American Journal of Mathematics and Statistics 2015, 5(4): 196-199
DOI: 10.5923/j.ajms.20150504.05

Rank Tests for Independence Against Weighted Alternative with Missing Values

Parameshwar V. Pandit[1,*], Savitha Kumari[2]

[1]Department of Statistics, Bangalore University, Bangalore, India
[2]Department of Statistics, SDM College, Ujire, India

Abstract One of the common problems of practical importance is that of determining whether there is independence between a pair of random variables. In this paper, the problem of testing independence of bivariate random variables against a weighted alternative model with possible missing values on both responses is considered. The model considered here is due to Shei, Bai and Tsai [9] which is the generalization of Hajek and Sidak [12] model with weighted contamination. A new rank test based on ranks is proposed and its asymptotic normality is established. Locally most powerful tests for the model is derived. The asymptotic null distributions of the test statistics are also provided for the purpose of practical use.

Keywords Independence tests, Locally most powerful rank test, Weighted dependence alternative, Missing values

1. Introduction

The problem of interest for the statisticians is that of determining whether there exists a relationship between two characteristics in a population. In the literature several authors attempted the quantification of the concept of stochastic dependence for bivariate distributions. Rank tests for independence based on complete data can be found in Spearman [1], Kendal [2], Bhuchongkul [3], Puri and Sen [4], Shirahata [5] among others. Iman and Conover [6] proposed a measure of dependence, which is the Pearson correlation coefficient computed on Savage [7] scores, reflects the importance on the top ranks. Shieh [8], proposed a weighted Kendal's tau statistic. Shieh, Bai and Tsai [9] proposed some rank tests and derived locally most powerful rank test for testing independence against a weighted contaminated alternative. Pandit [10] considered this problem with a different weighted alternative and derived locally most power rank test. However, in practical situations, some observations on either of the variables may be missing. In such a situation the tests mentioned above cannot be applied. Wei [11] derived a locally most powerful rank test for independence against the alternative given by Hajek and Sidak [12] in presence of missing values. In this paper, we propose rank tests and derive locally most powerful rank test for independence against weighted alternatives in presence of missing values.

Let $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n), (X_{n+1}, .), \dots, (X_{n+m}), (., Y_{n+1}), \dots, (., Y_{n+k})$ be a random sample from a bivariate distribution function F(x,y). The problem is to test $H_0 : F(x,y) = F_1(x).F_2(y), \quad for \ \ all \ (x,y)$. Here, The alternative considered is as below:

$$X = X^* + u(X^*)\Delta Z, \qquad Y = Y^* + \Delta Z \qquad (1)$$

where X^*, Y^* and Z^* are mutually independent and u(x) is monotone in x. (Shieh, Bai and Tsai [9]). The alternative due to Hajek and Sidak [12] is a particular case of (1).

Under (1), it is clear that if $\Delta = 0$, X and Y are independent and larger the Δ is, the more dependent are X and Y. Thus the constant Δ may be regarded as a dependence or mixing coefficient. The alternatives stated in (1) indicate the positive dependence of the random variables X and Y. If we assume negative dependence between X and Y, then the model (1) is

$$X = X^* + u(X^*)\Delta Z, \qquad Y = Y^* - \Delta Z \qquad (1^*)$$

where X^*, Y^* and Z^* are mutually independent and u(x) is monotone in x.

Let $R_1, R_2, \dots, R_{n+m}$ be the ranks of the first coordinates $X_1, X_2, \dots, X_{n+m}$ of the sample and $Q_1, Q_2, \dots, Q_{n+k}$ be the ranks of the second coordinates $Y_1, Y_2, \dots, Y_{n+k}$ of the sample. Rank tests developed for testing independence against the alternatives considered are the functions of R_i's and Q_j's, i=1,2,…, n+m, j=1,2,…, n+k. In section 2, we propose a new test for testing independence and the LMPR test for the alternative (1) is considered in section 2. The LMPR test for the alternative (1) is derived in section 3. In section 4, we give some remarks.

[*] Corresponding author:
panditpv12@gmail.com (Parameshwar V. Pandit)

Published online at http://journal.sapub.org/ajms

2. New Rank Test for Testing Independence

The model of dependence considered is that considered in Shei, Bai and Tsai [9]. The model for a random sample with missing values is as specified below. Let

$$X_i = X_i^* + \Delta u(X_i^*)Z_i, \quad Y_i = Y_i^* + \Delta Z_i, \quad i = 1,2,\dots,n$$
$$X_j = X_j^* + \Delta u(X_j^*)Z_j, \quad Y_l = Y_l^* + \Delta Z_{m+l}, \quad j = 1,2,\dots,m; l = 1,2,\dots,k.$$

The variables X^*, Y^* and Z are independent and Δ is a real nonnegative parameter.

First, we propose a test statistic for testing bivariate independence against model (1), which is the modified Spearman's coefficient to include missing observations. The statistic is defined by

$$W_s^* = \sum_{i=1}^{n} u_i \left(R_i - \frac{n+m+1}{2} \right)\left(Q_i - \frac{n+k+1}{2} \right).$$

Here, $u_i = I(i \le n^*)$, where $n^* = (n+1)p$ and $0 \le p \le 1$ is roughly the proportion of the observed items. The choice of p is to have less loss in significance level.

The asymptotic distribution of the statistic W_s^* is given in the following theorem.

Let *I(f)* denote the Fisher information, $I(f) = \int_{-\infty}^{\infty} \left(\frac{f'(x)}{f(x)} \right)^2 f(x)dx$.

Theorem 1: Assume that H_0 holds, *$I(f_{10})<\infty$ and $I(f_{20})<\infty$*. Then, $\dfrac{W_s^*}{\sqrt{\dfrac{pn(n^2-1)}{12}}}$ converges in distribution to standard normal.

Proof: Consider,

$$Var(W_s^*) = Var\left(\sum_{i=1}^{n} u_i \left(R_i - \frac{n+m+1}{2} \right)\left(Q_i - \frac{n+k+1}{2} \right) \right)$$

and under H_0 ,

$$Var(W_s^*) = E\left(\sum_{i=1}^{n} u_i \left(R_i - \frac{n+m+1}{2} \right)\left(Q_i - \frac{n+k+1}{2} \right) \right)^2$$
$$= E\left(\sum_{i=1}^{n^*} \left(R_i - \frac{n+m+1}{2} \right)\left(Q_i - \frac{n+k+1}{2} \right) \right)^2$$
$$\begin{aligned} &= n^* Var(R_1)Var(Q_1) \\ &\quad + n^*(n^*-1)Cov(R_1,R_2)Cov(Q_1,Q_2) \end{aligned}. \qquad (2)$$

Further, we have

$$E(R_1) = \frac{(n+m+1)}{2} \ , Var(R_1) = \frac{\lfloor (n+m)^2 - 1 \rfloor}{12},$$

$$E(Q_1) = \frac{(n+k+1)}{2} \ , Var(Q_1) = \frac{\lfloor (n+k)^2 - 1 \rfloor}{12},$$

$$E(R_1R_2) = \frac{[(n+m+1)(3(n+m)+2)]}{12}, \ .$$

$$Cov(R_1,R_2) = \frac{-(n+m+1)}{12}$$

$$E(Q_1Q_2) = \frac{[(n+k+1)(3(n+k)+2)]}{12},$$

$$Cov(Q_1,Q_2) = \frac{-(n+k+1)}{12}$$

Substituting these in equation (2) and as n→∞, we have

$$Var\left(\sqrt{n}W_s^* \Big/ \left[n(n^2-1)/12 \right] \right) \to p.$$

Assuming that *$I(f_{10})<\infty$ and $I(f_{20})<\infty$,* and applying Theorem V.1.6a in Hajek and Sidak [12], we have the required result.

3. Locally Most Powerful Rank Test for Weighted Alternative

Here, we consider the model of dependence which is the generalization of the model considered by Bhuchongkul [3], and, Hajek and Sidak [12]. The model for a random sample with missing values is as specified below. Let

$$X_i = X_i^* + \Delta u(X_i^*)Z_i, \ Y_i = Y_i^* + \Delta Z_i, \ i = 1,2,\dots,n$$
$$X_j = X_j^* + \Delta u(X_j^*)Z_j, \ Y_l = Y_l^* + \Delta Z_{m+l}, \ j = 1,2,\dots,m; l = 1,2,\dots,k.$$

The random variables X^*, Y^* and Z are independent; Δ is a real nonnegative parameter and w(x) monotone in x. Under the above model it is clear that if Δ =0, X and Y are independent.

Now, let X^* and Y^* have the p.d.f.s $f_{10}(x)$ and $f_{20}(y)$ respectively and the distribution of Z is arbitrary. Let $x^*=t(x, \Delta z)$ be the unique solution of the equation $x=x^*+w(x^*)\Delta z$, for given x and Δz. The joint p.d.f., q_Δ, of (X_1, Y_1), (X_2, Y_2), … , (X_n, Y_n), $(X_{n+1}, .)$,…, (X_{n+m}), $(. , Y_{n+1})$, …, $(. , Y_{n+k})$ is given by

$$q_\Delta = \prod_{i=1}^{n} h_\Delta(x_i, y_i). \prod_{j=n+1}^{n+m} f_{10}(x_j^*). \prod_{l=n+1}^{n+k} f_{20}(y_l - \Delta z),$$

where $h_\Delta(x,y) = \int_{-\infty}^{\infty} f_{10}(x^*).f_{20}(y-\Delta z).dM(z)$ and $M(z)$ is the distribution function of Z with mean μ_z and finite variance σ_z^2.

Let $X_{(i)}$ and $Y_{(i)}$ be the i-th order statistic of $\{X_1,X_2,\dots,X_{n+m}\}$ and $\{Y_1,Y_2,\dots,Y_{n+k}\}$ respectively. Further, let

$$a_{n+m}(r_i, f_{10}) = E\left\{-\left(\frac{(uf_{10})'}{f_{10}}\right)(X_{(i)})\right\}$$

and $a_{n+k}(q_i, f_{20}) = E\left\{-\left(\frac{f_{20}'}{f_{20}}\right)(Y_{(i)})\right\}$ denote the score functions corresponding to f_{10} and f_{20} respectively. In order to obtain LMPR test we assume the following conditions:

(i) The derivatives $(wf_{10})'$ and f_{20}' are continuous,

(ii) $\int_{-\infty}^{\infty}\left|(uf_{10})'(x)\right|dx < \infty$ and $\int_{-\infty}^{\infty}\left|f_{20}'(x)\right|dx < \infty$,

(iii) The missing observations on either the first coordinate or the second coordinate occur at random.

The following theorem states the LMPR test.

Theorem 2: Under the conditions (i) to (iii) and for the model (1), the test statistic $V_1 = \sum_{i=1}^{n} a_{n+m}(r_i, f_{10}).a_{n+k}(q_i, f_{20})$ with critical region $V_1 > c$, where *c* is a constant, is locally most powerful rank test for testing $H_0 : \Delta = 0$ against $H_1 : \Delta > 0$ at corresponding level of significance.

Corollary 1: If *f₁₀ and f₂₀* are from Logistic family, then the test based on W_s^* with critical region $W_s^* \geq c$, where *c* is a constant, is asymptotic LMPR test for testing $H_0 : \Delta = 0$ against $H_1 : \Delta > 0$ for model (1).

4. Power Comparisons

In this section, we compare the powers of the proposed test W_s^* with those of top-down statistic r_T and Kendall's τ. The alternative used is as in (1). For power comparisons, the alternatives considered are

1. X_i, Y_i and Z_i follow normal with mean zero and variance one.
2. X_i, Y_i and Z_i follow logistic(0,-1).

The correlation coefficient between X and Y denoted by ρ in terms of Δ is $\rho = \frac{p\Delta^2\sigma_z^2}{\sqrt{(\sigma_x^2 + p\Delta^2\sigma_z^2)(\sigma_y^2 + \Delta^2\sigma_z^2)}}$. Here it is to be noted that ρ=0 implies the independence. For simulation selected values of ρ are considered. The results are presented in table 1. In the table 1, m is the number of missing observations corresponding to x-values and k is the number of missing observations corresponding to y-values. From the above table it is easily seen that the test V is more powerful than the top-down statistic, $\mathbf{r_T}$ due to Iman and Conover [6] and Kendall's test, **τ**. Similar results are obtained for n=30, 50.

Table 1. Empirical Powers of new test W_s^*, top-down statistic r_T and Kendall's τ for n=20 and *p*=0.9

m	k	ρ	N(0,1)			Logistic(0,1)		
			W_s^*	r_T	τ	W_s^*	r_T	τ
2	2	0.1	0.1325	0.0994	0.0812	0.1578	0.0872	0.0905
		0.4	0.3242	0.1089	0.1582	0.3160	0.0951	0.1567
		0.7	0.5122	0.1279	0.2665	0.5283	0.1056	0.2381
2	3	0.1	0.2624	0.1552	0.1652	0.2764	0.2311	0.2153
		0.4	0.6234	0.5570	0.5284	0.7762	0.6732	0.5932
		0.7	0.9132	0.9981	0.8973	0.9372	0.8973	0.8693
3	3	0.1	0.2925	0.1072	0.0929	0.3078	0.1991	0.2125
		0.4	0.6356	0.1187	0.1696	0.6961	0.5862	0.5685
		0.7	0.9212	0.1286	0.2765	0.9383	0.8984	0.8791
3	4	0.1	0.3514	0.1645	0.1673	0.3454	0.2732	0.2053
		0.4	0.6531	0.5483	0.5096	0.7785	0.6578	0.5937
		0.7	0.9437	0.9104	0.8973	0.9478	0.9073	0.8774

5. Some Remarks and Conclusions

1. The paper considers the problem of independence against a weighted alternative when the data have missing values. The alternative considered is the generalization of Hajek and Sidak [12], accommodating weighted contamination.
2. The alternative considered here is the model used in Shei, Bai and Tsai [9].
3. A new rank test is developed and the asymptotic normality of the test statistic is established.
4. Locally most powerful rank(LMPR) test for this problem is derived for the alternative used in Shei, Bai and Tsai [9] and the LMPR derived in Wei [11] for the alternative due to Hajek and Sidak [12] can be obtained as a particular case.
5. The new test proposed here is shown to be LMPR when the marginal distributions of X^* and Y^* belong to logistic family.

REFERENCES

[1] Spearman, C. (1904). The proof and measurement of association between two things. Am. J.Psych. 15, 72-101.

[2] Kendall, M.G. (1962). Rank Correlation Methods. Griffin: London.

[3] Bhuchongkul, S. (1964). A class of nonparametric tests for independence in bivariate populations. Ann. Math. Statist. 35, 138-149.

[4] Puri, M.L. and Sen, P.K. (1971). Nonparametric Methods in Multivariate Analysis. Wiley: New York.

[5] Shirahata, S. (1974). Locally most powerful rank tests for independence. Bull Math. Stat. 16, 11-21.

[6] Iman, R.L. and Conover, W.J. (1987). A measure of top-down correlation. Technometrics, 29, 351-357.

[7] Savage, I.R (1956). Contributions to the theory of rank order statistics- two-sample case. Ann.Math.Statist.27,590-615.

[8] Shieh, G.S. (1998). A weighted Kendal's statistic. Stat. Prob. Letters, 39; 17-24.

[9] Shei, G.S., Bai, Z. and Tsai, W.Y. (2000). Rank tests for independence – with a weighted contamination alternative. Statistica Sinica, 10, 577-593.

[10] Pandit, P.V. (2006). Locally most powerful and other rank tests for independence-with a contaminated weighted alternative. Metrika, 64, 379-387.

[11] Wei, L.J. (1983). Tests for independence in the presence of missing values. Austral. J. Statist., 25, 85-90.

[12] Hajek, J. and Sidak, Z. (1967). Theory of Rank Tests. Academic Press; New York.