한국통계학회논문집 제3권 2호
1996년 8월 pp. 197-204

# A Bivariate Two Sample Rank Test for Mixture Distributions

Songyong Sim[1], Seungmin Lee[2]

## Abstract

We consider a two sample rank test for a bivariate mixture distribution based on Johnson's quantile score. The test statistic is simple to calculate and the exact distribution under the null hypothesis is obtained. A numerical example is given.

## 1. Introduction

When we compare two samples the usual alternative hypotheses are that the two populations $F$ and $G$ are not the same or that they are stochastically ordered. Often these alternative hypotheses are specified in terms of location or scale parameter. Median test is one of such tests. However, as Carrano and Moore (1982) found in a study of sister chromosome exchanges, it sometimes happens that the alternative is not a shift of the whole population but an increase in proportions of large responses. We may formulate this situation as follows; Let $F_1$ and $F_2$ be two distributions such that $F_1(x) \geq F_2(x)$ for all $x$. Let $Z_1, Z_2, \cdots, Z_m$ be a random sample from

$$F_\varepsilon(z) = (1-\varepsilon)F_1(z)+\varepsilon F_2(z) \tag{1}$$

for $0 < \varepsilon < 1$ and $Z_{m+1}, Z_{m+2}, \cdots, Z_{m+n}$ be a random sample from a distribution

$$F_{\varepsilon,\delta}(z)=(1-\varepsilon-\delta)F_1(z)+(\varepsilon+\delta)F_2(z) \tag{2}$$

for $0<\delta<1-\varepsilon$. We assume the two samples are independent. Johnson, Verrill, Moore (1987) suggested a locally most powerful rank test for testing $H_0: \delta = 0$ against $H_1: \delta > 0$ for $\delta$ in (2). The score function, called quantile score, they obtained is

$$\varphi_\varepsilon(u)=\begin{cases}0 & \text{if } 0<u\leq(1-\varepsilon)b_1+\varepsilon b_2 \\ 1 & \textit{otherwise}\end{cases}, \tag{3}$$

---

1) Full-time Lecturer, Department of Statistics, Hallym University, Chuncheon 200-702, Korea. This research was supported in part by '95 Hallym University Research Fund.

2) Associate Professor, Department of Statistics, Hallym University, Chuncheon 200-702, Korea.

where $b_1$ and $b_2$ are prefixed numbers such that $(1-\varepsilon)b_1+\varepsilon b_2$ assumes values between zero and one. Note that the quantile test generalized the univariate median test in the sense that the test gives weight 1 for all observations with ranks larger than some fixed number. Johnson and Sim (1995) extended quantile test to ordered alternatives.

For the bivariate one sample location problem, Chatterjee (1966) suggested a two dimensional extension of sign test. The bivariate signed test is associated with a two dimensional median, which is a paired marginal median. Chatterjee's test is based on the number of concordance and disconcordance of the first kind conditionally on the number of concordance. The test is conditionally distribution-free under the null hypothesis.

In this paper, we would like to extend the test to the case when we have bivariate mixture observations, for example, to the case when we observe the level of cholesterol and blood pressure from a patient at the same time. Suppose that $(X_i, Y_i)^T$ is an observation from

$$(1-\varepsilon-\delta)F_1(x,y)+(\varepsilon+\delta)F_2(x,y) \ . \tag{4}$$

The null hypothesis is $H_0: \delta = 0$ in (4) when underlying distribution a bivariate mixture distribution (4).

Section 2 gives formal description of the problem. In Section 3 we propose test statistic and its property. We applied our test to a real data set in Section 4.

## 2. Bivariate Two Sample Problem

Suppose $Z_i = (X_i, Y_i)^T$ for $i = 1, 2, \cdots, N$ with $N = m + n$ be the independent observations we make. Let $Z_i$, for $i = 1, 2, \cdots, m$ be the observations from the first sample and $Z_i$, for $i = m+1, m+2, \cdots, m+n$ be the observations from the second sample. Since two coordinates are usually dependent for bivariate observations, we do not assume that $X_i$ and $Y_i$ are independent while we assume that $Z_i$'s independent.

Then we can write an observation matrix $Z$ as follows;

$$Z=\begin{pmatrix} X_1 \cdots X_m \, X_{m+1} \cdots X_N \\ Y_1 \cdots Y_m \, Y_{m+1} \cdots Y_N \end{pmatrix} . \tag{5}$$

In order to formalize the hypotheses, we introduce a definition of stochastic ordering for two dimensional random vector. Let $Z = (X, Y)^T$ have a distribution function $F(x, y)$ and set

$$\overline{F}(x,y)=P[\ Z\geq z] \ =1-F(x,\infty)-F(\infty,y)+F(x,y) \ . \tag{6}$$

**Definition 1** (Bhattacharyya and Johnson, 1970)
A random variable $Z = (X, Y)^T$ with distribution $F$ is strongly smaller than $Z'$ with distribution $G$ ( $Z <_s Z'$) if $F \neq G$ and $F(x, y) \geq G(x, y)$ , $\overline{F}(x, y) \leq \overline{G}(x, y)$ hold for all $(x, y)$. $Z$ is weakly stochastically smaller (Type 3) than $Z'$ ( $Z <_{w3} Z'$) if $F \neq G$ and $F(x, y) \geq G(x, y)$ for all $(x, y)$.
$Z$ is weakly stochastically smaller (Type 1) than $Z$ ( $Z <_{w1} Z'$) if $F \neq G$ and for all $(x, y)$ $\overline{F}(x, y) \leq \overline{G}(x, y)$ .)

Let $F_1$ and $F_2$ are continuous bivariate distribution functions on the real plane. We assume stochastic ordering of the first type for $F_1$ and $F_2$. That is $F_1 \neq F_2$ and for $\overline{F_1}(x, y) \leq \overline{F_2}(x, y)$ all $(x, y)$. Suppose that $Z_i$, for $i = 1, 2, \cdots, m$ is a random sample from a distribution with a distribution function

$$F_\varepsilon(x, y) = (1 - \varepsilon)F_1(x, y) + \varepsilon F_2(x, y) \tag{7}$$

for $0 < \varepsilon < 1$ and $Z'$, for $i = m+1, m+2, \cdots, m+n$ is an independent random sample from a distribution function

$$F_{\varepsilon, \delta}(x, y) = (1-\varepsilon-\delta)F_1(x, y) + (\varepsilon+\delta)F_2(x, y) \tag{8}$$

for $0 < \delta < 1-\varepsilon$. We are interested in testing the null hypothesis

$$H_0 : \delta = 0 \quad \text{Or two samples are from the same distribution} \tag{9}$$

against the alternative

$$H_1 : \delta > 0 \quad \text{Or Second sample is stochastically larger} \tag{10}$$

Consider coordinatewise ranks of $X_i'$s and $Y_i'$s, for $i = 1, 2, \cdots, N$ of the observation matrix (5). Let $R_i$ be the rank of $X_i$ among $X_1, X_2, \cdots, X_N$ and $S_i$ be the rank of $Y_i$ among $Y_1, Y_2, \cdots, Y_N$. The rank matrix $R$ is

$$R = \begin{pmatrix} r(X_1) \cdots r(X_m) & r(X_{m+1}) \cdots r(X_N) \\ r(Y_1) \cdots r(Y_m) & r(Y_{m+1}) \cdots r(Y_N) \end{pmatrix} = \begin{pmatrix} R_1 \cdots R_N \\ S_1 \cdots S_N \end{pmatrix} . \tag{11}$$

Note that each row of $R$ in (11) is a combination of $1, 2, \cdots, N$.

We call two rank matrices of the form (11) are permutationally equivalent if one can be obtained from the other by a rearrangement of its column. Then the matrix $R$ is permutationally equivalent to another matrix $R^*$ if $R^*$ has the same columns as in $R$. Since $X_i$ and $Y_i$ are, in general, dependent on the joint distribution of the elements of $R$

will be dependent on the unknown (in practice) distribution function $F_\varepsilon$ even under the null hypothesis. However, the conditional distribution of $\boldsymbol{R}$ over the set of all possible permutations of $\boldsymbol{R}$, the distribution will be uniform under the null hypothesis. By denoting all the possible permutations of $\boldsymbol{R}$ as $S(\boldsymbol{R})$ we have

$$P[\ \boldsymbol{R} = \boldsymbol{r} \mid H_0] = \frac{1}{N!} \text{ for all } \boldsymbol{r} \in S(\boldsymbol{R}). \tag{12}$$

Let us denote P be the conditional permutational probability measure generated by the $N!$ equally likely possible permutations of the columns of the rank matrix $\boldsymbol{R}$.

## 3. Proposed Test

Let $\boldsymbol{A}$ be a $2\times N$ matrix obtained by applying the following score to the elements of rank matrix $\boldsymbol{R}$. The first row is obtained by

$$a(R_i/(N+1)) = \begin{cases} 1 & \text{if } R_i/(N+1) > b_1 \text{ or } R_i \geq [\ b_1(N+1)] \\ 0 & \textit{otherwise} \end{cases} \tag{13}$$

and the second row of $\boldsymbol{A}$ is obtained by

$$a(S_i/(N+1)) = \begin{cases} 1 & \text{if } S_i/(N+1) > b_2 \text{ or } S_i \geq [\ b_2(N+1)] \\ 0 & \textit{otherwise} \end{cases}. \tag{14}$$

In other words, we give weights 1 if the $k$-th coordinate is larger than $100b_k\%$ percentiles and weights 0 otherwise, for $k=1, 2$.

We call the matrix $\boldsymbol{A}$ a score matrix. Denote $c_k = [\ b_k(N+1)]$ . We have $(N-c_k)$ 1's along the $k$-th row of $\boldsymbol{A}(k = 1, 2)$. Note that each column of the score matrix $\boldsymbol{A}$ is either $(0,0)^T$, $(0,1)^T$, $(1,0)^T$ or $(1,1)^T$. Let $U_{ij}$ be the number of $(i, j)^T$ for $i, j = 0, 1$. Furthermore, each realization of score matrix $\boldsymbol{A}$ has the same probability under the null hypothesis. Since there exist $N!/(U_{00}!\,U_{01}!\,U_{10}!\,U_{11}!)$ different realizations under the permutation of the rank matrix $\boldsymbol{R}$, the probability for each realization under the null hypothesis is $(U_{00}!\,U_{01}!\,U_{10}!\,U_{11}!)/N!$

If the $i$-th column of $\boldsymbol{A}$ is $(1, 1)^T$, we have $X_1$ and $Y_i$ are greater than $100\,b_k$-th percentile of $X$'s and $Y$'s, respectively. Each row of the score matrix is coordinatewise score of $X$ and $Y$.

Let $U_1$ be the number of $(1, 1)^T$ from the first sample and $U_2$ be the number of

$(1, 1)^T$ from the second sample. That is, $U_1$ and $U_2$ are the number of observations with first and second coordinates are larger than the $100b_i\%$-th percentile for $i = 1, 2$, respectively.

We consider a test based on $U_2$ conditional on $U = U_1 + U_2$ and we reject $H_0$ for large values of $U_2|U$. Under the null hypothesis, the conditional permutational distribution of $U_2$ given $U$ is a hypergeometric distribution with its density

$$f(u_2|u) = \frac{\binom{m}{u_1}\binom{n}{u_2}}{\binom{N}{u}} \quad . \tag{15}$$

Under the null hypothesis, the conditional expectation and variance of $U_2$ given $U = u$ are $E[\ U_2 | U = u] = u\frac{n}{N}$ and $Var(U_2 | U = u) = u\frac{n}{N}\frac{N-n}{N}\frac{N-u}{N-1}$. The null distribution we obtained is (conditionally) distribution free. We note that this test is similar to Chatterjee's test in the sense we count concordance.

## 4. A Numerical Example

Johnson, Sim, Klein and Klein (1995) has bivariate observations. The data are proportions of cortical opacity in left and right eyes for different age groups. The original data are given in Table 1. As the data are proportions we may apply the usual transformation arcsin $(\sqrt{proportion})$. After applying the transformation the scatterplot of both eyes in two age groups are shown in Figure 1. Note that same values have only one plotting point in the figure. We used the quantile $b = 0.8$ for both the coordinates so that only those observations larger than 80 % percentiles of both the left and right eye lens opacity from the combined sample of size 100 are counted. It turns out that only 10 observations are above the 80 % percentile of both eyes. Among those one is of age less than 65 and nine are of age greater than or equal to 65. Hence the conditional permutational $p$-value is

$$\Pr[\ X \le 1] = \frac{\binom{50}{x}\binom{50}{10-x}}{\binom{100}{10}} = 0.00783 \ . \tag{16}$$

So we can conclude that we reject the null hypothesis at $\alpha = 0.01$.

Table 1 : Percent Area Involved by Cortical Opacity (50 observations each)

| Age 43-64 | | | | Age 65+ | | | |
|---|---|---|---|---|---|---|---|
| Right | Left | Right | Left | Right | Left | Right | Left |
| 0.0 | 0.0 | 0.0 | 0.0 | 11.0 | 0.5 | 25.0 | 0.0 |
| 0.0 | 0.0 | 0.0 | 0.5 | 0.0 | 0.0 | 50.0 | 3.0 |
| 0.5 | 0.0 | 1.0 | 1.0 | 31.0 | 35.0 | 0.0 | 0.0 |
| 1.0 | 0.0 | 1.0 | 0.0 | 0.0 | 1.0 | 1.0 | 0.0 |
| 1.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 | 5.0 | 1.0 |
| 1.0 | 0.0 | 2.0 | 0.0 | 1.0 | 0.0 | 1.0 | 1.0 |
| 0.5 | 0.0 | 0.0 | 0.5 | 3.0 | 0.0 | 99.0 | 95.0 |
| 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | 14.0 | 1.0 |
| 0.0 | 1.0 | 0.0 | 0.0 | 3.0 | 0.0 | 1.0 | 0.0 |
| 1.0 | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 | 1.0 | 2.0 |
| 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.5 | 6.0 | 0.0 |
| 0.0 | 0.0 | 0.0 | 1.0 | 8.0 | 1.0 | 2.0 | 10.0 |
| 4.0 | 0.5 | 1.0 | 0.0 | 6.0 | 60.0 | 18.0 | 1.0 |
| 2.0 | 10.0 | 0.0 | 0.0 | 0.0 | 2.0 | 0.0 | 2.0 |
| 1.0 | 1.0 | 1.0 | 0.0 | 0.0 | 2.0 | 1.0 | 2.0 |
| 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 4.0 | 2.0 |
| 1.0 | 0.0 | 1.0 | 0.0 | 1.0 | 1.0 | 1.0 | 0.0 |
| 0.0 | 0.0 | 0.5 | 1.0 | 0.0 | 2.0 | 90.0 | 20.0 |
| 1.0 | 1.0 | 1.0 | 1.0 | 9.0 | 1.0 | 25.0 | 1.0 |
| 0.0 | 0.0 | 0.0 | 0.0 | 0.5 | 0.0 | 80.0 | 65.0 |
| 0.0 | 0.0 | 5.0 | 2.0 | 17.0 | 12.0 | 0.0 | 0.0 |
| 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 9.0 | 18.0 |
| 1.0 | 0.0 | 0.0 | 0.0 | 2.0 | 5.0 | 0.0 | 0.0 |
| 0.0 | 0.0 | 1.0 | 0.0 | 0.5 | 1.0 | 5.0 | 0.0 |
| 0.0 | 0.0 | 0.0 | 0.0 | 3.0 | 4.0 | 1.0 | 1.0 |

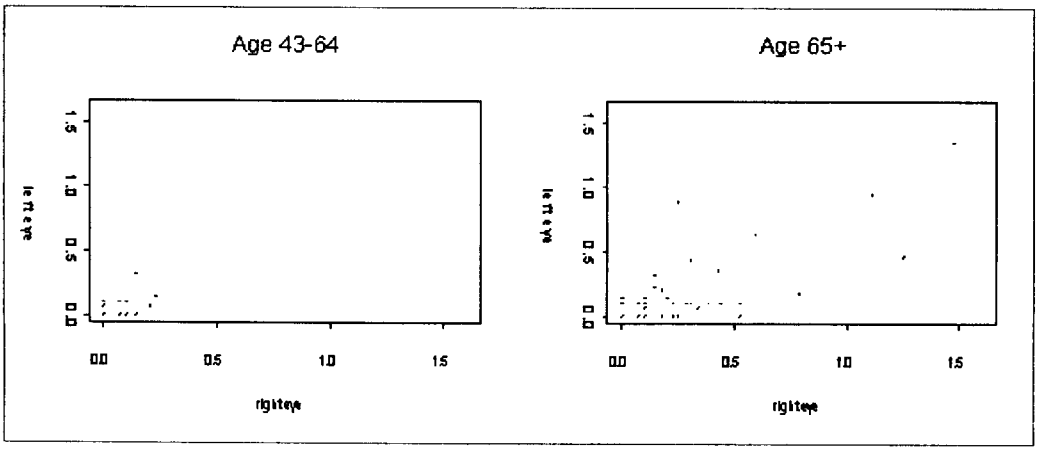


Figure 1 : Scatterplot of left and right eyes

## 5. Simulation study

To investigate the power of the proposed test, an empirical power study is done for bivariate mixture of two normal distributions $F_1$ and $F_2$. As there are too many parameters involved in our problem, we restrict ourselves to the case when $\varepsilon = 0$ and $F_1$ is normal with mean $\mathbf{0}$ and covariance $\begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$. Our choice of $F_2$ was bivariate normal with $\mu$ and the same covariance matrix as $F_1$. We change the values of correlation $r$, mean vector $\mu$ and values of $\delta$. Note that when $\delta = 0$, it is the case when the $H_0$ is true. We also changed the number of observations for two samples.

The number of rejections out of 1000 trials at 5% significance level for each combination of different conditions are given in Table 2. The numbers in the parentheses are the cases when the tests were impossible as the number of observations larger than both the coordinatewise 60% sample percentiles.

As we see in the table, the power increases as $\delta$ increases. Also as the shift $\mu$ increases, the power increases. One interesting aspect is that the power decreases as the correlation coefficient becomes large. One final remark on the simulation results is that the numbers of rejections under the null hypothesis are consistently smaller than 50. The reason is that the test is discrete and we do not use randomization.

## References

[1] Bhattacharyya, G. K. and Johnson, R. A. (1970). A Layer Rank Test for Ordered Bivariate Alternatives, *Annals of Mathematical Statistics*, Vol. 41, No. 1, 1296-1310.

[2] Carrano, A. and Moore, D. (1982). The Rationale and Methodology for Quantifying Sister Chromatid Exchange in Humans. In *Mutagenicity: New Horizons in Genetic Toxicology*, J. A. Heddle (ed), 268-304 New York: Academic Press.

[3] Chatterjee, S. K. (1966). A Bivariate Sign Test for Location, *Annals of Mathematical Statistics*, Vol. 37, 1771-1782.

[4] Johnson, R. A., Verrill, S., and Moore II, D. H. (1987). Two-Sample Rank Tests for Detecting Changes That Occur in a Small Proportion of the Treated Population, *Biometrics*, Vol. 43, 641-655

[5] Johnson, R. A., Sim, S., Klein, B. E. and Klein, R. (1995). A Multivariate Multi-Sample Quantile Test for Ordered Alternatives, *Tech. Rep.* No. 950, Department of Statistics, University of Wisconsin-Madison.

[6] Johnson, R. A. and Sim, S. (1995). Nonparametric Tests for Ordered $c$-Sample Contaminated Distributions, *Communications in Statistics - Theory and Methods*, Vol. 24, 861-879.

[7] Oja, H. (1983). Descriptive Statistics for Multivariate Data Statistics and Probability Letters 1, 327-332.

Table 2 : Number of rejections out of 1000 tests at 5 %

| ($n_1 = 50$, $n_2 = 50$) | | | | | |
|---|---|---|---|---|---|
| $\rho$ | $(\mu_1, \mu_2)$ | $\delta = 0.0$ | $\delta = 0.1$ | $\delta = 0.2$ | $\delta = 0.3$ |
| -0.7 | (1,1) | 8 (3) | 127 (1) | 350 (0) | 640 (0) |
| | (2,2) | 11 (5) | 369 (0) | 862 (0) | 985 (0) |
| | (3,3) | 11 (7) | 420 (0) | 892 (0) | 995 (0) |
| -0.3 | (1,1) | 26 (0) | 117 (0) | 297 (0) | 518 (0) |
| | (2,2) | 31 (0) | 269 (0) | 703 (0) | 940 (0) |
| | (3,3) | 26 (0) | 259 (0) | 744 (0) | 981 (0) |
| 0.0 | (1,1) | 21 (0) | 104 (0) | 209 (0) | 358 (0) |
| | (2,2) | 27 (0) | 182 (0) | 532 (0) | 847 (0) |
| | (3,3) | 31 (0) | 210 (0) | 627 (0) | 924 (0) |
| 0.3 | (1,1) | 27 (0) | 69 (0) | 162 (0) | 258 (0) |
| | (2,2) | 22 (0) | 150 (0) | 423 (0) | 727 (0) |
| | (3,3) | 38 (0) | 169 (0) | 497 (0) | 846 (0) |
| 0.7 | (1,1) | 31 (0) | 95 (0) | 205 (0) | 382 (0) |
| | (2,2) | 27 (0) | 146 (0) | 421 (0) | 732 (0) |
| | (3,3) | 26 (0) | 159 (0) | 487 (0) | 821 (0) |
| ($n_1 = 100$, $n_2 = 100$) | | | | | |
| -0.7 | (1,1) | 20 (0) | 293 (0) | 712 (0) | 925 (0) |
| | (2,2) | 19 (0) | 684 (0) | 992 (0) | 1000 (0) |
| | (3,3) | 13 (0) | 768 (0) | 998 (0) | 1000 (0) |
| -0.3 | (1,1) | 28 (0) | 213 (0) | 567 (0) | 830 (0) |
| | (2,2) | 37 (0) | 476 (0) | 949 (0) | 998 (0) |
| | (3,3) | 29 (0) | 557 (0) | 967 (0) | 1000 (0) |
| 0.0 | (1,1) | 36 (0) | 168 (0) | 400 (0) | 681 (0) |
| | (2,2) | 36 (0) | 340 (0) | 830 (0) | 988 (0) |
| | (3,3) | 43 (0) | 395 (0) | 906 (0) | 997 (0) |
| 0.3 | (1,1) | 37 (0) | 115 (0) | 319 (0) | 495 (0) |
| | (2,2) | 35 (0) | 293 (0) | 693 (0) | 944 (0) |
| | (3,3) | 39 (0) | 341 (0) | 804 (0) | 986 (0) |
| 0.7 | (1,1) | 35 (0) | 147 (0) | 408 (0) | 657 (0) |
| | (2,2) | 37 (0) | 298 (0) | 749 (0) | 966 (0) |
| | (3,3) | 30 (0) | 328 (0) | 810 (0) | 990 (0) |