, & \text{otherwise} \end{cases}$$

The sum of $(x_{i1}, x_{i2}, \cdots, x_{ic})$ over all distinct sub-samples of size $c, 1 \leq i_1 \leq i_2 \cdots \leq i_c \leq m$, is equivalently given by

$$V_1 = \sum_{k=c}^{m} b_c(k)\gamma_k, \quad \text{where} \quad \gamma_k = \begin{cases} 1, & \text{if } X_{|k|} > 0 \\ \\ 0, & \text{otherwise.} \end{cases}$$

On similar lines, we can prove that the sum of $h_2(y_{j1}, y_{j2}, \cdots, y_{jd})$ over all distinct sub-samples of size $d, 1 \leq j_1 \leq j_2 \cdots \leq j_d \leq n$, is equivalently given by

$$V_2 = \sum_{k=d}^{n} b_d(k)\xi_k, \quad \text{where} \quad \xi_k = \begin{cases} 1, & \text{if } Y_{|k|} > 0 \\ \\ 0, & \text{otherwise.} \end{cases}$$

Hence $Y_{|1|}, \cdots, Y_{|n|}$ are the observations arranged in increasing absolute value. Then the statistic $V(c,d)$ can be written as

$$\begin{aligned} \begin{pmatrix} m \\ c \end{pmatrix} \begin{pmatrix} n \\ d \end{pmatrix} V(c,d) &= \begin{pmatrix} m \\ c \end{pmatrix} V_2 - \begin{pmatrix} n \\ d \end{pmatrix} V_1 \\ &= \begin{pmatrix} m \\ c \end{pmatrix} \sum_{k=d}^{n} b_d(k)\xi_k - \begin{pmatrix} n \\ d \end{pmatrix} \sum_{k-c}^{m} b_c(k)\gamma_k. \end{aligned}$$

3. Results

3.1. Distributional Properties of $V(c,d)$

The mean of $V(c,d)$ is given by

$$\mu(\theta) = E[V(c,d)] = \int_{-\theta}^{\infty} [F(y) - F(-y-\theta)]^{d-1} f(y)dy - \frac{1}{2} > 0.$$

Under H_0, $E[V(c,d)] = 0$.

However, under H_1

$$E[V(c,d)] = \int_{-\theta}^{\infty} [F(y) - F(-y-\theta)]^{d-1} f(y)dy - \frac{1}{2} > 0, \quad \text{for} \quad d \geq 2.$$

Here, one can notice that $V(c,d)$ is distribution free under H_0. Since $V(c,d)$ is a U-statistic with a square integrable kernel, the asymptotic normality of $V(c,d)$ follows from Lehmann [6]. Under H_0, $\sqrt{N}\, V(c,d)$ is asymptotically normal with mean zero and variance given by

$$\sigma^2_{c,d} = \frac{c^2\zeta_{10}}{\lambda} + \frac{d^2\zeta_{01}}{1-\lambda},$$

where

$$\zeta_{10} = \frac{1}{4(2c-1)}, \quad \zeta_{01} = \frac{1}{4(2d-1)} \quad \text{and} \quad \lambda = \lim_{N\to\infty} \frac{m}{N}, \; N = c + d.$$

3.2 Asymptotic Relative Efficiency

For the sequence of Pitman alternatives $\theta_N = \theta\sqrt{N}$ the efficacy of $V(c,d)$ is given by

$$e^2_v(d) = 4d^2(d-1)^2 \left\{ \int_0^{\infty} [2F(y) - 1]^{d-2} f^2(y)dy \right\}^2 / \sigma^2_{c,d}, \quad d \geq 2.$$

When sub sample sizes are equal i.e, $c = d = r$, the efficacy of $V(r)$ is given by

$$e^2_v(r) = 16(r-1)^2(2r-1) \left\{ \int_0^{\infty} [2F(y) - 1]^{r-2} f^2(y)dy \right\}^2 / \lambda(1-\lambda), \quad r \geq 2.$$

Table 1, lists asymptotic relative efficiency of $V(r)$ with respect to the two sample t test T for various continuous distributions for some values of r. Table 2, gives the ARE's of $V(r)$ with respect to Mann Whitney test M. The ARE's of $V(c,d)$ with respect to Deshpande and Kochar [2] Test $L(c,d)$ for $c = 1$ and $d = 2, 3, 4$ for equal sample sizes are given in Table 3. In Table 4, the ARE's of $V(c,d)$ with respect to Stephenson and Ghosh [12] test $U(c,d)$ are presented.

Table 1: Asymptotic Relative Efficiency of $V(r)$ relative to T

Density	$r=2$	$r=3$	$r=4$	Comment
Cauchy	0.3040	0.1792	0.1093	Max at $r=1$
Laplace	1.5000	1.1111	0.8750	Max at $r=1$
Logestic	1.0966	1.0281	0.9212	Max at $r=2$
Normal	0.9549	0.9774	0.9386	Max at $r=3$
Triangular	0.8889	0.9481	0.9752	Increasing in r
Parabolic	0.8640	1.0635	1.1836	Increasing in r
Uniform	1.0000	1.6667	2.3333	Increasing in r
Inv. Triangular	2.6667	6.4000	10.2857	Increasing in r

Table 2 : Asymptotic Relative Efficiency of $V(r)$ relative to M

Density	$r=2$	$r=3$	$r=4$
Cauchy	0.1014	0.0597	0.0363
Laplace	1.0000	0.7407	0.5833
Logestic	1.0000	0.9375	0.8401
Normal	1.0000	1.0236	0.9829
Triangular	1.0000	1.0667	1.0971
Parabolic	1.9812	2.4387	2.7141
Uniform	4.6182	5.7731	8.0821
Inv. Triangular	5.6558	13.5737	21.8148

Table 3 : Asymptotic Relative Efficiency of $V(1,d)$ relative to $L(1,d)$

Density	$d=2$	$d=3$	$d=4$
Laplace	1.1429	0.9524	0.8632
Logestic	1.1428	1.2054	1.175
Normal	1.1427	1.3158	1.3379
Uniform	1.1428	2.2948	2.6437

Table 4 : Asymptotic Relative Efficiency of $V(1,d)$ relative to $U(1,d)$

Density	$d=2$	$d=3$	$d=4$
Cauchy	1.0000	0.6288	0.4808
Laplace	1.0000	0.7407	0.7347
Logestic	1.0000	1.0000	1.0000
Normal	1.0000	1.0918	1.1389
Triangular	1.0000	1.1378	1.2461
Uniform	1.0000	1.7778	2.2500

4. Discussions

1. It follows from Lehmann [6] that the test is consistent for testing H_0 against H_1. Since expected value of $V(c,d)$ under H_1 is greater than its expected value under H_0 and the asymptotic distribution of the test statistic is normal.

2. The proposed class of test statistic is constructed in such a way that the statistic is based on the signs of the observations that is the largest value in subsamples of sizes c and d taken from X and Y samples.

3. The performances of the members of our class are better than Mann-Whitney statistics for light tailed distributions or the distributions with finite range.

4. For medium tailed distribution, the members of our class are better for $r > 2$.

5. For $r = 2$, our test is the best test statistics for light as well as heavy tailed distribution except for Cauchy distribution.

6. It can be seen that ARE's increase with increase in the sub sample size r. For heavy tailed distributions, ARE's decrease as r increases and for medium tailed distributions, ARE's increase and then decrease as r increases.

7. The performance of the members of our class is better as compared to Deshpande and Kochar's [2] test for both heavy and light tailed distributions.

8. The performance of the members of our class is better as compared to Stephenson and Ghosh [12] test for light and medium tailed distributions.

Acknowledgement

The second author would like to thank the University Grants Commission for its support under FDP scheme.

References

[1] Ahmad I. A., A class of Mann-Whitney-Wilcoxon type statistics. The American Statistician. 50 (1996), 324-327.

[2] Deshpande J. V. and Kochar S. C., Some competitors of Wilcoxon-Mann Whitney test for the location alternatives. J. Indian Statist. Assoc. 19 (1982), 9-18.

[3] Gastwirth J. L., Percentile modifications of two sample rank tests. J. Amer. Statist. Assoc, 60 (1965), 1127-1141.

[4] Hoeffding W., A class of statistics with asymptotically normal distribution. Ann. Math. Statist., 19 (1948), 293-325.

[5] Hogg R. V., Fisher D. M. and Randles R. H., A two-sample adaptive distribution-free test. J. Amer. Statist. Assoc., 70 (1975), 656-661.

[6] Lehmann E. L., Consistency and unbiasedness of certain nonparametric tests. Ann. Math. Statist., 22 (1951), 165-179.

[7] Mann H. B. and Whitney D. R., On a Test of whether one of the two random variables is stochastically larger than the other. Ann. Math. Statist, 18 (1947), 50-60.

[8] Mathisen H. C., A method of testing the hypothesis that two samples are from the same population. Ann. Math. Statist, 14 (1943), 188-94.

[9] Randles R. H. and Wolfe D. A., Introduction to the Theory of Non-parametric Statistics. John Wiley and Sons, New York. (1979).

[10] Shetty I. D. and Govindarajulu Z., A two-sample test for location. Commun. Statist.- Theor. Meth. 17(7) (1988), 2389-2401.

[11] Shetty I. D. and Bhat S. V., A note on the generalization of Mathisens median test. Statistics and Prob. Letters, 19 (1994), 199-204.

[12] Stephenson W. R. and Gosh M., Two sample non-parametric tests based on subsamples. Commun. Statist. -Theor. Meth., 14 (1985), 1669-1684.

[13] Xie J. and Priebe C. E., Generalizing the Mann-Whitney Wilcoxon Statistic. Non-parametric Statistics, 12 (2000), 661-682.