LEE DISTANCE IN TWO-SAMPLE RANK TESTS

Nikolay I. Nikolov
Institute of Mathematics and Informatics, Bulgarian Academy of Sciences
Sofia, BULGARIA
e-mail: n.nikolov@math.bas.bg

Abstract

In nonparametric statistic there are various procedures to constructing rank tests via metrics on the permutation group. In this paper Critchlow's unified approach is applied to Lee distance. The two-sample location problem is considered and the distribution of the test statistic under null hypothesis is derived and studied.

1 Introduction

Let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ be two independent random samples with continuous distribution functions $F(x)$ and $G(x)$, respectively. We consider rank tests for the two-sample location problem of testing the null hypothesis H_0 against the alternative H_1

$$H_0 : F(x) \equiv G(x)$$
$$H_1 : F(x) \geq G(x),$$

with strict inequality for some x. Let $\alpha(i)$ be the rank of X_i for $i = 1, 2, \ldots, m$ and $\alpha(m+j)$ be the rank of Y_j for $j = 1, 2, \ldots, n$ among $X_1, X_2, \ldots, X_m, Y_1, Y_2, \ldots, Y_n$. Then $\alpha = (\alpha(1), \alpha(2), \ldots, \alpha(m+n))$ is the rank vector of all observations and $\alpha \in S_{m+n}$, where S_{m+n} is the permutation group generated by the first $m+n$ natural integers. The class of permutations, which are most in agreement with the alternative H_1 is $E = S_m \times S_n = \{\pi \in S_{m+n} : \pi(i) \leq m, \forall i \leq m\}$. The left coset $[\alpha] = \alpha(S_m \times S_n) = \{\alpha \circ \pi : \pi \in S_m \times S_n\}$ consists of all permutations in S_{m+n} which are equivalent to α. Many rank statistics could be obtained by using distances between sets of permutation. Critchlow [2] proposed a unified approach to constructing nonparametric tests which produces many well-known rank statistics. The method is based on finding the minimum interpoint distance between the class of equivalence $[\alpha]$ and the extremal set E:

$$d([\alpha], E) = \min_{\substack{\pi \in [\alpha] \\ \sigma \in E}} d(\pi, \sigma), \tag{1}$$

where d is an arbitrary metric on S_{m+n}. The proposed test rejects the null hypothesis H_0 for small values of the statistic $d([\alpha], E)$. This contrasts with the structure of some parametric test, where H_0 is rejected if the distance from H_0 is large. Since the minimal value of the proposed test statistic is zero and $d([\alpha], E) = 0$ if and only if $d(\alpha, \sigma) = 0$ for some $\sigma \in E$, the strongest evidence for rejecting H_0 occurs if and only if the observed permutation α is equivalent to some extremal permutation $\sigma \in E$.

2 Lee distance on S_N

The goal of this paper is to derive and study the rank test statistic in (1) induced by the Lee distance function on S_N:

$$L(a,b) = \sum_{i=1}^{N} \min\left(\mid a(i) - b(i) \mid, N - \mid a(i) - b(i) \mid\right).$$

In nonparametric statistics the right-invariance of a metric is necessary requirement since it means that the distance between rankings does not depend on the labelling of the observations.

Definition 1. The metric d on S_N is called right-invariant, if and only if $d(\alpha,\beta) = d(\alpha\circ\gamma, \beta\circ\gamma)$ for all $\alpha,\beta,\gamma \in S_N$.

Deza and Huang [4] includes extensive discussion of some metrics on the permutation group S_N which are widely used in applied scientific and statistical problems. Critchlow [2] obtained the minimal value defined by (1) for four basic distance functions: Spearman's footrule, Ulam distance, Kendall's tau and Hammning distance, and proved that the induced test statistics are equivalent to some familiar rank statistics. Stoimenova [6] derived the test statistic induced by Chebyshev metric. More properties of these distances can be found in Critchlow [1, 3], Deza [4] and Diaconis [5].

Since $L(a,b)$ is right-invariant it follows

$$\begin{aligned} L([\alpha], E) &= \min_{\substack{\pi\in[\alpha]\\ \sigma\in E}} L(\pi,\sigma) = \min_{\pi\in[\alpha]} L(\pi, e) \\ &= \min_{\pi\in[\alpha]} \left\{ \sum_{i=1}^{m+n} \min\left(\mid a(i) - i \mid, m+n - \mid a(i) - i \mid\right) \right\}, \end{aligned} \tag{2}$$

where $e = (1,2,\ldots,m+n)$ is the identity permutation. After solving the optimal problem (2), $L([\alpha], E)$ can be expressed as

$$\begin{aligned} L([\alpha], E) = 2 \sum_{i\in K_m} \min\big(&\mid \alpha(i) - \gamma_n^{-1}(k+1-\gamma_m(\alpha(i))) \mid, \\ &m+n - \mid \alpha(i) - \gamma_n^{-1}(k+1-\gamma_m(\alpha(i))) \mid\big) \end{aligned} \tag{3}$$

where

$$\begin{aligned} K_m &= \{i \in \{1,2,\ldots,m\} : \alpha(i) > m\} \quad , \\ K_n &= \{i \in \{m+1, m+2, \ldots, m+n\} : \alpha(i) \le m\} \quad , \end{aligned} \tag{4}$$

k is the number of elements of K_m $(k = \mid K_m \mid = \mid K_n \mid)$, $\gamma_m(\alpha(i))$ is the rank of $\alpha(i)$ among $\{\alpha(i) : i \in K_m\}$, $\gamma_n(\alpha(i))$ is the rank of $\alpha(i)$ among $\{\alpha(i) : i \in K_n\}$ and γ^{-1} is the inverse of γ, i.e. $\gamma^{-1}(\gamma(\alpha(i))) = \alpha(i)$. The statistic $L([\alpha], E)$ is equivalent to

$$L := \frac{L([\alpha], E)}{2}. \tag{5}$$

There is an interpretation of the rank test statistic L in terms of graph theory. Let C be a simple cycle graph with vertices $\{i\}_{i=1}^{m+n}$ and edges $\bigcup_{i=1}^{m+n-1}\{i, i+1\}$ and $\{m+n, 1\}$. Then L is the minimum sum of distances over C between the elements of K_m and the elements of K_n. An example when $m = 6$, $n = 4$, $K_m = \{3, 5\}$ and $K_n = \{8, 9\}$ is illustrated on Figure 1. In this case $L = (10 - \mid 3 - 9 \mid) + \mid 5 - 8 \mid = 4 + 3 = 7$.

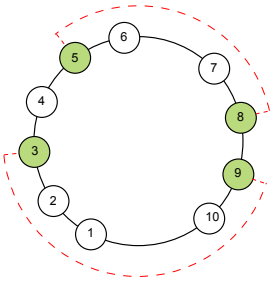

Figure 1: Lee distance on C.

The value of L depends not only on the elements in K_m and K_n, but also on the way in which their elements are paired. Formula (3) gives that the minimal sum of distances between pairwise elements of K_m and K_n is obtained when the smallest element of K_m is combined with the largest element of K_n, the second smallest element of K_m is combined with the second largest element of K_n, ..., the largest element of K_m is combined with the smallest element of K_n. Using this fact the distribution of the test statistic could be calculated for fixed number k of elements in K_m and K_n, $k = \mid K_m \mid = \mid K_n \mid$. Let $[K_m \times K_n]^*$ be the described above set of pairs and $s-1$ be the number of pairs $(x, y) \in [K_m \times K_n]^*$ for which the shortest path on C goes over the edge $\{m, m+1\}$. Obviously, s is between 1 and $k+1$. If for some pair $(x, y) \in [K_m \times K_n]^*$ the paths over $\{m, m+1\}$ and over $\{m+n, 1\}$ are with the same length, then the path over $\{m+n, 1\}$ is considered to be the shortest. For $i = 0, 1, \ldots, s-1$ let $a_i^{(m)}$ be the number of elements in $\{1, 2, \ldots, m\} \backslash K_m$ which are in the shortest path of exactly i pairs $(x, y) \in [K_m \times K_n]^*$ connected by the edge $\{m, m+1\}$. For $j = 1, 2, \ldots, k-s+1$ let $b_j^{(m)}$ be the number of elements in $\{1, 2, \ldots, m\} \backslash K_m$ which are in the shortest path of exactly j pairs $(x, y) \in [K_m \times K_n]^*$ connected by the edge $\{m+n, 1\}$. Similarly the numbers $\{a_i^{(n)}\}_{i=0}^{s-1}$ and $\{b_j^{(n)}\}_{j=1}^{k-s+1}$ are defined for the set $\{m+1, m+2, \ldots, m+n\} \backslash K_n$. An illustration of the used notation is shown on Figure 2.

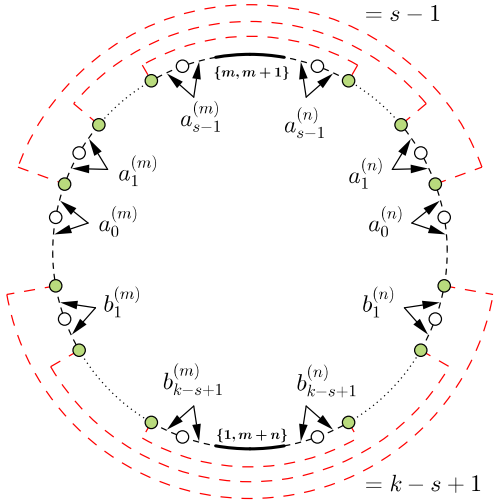

Figure 2: Notations.

For the considered example on Figure 1, $m = 6$, $n = 4$, $[K_m \times K_n]^* = \{(3, 9), (5, 8)\}$, $s = 2$, $a_0^{(m)} = 1 = \mid \{4\} \mid$, $a_1^{(m)} = 1 = \mid \{6\} \mid$, $b_1^{(m)} = 2 = \mid \{1, 2\} \mid$, $a_0^{(n)} = 0$, $a_1^{(n)} = 1 = \mid \{7\} \mid$ and $b_1^{(n)} = 1 = \mid \{10\} \mid$.

Theorem 1. *Let L be defined by (5) and $K = \mid K_m \mid$ be the number of elements of the set K_m, defined by (4). Then the joint distribution of L and K under H_0 is given by*

$$P(L = l, K = k) = \begin{cases} \dfrac{m!n!}{(m+n)!} & , \text{ for } l = 0 \text{ and } k = 0 \\ \displaystyle\sum_{s} \sum_{a,b} \frac{m!n!}{(m+n)!} & , \text{ for } 1 \leq k \leq \min(m, n) \text{ and} \end{cases} \tag{6}$$

$\left[\frac{k^2+1}{2}\right] \leq l \leq \left[\frac{(m+n-k)\,k+1}{2}\right]$, *where* $[x]$ *is the integer part of* x.

The first summation in (6) is taken over all s such that $(s-1)^2+(k-s+1)^2 \leq l$. The second summation is over all nonnegative integers $\{a_i^{(m)}\}_{i=0}^{s-1}$, $\{a_i^{(n)}\}_{i=0}^{s-1}$, $\{b_j^{(m)}\}_{j=1}^{k-s+1}$ and $\{b_j^{(n)}\}_{j=1}^{k-s+1}$ that satisfy:

(i) $\sum_{i=0}^{s-1} a_i^{(m)} + \sum_{j=0}^{k-s+1} b_j^{(m)} = m-k$ **(ii)** $\sum_{i=0}^{s-1} a_i^{(n)} + \sum_{j=0}^{k-s+1} b_j^{(n)} = n-k$

(iii) $l = (s-1)^2 + (k-s+1)^2 + \sum_{i=0}^{s-1} i\left(a_i^{(m)} + a_i^{(n)}\right) + \sum_{j=0}^{k-s+1} j\left(b_j^{(m)} + b_j^{(n)}\right)$

(iv) $2\,(s-1) + \sum_{i=0}^{s-1}\left(a_i^{(m)} + a_i^{(n)}\right) \geq 2\,(k-s) + \sum_{j=0}^{k-s+1}\left(b_j^{(m)} + b_j^{(n)}\right)$, if $s \in \{1,2,\ldots,k\}$

(v) $2\,(s-2) + \sum_{i=1}^{s-1}\left(a_i^{(m)} + a_i^{(n)}\right) < 2\,(k-s+1) + a_0^{(m)} + a_0^{(n)} + \sum_{j=0}^{k-s+1}\left(b_j^{(m)} + b_j^{(n)}\right)$,

if $s \in \{2,3,\ldots,k+1\}$. The indexes $b_0^{(m)}$ and $b_0^{(n)}$ are defined to be zeros, $b_0^{(m)} := 0$, $b_0^{(n)} := 0$, for completeness in conditions **(i)-(v)**.

Given the joint distribution of L and K the marginal distribution of L under H_0 can be easily derived.

Acknowledgements: This work was supported by the grant I02/19 of the Bulgarian National Science Fund.

References

[1] Critchlow D. (1985). *Metric Methods for Analyzing Partially Ranked Data*. Lecture Notes in Statist.,No. **34**, Springer, New York.

[2] Critchlow D. E. (1986). A Unified Approach to Constructing Nonparametric Tests, Tech. Report. No. **86-15**, Dept. of Statistics, Purdue University, Indiana.

[3] Critchlow D. E. (1992). On rank statistics: An Approach via Metrics on the Permutation Group, *J. Statist. Plann. Inference*, Vol. **32**, pp. 325-346.

[4] Deza M., Huang T.(1998). Metrics on Permutations, a Survey, *Journal of Combinatoric, Information and System Sciences*, Vol. **23**, pp. 173-185.

[5] Diaconis P. (1988). *Group Representations in Probability and Statistics*. IMS Lecture Notes - Monograph Series, Vol. **11**, Hayward, Carifornia.

[6] Stoimenova E.(2000). Rank Tests Based on Exceeding Observations, *Ann. Inst. Stat. Math.*, Vol. **52-2**, pp. 255-266.