7. Independence Problem

The problem of interest

- Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a sample of random variable pairs.
- Each pair (X_i, Y_i) is observed from subject i and has a bivariate distribution with a joint cdf $F(x, y)$, $i = 1, \ldots, n$.
- The question of interest in this section is whether $X_1, \ldots, X_n$ are independent of $Y_1, \ldots, Y_n$, or equivalently, $F(x, y) = F_X(x)F_Y(y)$ for all $x, y \in \mathbb{R}$, where F_X and F_Y are the marginal cdf's of X_i and Y_i, respectively.
- If independence is accepted, then we can treat $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ as two independent samples with their respective marginal distributions; otherwise a joint distribution is needed to draw inference.
- An example of possible dependence between X_i and Y_i is that they represent the lifetimes of two persons genetically connected.

7.1 Sign tests of independence

Assumption 7.1: The pairs of random variables $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. with a continuous bivariate cdf $F(x, y)$.

Null hypothesis: The null hypothesis assumes that $X_1, \ldots, X_n$ are independent of $Y_1, \ldots, Y_n$. It can be formally expressed by

$$H_0 : F(x, y) = F(x)G(y) \text{ for all } x, y \in \mathbb{R}, \tag{7.1}$$

where $F(x)$ and $G(y)$ are the marginal cdf's of X's and Y's, respectively.

Kendall correlation coefficient: This is also called *Kendall's tau*, defined by

$$\begin{aligned} \tau &= \Pr\big((X_1 - X_2)(Y_1 - Y_2) > 0\big) - \Pr\big((X_1 - X_2)(Y_1 - Y_2) < 0\big) \\ &= 2\Pr\big((X_1 - X_2)(Y_1 - Y_2) > 0\big) - 1 \end{aligned} \tag{7.2}$$

Because $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. and continuous,

$$\Pr\big(X_1 > X_2\big) = \Pr\big(X_1 < X_2\big) = \frac{1}{2} \quad \text{and} \quad \Pr\big(Y_1 > Y_2\big) = \Pr\big(Y_1 < Y_2\big) = \frac{1}{2} \tag{7.3}$$

Alternative hypotheses

If the null hypothesis H_0 in (7.1) is true, then by (7.2) and (7.3),

$$\begin{aligned}\tau &= 2[\Pr(X_1 > X_2, Y_1 > Y_2) + \Pr(X_1 < X_2, Y_1 < Y_2)] - 1 \\ &= 2\Pr(X_1 > X_2)\Pr(Y_1 > Y_2) + 2\Pr(X_1 < X_2)\Pr(Y_1 < Y_2) - 1 \\ &= 2\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) + 2\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) - 1 = \frac{1}{2} + \frac{1}{2} - 1 = 0\end{aligned}$$

Thus $\tau = 0$ under H_0 and $\tau \neq 0$ indicates dependence between (X_i, Y_i), $i = 1, \ldots, n$. So we will consider the following alternative hypotheses:

$$H_1: \tau > 0, \quad H_1: \tau < 0 \quad \text{and} \quad H_1: \tau \neq 0. \tag{7.4}$$

Remark 7.1 While H_0 implies $\tau = 0$, the converse is not true. It is possible to have $\tau = 0$ even if X_i and Y_i are dependent (H_0 is false). Thus we will keep the null hypothesis H_0 as in (7.1) instead of replacing it by $H_0: \tau = 0$, and consider the alternative hypotheses in (7.4).

Test statistic: For $1 \le u \ne v \le n$, define

$$Q_{uv} = Q_{vu} = \begin{cases} 1 & \text{if } (X_u - X_v)(Y_u - Y_v) > 0 \\ -1 & \text{if } (X_u - X_v)(Y_u - Y_v) < 0 \end{cases} \tag{7.5}$$

Then the *Kendall statistic* K to test H_0 is defined by

$$K = \sum_{u<v}^{n} Q_{uv} = \sum_{v=2}^{n} \sum_{u=1}^{v-1} Q_{uv} = \sum_{u=1}^{n-1} \sum_{v=u+1}^{n} Q_{uv} \tag{7.6}$$

Calculation of K: Assume no ties. Let $(R_1, \ldots, R_n)$ be the ranks of $X_1, \ldots, X_n$ and $(S_1, \ldots, S_n)$ the ranks of $Y_1, \ldots, Y_n$ (in ascending order). Then

$$Q_{uv} = \begin{cases} 1 & \text{if } (R_u - R_v)(S_u - S_v) > 0 \\ -1 & \text{if } (R_u - R_v)(S_u - S_v) < 0 \end{cases}$$

Thus $Q_{uv} = 1$ if (R_u, R_v) and (S_u, S_v) are in the same order; otherwise $Q_{uv} = -1$. We can arrange $(R_1, \ldots, R_n) = (1, \ldots, n)$ and consider each of the $n!$ permutations $(S_1, \ldots, S_n)$ of $(1, \ldots, n)$. Then for $u < v$, $Q_{uv} = 1$ if $S_u < S_v$; $Q_{uv} = -1$ if $S_u > S_v$.

Let $S_u < S_v$ ($S_u > S_v$) denote a pair (S_u, S_v) with $1 \le u < v \le n$ and $S_u < S_v$ ($S_u > S_v$). The total number of pairs $u < v$ is $n(n-1)/2$. Then we can calculate K by

$$K = (\text{No. of pairs } S_u < S_v) - (\text{No. of pairs } S_u > S_v) \tag{7.7}$$

For example, if $(X_1, Y_1), \ldots, (X_n, Y_n) = (1,-6), (4,3), (-1,-2), (2,5), (-3,8)$, then we arrange the data as $(-3,8), (-1,-2), (1,-6), (2,5), (4,3)$, so that

$$(R_1, \ldots, R_5) = (1,2,3,4,5) \quad \text{and} \quad (S_1, \ldots, S_5) = (5,2,1,4,3)$$

There are 4 pairs $S_u < S_v$: (2,4), (2,3), (1,4), (1,3), and the total number of pairs $u < v$ is $5(4)/2 = 10$. Hence by (7.7), $K = 4 - (10 - 4) = 4 - 6 = -2$.

Null distribution of K

Under H_0, each permutation $(S_1, \ldots, S_n)$ of $(1, \ldots, n)$ is equally likely to occur with probability $1/n!$. Hence the null distribution of K is given by:

$$\Pr(K = k) = \frac{\text{No. of permutations } (S_1, \ldots, S_n) : K = k}{n!} \tag{7.8}$$

Example 7.1 Let $n=4$. Then $n!=4!=24$, $(R_1,R_2,R_3,R_4)=(1,2,3,4)$ and the total number of pairs $u<v$ is $4(3)/2=6$.

The distribution of K is calculated in the following table:

$K=k$	(S_1,S_2,S_3,S_4)	$\Pr(K=k)$
$K=6-0=6$	$(1,2,3,4)$	$1/24$
$K=5-1=4$	$(2,1,3,4)$, $(1,3,2,4)$, $(1,2,4,3)$	$3/24$
$K=4-2=2$	$(2,3,1,4)$, $(2,1,4,3)$, $(3,1,2,4)$, $(1,3,4,2)$, $(1,4,2,3)$	$5/24$
$K=3-3=0$	$(3,2,1,4)$, $(2,3,4,1)$, $(2,4,1,3)$, $(3,1,4,2)$, $(1,4,3,2)$ $(4,1,2,3)$	$6/24$
$K=2-4=-2$	$(4,1,3,2)$, $(3,4,1,2)$, $(4,2,1,3)$, $(2,4,3,1)$, $(3,2,4,1)$	$5/24$
$K=1-5=-4$	$(4,3,1,2)$, $(4,2,3,1)$, $(3,4,2,1)$	$3/24$
$K=0-6=-6$	$(4,3,2,1)$	$1/24$

Hence $\Pr(K\geq 6)=1/24$, $\Pr(K\geq 4)=(3+1)/24=4/24$, and so on.

Mean of K

Let $1 \le u \ne v \le n$. By the i.i.d. assumptions of $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$,

$$\begin{aligned} \mathrm{E}[Q_{uv}] &= \mathrm{E}[Q_{12}] = \Pr(Q_{12} = 1) - \Pr(Q_{12} = -1) \\ &= \Pr\big((X_1 - X_2)(Y_1 - Y_2) > 0\big) - \Pr\big((X_1 - X_2)(Y_1 - Y_2) < 0\big) \\ &= \tau \quad \text{(Kendall's tau)}, \end{aligned} \tag{7.9}$$

It follows that the mean of K is given by

$$\mathrm{E}[K] = \sum_{u<v}^{n} \mathrm{E}[Q_{uv}] = \binom{n}{2}\tau = \frac{n(n-1)}{2}\tau \tag{7.10}$$

In particular, $\mathrm{E}_0[K] = 0$ under the null hypothesis H_0 of independence.

By (7.10), an unbiased estimator of τ is given by

$$\bar{K} = \frac{2K}{n(n-1)} \Rightarrow \mathrm{E}[\bar{K}] = \tau \tag{7.11}$$

Variance of K

The variance of K is

$$\mathrm{Var}(K) = \mathrm{Var}\left(\sum_{u<v}^{n} Q_{uv}\right) = \sum_{u<v}^{n} \mathrm{Var}(Q_{uv}) + \sum_{\substack{s<u;t<v \\ (s,u)\neq(t,v)}} \mathrm{Cov}(Q_{su}, Q_{tv}) \qquad (7.12)$$

By the i.i.d. assumptions of $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$, the definition of Q_{ij} in (7.5) implies that for $1 \le u \ne v \le n$,

$$\Pr(Q_{uv}^2 = 1) = 1 \;\Rightarrow\; \mathrm{E}[Q_{uv}^2] = 1, \qquad (7.13)$$

$$\mathrm{Var}(Q_{uv}) = \mathrm{Var}(Q_{12}) = \mathrm{E}[Q_{12}^2] - \left(\mathrm{E}[Q_{12}]\right)^2 = 1 - \tau^2, \qquad (7.14)$$

$$\mathrm{Cov}(Q_{tu}, Q_{tv}) = \mathrm{Cov}(Q_{tu}, Q_{uv}) = \mathrm{Cov}(Q_{tv}, Q_{uv}) = \mathrm{Cov}(Q_{12}, Q_{13}) \qquad (7.15)$$

for $1 \le t < u \ne v \le n$, and

$$\mathrm{Cov}(Q_{su}, Q_{tv}) = 0 \;\text{ for distinct } s, t, u, v \in \{1, \ldots, n\}. \qquad (7.16)$$

By (7.15) – (7.16), the last sum of covariances (over s,t,u,v) in (7.12) equals

$$\sum_{t<u\neq v} \mathrm{Cov}(Q_{tu},Q_{tv}) + \sum_{t<u<v} \mathrm{Cov}(Q_{tu},Q_{uv}) + \sum_{t\neq u<v} \mathrm{Cov}(Q_{tv},Q_{uv})$$

$$= \sum_{t<u\neq v} 3\mathrm{Cov}(Q_{tu},Q_{tv}) = 3\sum_{t=1}^{n-2} \sum_{t<u\neq v}^{n} \mathrm{Cov}(Q_{tu},Q_{tv}) \qquad (7.17)$$

In (7.17), $t<u\neq v$ in the first sum is a shorthand of $t<u$, $t<v$, $u\neq v$; similarly, $t\neq u<v$ represents $t\neq u$, $t<v$, $u<v$. Moreover, $\sum_{t<u\neq v}$, $\sum_{t<u<v}$ and $\sum_{t\neq u<v}$ are triple sums over t,u,v, while $\sum_{t<u\neq v}^{n}$ is a double sum over u,v for a given t.

By (7.2),

$$\Pr(Q_{12}=1) = \Pr\big((X_1-X_2)(Y_1-Y_2)>0\big) = \frac{1+\tau}{2}$$

and

$$\Pr(Q_{12}=-1) = 1-\Pr(Q_{12}=1) = 1-\frac{1+\tau}{2} = \frac{1-\tau}{2}$$

Define

$$\delta = \Pr(Q_{12} = 1, Q_{13} = 1) \tag{7.18}$$

Then

$$\Pr(Q_{12} = 1, Q_{13} = -1) = \Pr(Q_{12} = 1) - \Pr(Q_{12} = 1, Q_{13} = 1) = \frac{1+\tau}{2} - \delta$$
$$= \Pr(Q_{12} = -1, Q_{13} = 1)$$

and

$$\Pr(Q_{12} = -1, Q_{13} = -1) = \Pr(Q_{12} = -1) - \Pr(Q_{12} = -1, Q_{13} = 1)$$
$$= \frac{1-\tau}{2} - \left(\frac{1+\tau}{2} - \delta\right) = \delta - \tau$$

It follows that

$$\Pr(Q_{12}Q_{13} = 1) = \Pr(Q_{12} = 1, Q_{13} = 1) + \Pr(Q_{12} = -1, Q_{13} = -1) = 2\delta - \tau$$

and

$$\Pr(Q_{12}Q_{13} = -1) = 2\Pr(Q_{12} = 1, Q_{13} = -1) = 2\left(\frac{1+\tau}{2} - \delta\right) = 1 + \tau - 2\delta$$

Consequently,

$$\mathrm{E}[Q_{12}Q_{13}] = \Pr(Q_{12}Q_{13} = 1) - \Pr(Q_{12}Q_{13} = -1) = 4\delta - 1 - 2\tau \tag{7.19}$$

and

$$\mathrm{Cov}(Q_{12}, Q_{13}) = \mathrm{E}[Q_{12}Q_{13}] - \tau^2 = 4\delta - 1 - 2\tau - \tau^2 = 4\delta - (1+\tau)^2 \tag{7.20}$$

Combine (7.12), (7.14) – (7.17) and (7.20), we obtain

$$\begin{aligned}
\mathrm{Var}(K) &= \mathrm{Var}\left(\sum_{u<v}^{n} Q_{uv}\right) = \sum_{u<v}^{n} \mathrm{Var}(Q_{uv}) + 3\sum_{t=1}^{n-2} \sum_{t<u\neq v}^{n} \mathrm{Cov}(Q_{tu}, Q_{tv}) \\
&= \sum_{u<v}^{n} (1-\tau^2) + 3\sum_{t=1}^{n-2} \sum_{t<u\neq v}^{n} \left[4\delta - (1+\tau)^2\right] \\
&= \frac{n(n-1)}{2}(1-\tau^2) + 3\left[4\delta - (1+\tau)^2\right]\sum_{t=1}^{n-2}(n-t)(n-t-1)
\end{aligned} \tag{7.21}$$

In (7.21), given $t \in \{1, 2, \ldots, n-2\}$, the sum over $u \neq v > t$ has $(n-t)(n-t-1)$ terms (the number of pairs $u \neq v$ taken from $n-t$ numbers $\{t+1, \ldots, n\}$).

Let $k=n-t$. Then $t=1 \Rightarrow k=n-1$ and $t=n-2 \Rightarrow k=2$. Hence

$$\sum_{t=1}^{n-2}(n-t)(n-t-1)=\sum_{k=2}^{n-1}k(k-1)=\sum_{k=1}^{n-1}\left(k^2-k\right)=\sum_{k=1}^{n-1}k^2-\sum_{k=1}^{n-1}k$$

$$=\frac{n(n-1)(2n-1)}{6}-\frac{n(n-1)}{2}=\frac{n(n-1)(2n-1-3)}{6}$$

$$=\frac{n(n-1)(2n-4)}{6}=\frac{n(n-1)(n-2)}{3} \qquad (7.22)$$

Substituting (7.22) into (7.21) leads to

$$\operatorname{Var}(K)=\frac{n(n-1)}{2}\left(1-\tau^2\right)+n(n-1)(n-2)\left[4\delta-(1+\tau)^2\right]$$

$$=\frac{n(n-1)}{2}\left\{1-\tau^2+2(n-2)\left[4\delta-(1+\tau)^2\right]\right\} \qquad (7.23)$$

and

$$\operatorname{Var}_0(K)=\frac{n(n-1)}{2}\left\{1+2(n-2)[4\delta-1]\right\} \text{ under } H_0 \qquad (7.24)$$

Null mean and variance of K

Under the null hypothesis H_0, $\tau = 0$. Hence by (7.10),

$$\mathrm{E}_0[K] = \frac{n(n-1)}{2}\tau = 0 \tag{7.25}$$

Next, since X_1, X_2, X_3 are i.i.d.,

$$\Pr(X_i < X_j < X_k) = \frac{1}{6} \text{ for all permutations } (i, j, k) \text{ of } (1,2,3) \tag{7.26}$$

It follows that

$$\Pr(X_1 > X_2, X_1 > X_3) = \Pr(X_2 < X_3 < X_1) + \Pr(X_3 < X_2 < X_1) = \frac{1}{3} \tag{7.27}$$

and

$$\Pr(X_1 < X_2, X_1 < X_3) = \Pr(X_1 < X_2 < X_3) + \Pr(X_1 < X_3 < X_2) = \frac{1}{3} \tag{7.28}$$

The results in (7.26) – (7.28) also hold with Y_1, Y_2, Y_3 in place of X_1, X_2, X_3.

Then by (7.26) – (7.28) and the independence between X_1, X_2, X_3 and Y_1, Y_2, Y_3 under H_0, the δ defined in (7.18) is calculated by

$$\begin{aligned}
\delta &= \Pr(Q_{12}=1, Q_{13}=1) = \Pr\big((X_1-X_2)(Y_1-Y_2)>0, (X_1-X_3)(Y_1-Y_3)>0\big) \\
&= \Pr(X_1 > X_2, X_1 > X_3)\Pr(Y_1 > Y_2, Y_1 > Y_3) \\
&\quad + \Pr(X_2 < X_1 < X_3)\Pr(Y_2 < Y_1 < Y_3) + \Pr(X_3 < X_1 < X_2)\Pr(Y_3 < Y_1 < Y_2) \\
&\quad + \Pr(X_1 < X_2, X_1 < X_3)\Pr(Y_1 < Y_2, Y_1 < Y_3) \\
&= \left(\frac{1}{3}\right)^2 + \left(\frac{1}{6}\right)^2 + \left(\frac{1}{6}\right)^2 + \left(\frac{1}{3}\right)^2 = \frac{4+1+1+4}{36} = \frac{10}{36} = \frac{5}{18} \qquad (7.29)
\end{aligned}$$

It follows from (7.24) and (7.29) that

$$\begin{aligned}
\mathrm{Var}_0(K) &= \frac{n(n-1)}{2}\left\{1 + 2(n-2)\left[4\times\frac{5}{18} - 1\right]\right\} = \frac{n(n-1)}{2}\left[1 + \frac{2}{9}(n-2)\right] \\
&= \frac{n(n-1)}{18}(9 + 2n - 4) = \frac{n(n-1)(2n+5)}{18} \qquad (7.30)
\end{aligned}$$

Asymptotic distribution of K

By the central limit theorem together with (7.25) and (7.30),

$$K^* = \frac{K - \mathrm{E}_0[K]}{\sqrt{\mathrm{Var}_0(K)}} = \frac{K}{\sqrt{n(n-1)(2n+5)/18}} \to_d N(0,1) \text{ as } n \to \infty \qquad (7.31)$$

under H_0 in (7.1), where "$\to_d$" denotes convergence in distribution.

Average paired sign

An equivalent version of the Kendall statistic K is the average over paired sign statistics $\{Q_{uv}, u < v\}$:

$$\bar{K} = \frac{2}{n(n-1)} \sum_{u<v}^{n} Q_{uv} = \frac{2K}{n(n-1)} \qquad (7.32)$$

By (7.11) and (7.30),

$$\mathrm{E}[\bar{K}] = \tau, \quad \mathrm{E}_0[\bar{K}] = 0 \text{ and } \mathrm{Var}_0(\bar{K}) = \frac{2(2n+5)}{9n(n-1)} \qquad (7.33)$$

Rejection rule

For the null hypothesis H_0 stated in (7.1), the *Kendall test* for independence at level α is given by the following rules:

- Reject H_0 for $H_1: \tau > 0$ if $\bar{K} \geq k_\alpha$;
- Reject H_0 for $H_1: \tau < 0$ if $\bar{K} \leq -k_\alpha$;
- Reject H_0 for $H_1: \tau \neq 0$ if $|\bar{K}| \geq k_{\alpha/2}$, where $\Pr\left(\bar{K} \geq k_\alpha\right) = \alpha$ under H_0.

In Example 7.1, $\Pr\left(\bar{K} \geq 2(6)/4(3) = 1\right) = \Pr\left(K \geq 6\right) = 1/24 \;\; \Rightarrow \;\; k_{1/24} = 1$.

Approximate rejection rule

By (7.31), the approximate rules to test H_0 at level α are as follows:

- Reject H_0 for $H_1: \tau > 0$ if $K^* \geq z_\alpha$;
- Reject H_0 for $H_1: \tau < 0$ if $K^* \leq -z_\alpha$;
- Reject H_0 for $H_1: \tau \neq 0$ if $\left|K^*\right| \geq z_{\alpha/2}$.

Example 7.2 Table 8.1 in Example 8.1 of the textbook (page 398) presents paired data (X_i, Y_i), where X_i is the Hunter L lightness value and Y_i is the average of 80 panel scores for lot i of canned tuna, $i = 1, \ldots, 9$.

The following table shows the original and rearranged data (X_i, Y_i) (in increasing order of X_i), and the ranks (R_i, S_i) of the rearranged data:

Original data		Rearranged data		Ranks	
X_i	Y_i	X_i	Y_i	R_i	S_i
44.4	2.6	41.9	2.5	1	1
45.9	3.1	44.1	4.0	2	7
41.9	2.5	44.4	2.6	3	2
53.3	5.0	44.7	3.6	4	5
44.7	3.6	45.2	2.8	5	3
44.1	4.0	45.9	3.1	6	4
50.7	5.2	50.7	5.2	7	9
45.2	2.8	53.3	5.0	8	8
60.1	3.8	60.1	3.8	9	6

The question of interest is whether the Hunter L value (a measure of quality) is positively correlated with the panel score (representing consumer preference). We can apply the Kendall test of H_0 against $H_1: \tau > 0$ for this question.

From the column for S_i in the above table, we can find 10 pairs $S_u > S_v$ with $u < v$ ($\Rightarrow$ $Q_{uv} = -1$):

$$(S_u, S_v) = (7,2), (7,5), (7,3), (7,4), (7,6), (5,3), (5,4), (9,8), (9,6), (8,6)$$

The total number of $u < v$ pairs is $9(8)/2 = 36$. Hence $K = (36-10)-10 = 16$ and $\bar{K} = 16/36 = 4/9$. By R, the p-value for $\tau > 0$ is $\Pr(K \geq 16) = \Pr\left(\bar{K} \geq 4/9\right) = 0.060$.

If we use the large-sample approximation, then

$$K^* = \frac{K}{\sqrt{n(n-1)(2n+5)/18}} = \frac{16}{\sqrt{9(8)(23)/18}} = 1.668$$

Hence the approximate p-value for $\tau > 0$ is $\Pr(Z \geq 1.668) = 0.048$ ($Z \sim N(0,1)$). Both p-values point to moderate evidence (not very strong) that the Hunter L value is positively correlated with the panel score.

Ties: If there are ties among $\{X_1,\ldots,X_n\}$ and/or $\{Y_1,\ldots,Y_n\}$, we define

$$Q_{uv} = Q_{vu} = \begin{cases} 1 & \text{if } (X_u - X_v)(Y_u - Y_v) > 0 \\ 0 & \text{if } (X_u - X_v)(Y_u - Y_v) = 0 \\ -1 & \text{if } (X_u - X_v)(Y_u - Y_v) < 0 \end{cases}$$

The definition of K in (7.6) remains valid.

Then the rejection rules based on k_α from the exact distribution of K with no ties can still be applied, but with an approximate level α.

To apply the approximate rejection rules using K^* based on z_α, the null mean $\mathrm{E}_0[K]$ is not affected by ties, but the null variance $\mathrm{Var}_0(K)$ should be adjusted to equation (8.18) on page 397 of the textbook.

The exact distribution of K conditional on observed ties can be worked out by the same method of enumeration as in the case with no ties. Then the critical point for the exact level α of significance can be determined accordingly.

Example 7.3 Consider $n=4$ with $n!=4!=24$. Let $(R_1,R_2,R_3,R_4)=(1,2,3,4)$ and (S_1,S_2,S_3,S_4) a permutation of $(1,2.5,2.5,4)$. Then conditional on the ties in (Y_1,Y_2,Y_3,Y_4), the distribution of K under H_0 is calculated below:

$K=k$	(S_1,S_2,S_3,S_4)	$\Pr(K=k)$
$K=5-0=5$	$(1,2.5,2.5,4)$, $(1,2.5,2.5,4)$	$2/24$
$K=4-1=3$	$(2.5,1,2.5,4)\times 2$, $(1,2.5,4,2.5)\times 2$	$4/24$
$K=3-2=1$	$(2.5,2.5,1,4)\times 2$, $(1,4,2.5,2.5)\times 2$, $(2.5,1,4,2.5)\times 2$	$6/24$

and $\Pr(K=k)=\Pr(K=-k)$ for $k=-1,-3,-5$. $\mathrm{E}_0[K]=0$ by symmetry, and

$$\mathrm{Var}_0(K)=\frac{5^2(2)+3^2(4)+1^2(6)}{24}\times 2=\frac{92}{12}=\frac{23}{3}<\frac{26}{3}=\frac{12(13)}{18}\quad\text{(with no ties)}$$

The same result follows from equation (8.18) on page 397 of the textbook:

$$\mathrm{Var}_0(K)=\frac{4(3)(8+5)-2(1)(4+5)}{18}=\frac{156-18}{18}=\frac{26-3}{3}=\frac{23}{3}$$

Calculations of Kendall statistic with ties

Case 1. There are ties among $(Y_1,\ldots,Y_n)$, but not among $(X_1,\ldots,X_n)$.

Then we can keep $(R_1,\ldots,R_n)=(1,2,\ldots,n)$ and calculate K by (7.7):

$$K=\text{No.}\{Q_{uv}=1\}-\text{No.}\{Q_{uv}=-1\}=\text{No.}\{u<v:S_u<S_v\}-\text{No.}\{u<v:S_u>S_v\}$$

But because $Q_{uv}=0$ for $u<v$ if $Y_u=Y_v$,

$$\text{No.}\{Q_{uv}=-1\}=\frac{n(n-1)}{2}-\text{No.}\{Q_{uv}=1\}-\text{No.}\{Q_{uv}=0\}\ \Rightarrow$$

$$K=\text{No.}\{Q_{uv}=1\}-\left[\frac{n(n-1)}{2}-\text{No.}\{Q_{uv}=1\}-\text{No.}\{Q_{uv}=0\}\right]$$

$\text{No.}\{Q_{uv}=0\}$ can be determined as follows: each tied group of size t contributes $t(t-1)/2$ to $\text{No.}\{Q_{uv}=0\}$. For example, if $(R_1,\ldots,R_6)=(1,\ldots,6)$ and $(S_1,\ldots,S_6)$ $=(2,5.5,2,4,2,5.5)$, then $\text{No.}\{Q_{uv}=0\}=2(1)/2+3(2)/2=1+3=4$ and

$$\text{No.}\{Q_{uv}=1\}=3+2+1+1=7\ \Rightarrow\ K=7-\left[\frac{6(5)}{2}-7-4\right]=7-4=3$$

Case 2. There are ties among $(X_1,\ldots,X_n)$, but not among $(Y_1,\ldots,Y_n)$.

This case can be handled in the same way as Case 1 by switching $(X_1,\ldots,X_n)$ and $(Y_1,\ldots,Y_n)$, or $(R_1,\ldots,R_n)$ and $(S_1,\ldots,S_n)$.

Case 3. There are ties among both $(X_1,\ldots,X_n)$ and $(Y_1,\ldots,Y_n)$.

In this case, we can take $(R_1,\ldots,R_n)$ in nondecreasing order and skip tied ranks in $(R_1,\ldots,R_n)$. For example, if

$$(R_1,\ldots,R_8)=(1,2.5,2.5,4,5,7,7,7) \text{ and } (S_1,\ldots,S_8)=(2,4,7,1,4,8,6,4),$$

we skip pairs $uv=23$, 67, 68, 78 and count

$Q_{uv}=1$ for $uv=12$, 13, 15, 16, 17, 18, ~~23~~, 26, 27, 36, 45, 46, 47, 48, 56, 57;

$Q_{uv}=-1$ for $uv=14$, 24, 34, 35, 37, 38, ~~67~~, ~~68~~, ~~78~~

$$\Rightarrow \text{No.}\{Q_{uv}=1\}=6+2+1+4+2=15 \quad \text{and} \quad \text{No.}\{Q_{uv}=-1\}=6$$

Thus $K=15-6=9$ (if there were no ties in X_i's, K would be $16-9=7$).

7.2 Estimation of Kendall correlation coefficient

An estimator of the Kendall correlation coefficient τ is given by

$$\hat{\tau} = \bar{K} = \frac{2K}{n(n-1)} \tag{7.34}$$

By (7.11) or (7.33), $\mathrm{E}[\hat{\tau}] = \mathrm{E}[\bar{K}] = \tau$. Hence $\hat{\tau}$ is an unbiased estimator of τ.

The variance of $\hat{\tau}$ can be obtained from (7.23) as

$$\mathrm{Var}(\hat{\tau}) = \frac{4\mathrm{Var}(K)}{n^2(n-1)^2} = \frac{4}{n(n-1)}\left\{\frac{1-\tau^2}{2} + (n-2)\left[4\delta - (1+\tau)^2\right]\right\}, \tag{7.35}$$

where δ is defined in (7.18):

$$\delta = \Pr(Q_{12} = 1, Q_{13} = 1) = \Pr\big((X_1 - X_2)(Y_1 - Y_2) > 0, (X_1 - X_3)(Y_1 - Y_3) > 0\big)$$

To estimate $\mathrm{Var}(\hat{\tau})$, let

$$C_i = \sum_{t \neq i}^{n} Q_{it}, \quad i = 1, \ldots, n, \quad \text{and} \quad \bar{C} = \frac{1}{n}\sum_{i=1}^{n} C_i \tag{7.36}$$

It is easy to see that

$$\sum_{i=1}^{n} C_i = \sum_{i=1}^{n}\sum_{t\neq i}^{n} Q_{it} = 2\sum_{1\le i<t\le n} Q_{it} = 2K \;\Rightarrow\; \overline{C} = \frac{2K}{n} = (n-1)\hat{\tau} \tag{7.37}$$

Hence

$$\sum_{i=1}^{n}(C_i-\overline{C})^2 = \sum_{i=1}^{n} C_i^2 - n\overline{C}^2 = \sum_{i=1}^{n} C_i^2 - \frac{4}{n}K^2 = \sum_{i=1}^{n} C_i^2 - n(n-1)^2\hat{\tau}^2 \tag{7.38}$$

For C_i defined in (7.36), since $Q_{it}^2 = 1$ if $i \neq t$,

$$C_i^2 = \sum_{s\neq i}^{n}\sum_{t\neq i}^{n} Q_{is}Q_{it} = \sum_{t\neq i}^{n} Q_{it}^2 + \sum_{s\neq t\neq i}^{n} Q_{is}Q_{it} = n-1+\sum_{s\neq t\neq i}^{n} Q_{is}Q_{it},\quad i=1,\ldots,n.$$

This together with (7.38) yields

$$\sum_{i=1}^{n}(C_i-\overline{C})^2 = n(n-1) + \sum_{i=1}^{n}\sum_{s\neq t\neq i}^{n} Q_{is}Q_{it} - n(n-1)^2\hat{\tau}^2, \tag{7.39}$$

where the sum over $1 \le s \neq t \neq i \le n$ has $n(n-1)(n-2) = O(n^3)$ terms of $Q_{is}Q_{it}$.

By the i.i.d. assumption of $(X_1, Y_1), \ldots, (X_n, Y_n)$, $Q_{is}Q_{it} \sim Q_{12}Q_{13}$ for $s \neq t \neq i$ and $Q_{is}Q_{it}$ is independent of $Q_{ju}Q_{jv}$ if $i \neq j, s \neq u$ and $t \neq u$. Thus

$$\lim_{n\to\infty} \frac{1}{n(n-1)^2} \sum_{i=1}^{n} \sum_{s\neq t\neq i}^{n} Q_{is}Q_{it} = \mathrm{E}[Q_{12}Q_{13}] \tag{7.40}$$

in probability by the law of large numbers. Similarly,

$$\lim_{n\to\infty} \hat{\tau} = \lim_{n\to\infty} \frac{2K}{n(n-1)} = \lim_{n\to\infty} \frac{2}{n(n-1)} \sum_{u<v}^{n} Q_{uv} = \mathrm{E}[Q_{12}] = \tau \tag{7.41}$$

in probability. Combine (7.39) – (7.41) with (7.20), we obtain

$$\lim_{n\to\infty} \frac{1}{n(n-1)^2} \sum_{i=1}^{n} (C_i - \overline{C})^2 = \lim_{n\to\infty} \left[\frac{1}{n-1} + \frac{1}{n(n-1)^2} \sum_{i=1}^{n} \sum_{s\neq t\neq i}^{n} Q_{is}Q_{it} - \hat{\tau}^2 \right]$$

$$= \mathrm{E}[Q_{12}Q_{13}] - \tau^2 = \mathrm{Cov}(Q_{12}, Q_{13}) = 4\delta - (1+\tau)^2 \tag{7.42}$$

in probability.

It follows from (7.42) that $4\delta-(1+\tau)^2$ can be consistently estimated by

$$\frac{1}{n(n-1)^2}\sum_{i=1}^{n}(C_i-\overline{C})^2 \tag{7.43}$$

Substitute (7.43) for $4\delta-(1+\tau)^2$ in (7.35), an estimator of $\mathrm{Var}(\hat{\tau})$ is given by

$$\hat{\sigma}^2=\frac{2}{n(n-1)}\left[\frac{2(n-2)}{n(n-1)^2}\sum_{i=1}^{n}(C_i-\overline{C})^2+1-\hat{\tau}^2\right] \tag{7.44}$$

By (7.35), (7.42), (7.44) and the central limit theorem,

$$\lim_{n\to\infty}\frac{\hat{\sigma}^2}{\mathrm{Var}(\hat{\tau})}=1 \quad \text{and} \quad \frac{\hat{\tau}-\tau}{\hat{\sigma}}\to_d N(0,1) \quad \text{as } n\to\infty. \tag{7.45}$$

Hence an approximate $100(1-\alpha)\%$ confidence interval of τ is given by

$$(\tau_L,\tau_U)=\hat{\tau}\pm z_{\alpha/2}\hat{\sigma}=\left(\hat{\tau}-z_{\alpha/2}\hat{\sigma},\hat{\tau}+z_{\alpha/2}\hat{\sigma}\right) \tag{7.46}$$

Example 7.4 In Example 7.2, since $n=9$ and $K=16$, by (7.34),

$$\hat{\tau}=\frac{2K}{n(n-1)}=\frac{2(16)}{9(8)}=\frac{4}{9}$$

To estimate the variance of $\hat{\tau}$ and obtain confidence interval of τ, let

$$C_i^+ = \text{No. of } j<i:\ S_j<S_i \text{ and } j>i:\ S_j>S_i;\quad C_i^- = n-1-C_i^+,\ i=1,\ldots,n.$$

Then

$$C_i = C_i^+ - C_i^- = C_i^+ - \left(n-1-C_i^+\right) = 2C_i^+ - n+1,\ i=1,\ldots,n.$$

In Example 7.2, $(S_1,\ldots,S_9)=(1,7,2,5,3,4,9,8,6)$. Hence $n-1=8$,

$S_1=1<S_j$ for $j=2,\ldots,9 \Rightarrow C_1^+=8,\ C_1^-=9-1-8=0 \Rightarrow C_1=8-0=8$

$S_1=1<S_2=7<9,8\ (S_7,S_8) \Rightarrow C_2^+=3,\ C_2^-=8-3=5 \Rightarrow C_2=3-5=-2$

$S_1=1<S_3=2<S_j$ for $j=4,5,\ldots,9 \Rightarrow C_3^+=7 \Rightarrow C_3=7-1=6$

Similarly,

$$C_4 = 5-3=2,\ C_5 = C_6 = C_7 = C_8 = 6-2=4 \ \text{ and } \ C_9 = 5-3=2.$$

Thus by (7.38),

$$\sum_{i=1}^{n}(C_i-\overline{C})^2 = \sum_{i=1}^{9} C_i^2 - \frac{4}{9}K^2 = 8^2+(-2)^2+6^2+2\times 2^2+4\times 4^2-\frac{4}{9}\times 16^2$$

$$=176-\frac{1024}{9}=\frac{1584-1024}{9}=\frac{560}{9}$$

Then $\text{Var}(\hat{\tau})$ is estimated by (7.44):

$$\hat{\sigma}^2 = \frac{2}{9(8)}\left[\frac{2(7)}{9(8)^2}\times\frac{560}{9}+1-\left(\frac{4}{9}\right)^2\right]=0.0643$$

By (7.46), an approximate 90% confidence interval of τ is

$$(\tau_L,\tau_U)=\hat{\tau}\pm z_{0.05}\hat{\sigma}=\frac{4}{9}\pm 1.645\sqrt{0.0643}=(0.0273, 0.8616)$$

7.3 Rank tests of independence

Assumption 7.1 and the null hypothesis in (7.1) remain valid. Assume no ties. Define the *Spearman rank correlation coefficient* as

$$r_s = \frac{12}{n(n^2-1)} \sum_{i=1}^{n} \left(R_i - \frac{n+1}{2}\right)\left(S_i - \frac{n+1}{2}\right) = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} D_i^2, \qquad (7.47)$$

where (R_i, S_i) are the ranks of (X_i, Y_i) and $D_i = S_i - R_i$, $i = 1, \ldots, n$. Note that

$$\sum_{i=1}^{n} \left(R_i - \frac{n+1}{2}\right)\left(S_i - \frac{n+1}{2}\right) = \sum_{i=1}^{n} \left[R_i S_i - \frac{n+1}{2}(R_i + S_i) + \left(\frac{n+1}{2}\right)^2 \right]$$

$$= \sum_{i=1}^{n} R_i S_i - \frac{n+1}{2}\left[\frac{n(n+1)}{2} + \frac{n(n+1)}{2}\right] + n\left(\frac{n+1}{2}\right)^2 = \sum_{i=1}^{n} R_i S_i - \frac{n(n+1)^2}{4}$$

Hence an equivalent alternative formula to (7.47) is given by

$$r_s = \frac{12}{n(n^2-1)} \sum_{i=1}^{n} R_i S_i - \frac{12n(n+1)^2}{4n(n^2-1)} = \frac{12}{n(n^2-1)} \sum_{i=1}^{n} R_i S_i - 3\frac{n+1}{n-1} \qquad (7.48)$$

Mean and variance of r_s

By Assumption (7.1), $\Pr(R_i = j) = \Pr(S_i = j) = 1/n$ for all $i, j \in \{1,\ldots,n\}$. Hence $E_0[R_i] = E_0[S_i] = (n+1)/2 \;\Rightarrow\; E_0[R_iS_i] = E_0[R_i]E_0[S_i] = (n+1)^2/4$ under H_0.
Then (7.48) implies

$$E_0[r_s] = \frac{12}{n(n^2-1)} \cdot \frac{n(n+1)^2}{4} - 3\frac{n+1}{n-1} = 0 \tag{7.49}$$

We can rearrange (X_i, Y_i) such that $R_i = i$, $i = 1,\ldots,n$. By the same arguments for equation (6.11) in Section 6 (with n in place of k), we get

$$\mathrm{Var}_0\left(\sum_{i=1}^{n} R_iS_i\right) = \mathrm{Var}_0\left(\sum_{i=1}^{n} iS_i\right) = \frac{n^2(n+1)^2(n-1)}{144} \tag{7.50}$$

It follows from (7.48) and (7.50) that

$$\mathrm{Var}_0(r_s) = \frac{144}{n^2(n^2-1)^2}\mathrm{Var}_0\left(\sum_{i=1}^{n} R_iS_i\right) = \frac{(n+1)^2(n-1)}{(n^2-1)^2} = \frac{1}{n-1} \tag{7.51}$$

Rejection rule: Let $r = \text{Corr}(X,Y)$ denote the correlation coefficient between X and Y. The Spearman test has the following rejection rules at level α:

- Reject H_0 for $H_1: r > 0$ if $r_s \geq r_{s,\alpha}$;
- Reject H_0 for $H_1: r < 0$ if $r_s \leq -r_{s,\alpha}$;
- Reject H_0 for $H_1: r \neq 0$ if $|r_s| \geq r_{s,\alpha/2}$, where $\Pr(r_s \geq r_{s,\alpha}) = \alpha$ under H_0.

The value of $r_{s,\alpha}$ and the p-value of the test can be obtained from the distribution of r_s by taking $(R_1,\ldots,R_n) = (1,\ldots,n)$. Then the probability for each value of r_s is equal to the number of $(S_1,\ldots,S_n)$ that assign this value to r_s divided by $n!$.

Example 7.5 Let $n = 4$ and $(R_1,R_2,R_3,R_4) = (1,2,3,4)$. Then $n! = 24$ and

$$(S_1,\ldots,S_4) = (2,3,4,1) \Rightarrow (D_1,\ldots,D_4) = (2-1,3-2,4-3,1-4) = (1,1,1,-3) \Rightarrow$$

$$r_s = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n} D_i^2 = 1 - \frac{6(1+1+1+9)}{4(16-1)} = 1 - \frac{12}{10} = -0.2 \quad \text{by (7.47)}$$

Similarly, $r_s = -0.2$ for $(S_1,\ldots,S_4) = (4,1,2,3)$. Thus $\Pr(r_s = -0.2) = 2/24$.

The full distribution of r_s with $n = 4$ is presented in the flowing table:

r_s	(S_1, S_2, S_3, S_4)	Probability
–1.0	(4,3,2,1)	1/24
–0.8	(3,4,2,1), (4,2,3,1), (4,3,1,2)	3/24
–0.6	(3,4,1,2)	1/24
–0.4	(2,4,3,1), (3,2,4,1), (4,1,3,2), (4,2,1,3)	4/24
–0.2	(2,3,4,1), (4,1,2,3)	2/24
0.0	(2,4,1,3), (3,1,4,2)	2/24
0.2	(1,4,3,2), (3,2,1,4)	2/24
0.4	(1,3,4,2), (1,4,2,3), (2,3,1,4), (3,1,2,4)	4/24
0.6	(2,1,4,3)	1/24
0.8	(1,2,4,3), (1,3,2,4), (2,1,3,4)	3/24
1.0	(1,2,3,4)	1/24

Thus $\Pr(r_s \geq 1) = 1/24 \;\Rightarrow\; r_{s,1/24} = 1$, $\Pr(r_s \geq 0.8) = 4/24 \;\Rightarrow\; r_{s,4/24} = 0.8$, etc.

Large-sample approximation:

By the central limit theorem together with (7.49) and (7.51),

$$r_s^* = \frac{r_s - \mathrm{E}_0[r_s]}{\sqrt{\mathrm{Var}_0(r_s)}} = r_s\sqrt{n-1} \to_d N(0,1) \text{ as } n \to \infty \tag{7.52}$$

Approximate rejection rule:

Let r be the correlation coefficient between X and Y:

$$r = \mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

From (7.52), the approximate rules to test H_0 at level α are as follows:

- Reject H_0 for $H_1: r > 0$ if $r_s^* \geq z_\alpha$;
- Reject H_0 for $H_1: r < 0$ if $r_s^* \leq -z_\alpha$;
- Reject H_0 for $H_1: r \neq 0$ if $|r_s^*| \geq z_{\alpha/2}$.

Ties: If there are ties among $(X_1,\ldots,X_n)$ and/or $(Y_1,\ldots,Y_n)$, assign average ranks to tied values. Let

$$A=\sum_{i=1}^{g}t_i(t_i^2-1) \quad \text{and} \quad B=\sum_{j=1}^{h}u_j(u_j^2-1), \tag{7.53}$$

where g and h are the numbers of tied groups, t_j and u_j are the numbers of tied values in group j for $(X_1,\ldots,X_n)$ and $(Y_1,\ldots,Y_n)$, respectively. Adjust (7.47) to

$$r_s=\frac{1}{\sqrt{n(n^2-1)-A}\sqrt{n(n^2-1)-B}}\left[n(n^2-1)-6\sum_{i=1}^{n}D_i^2-\frac{1}{2}(A+B)\right] \tag{7.54}$$

Then the above rejection rules are valid approximately. The distribution of r_s conditional on ties can be worked out in a similar way to the case with no ties.

From (7.53) we see that groups with $t_j=1$ and $u_j=1$ can be ignored.

If there are no ties, then $A=B=0$ and (7.54) reduces to (7.47).

Example 7.6 In Example 8.5 of the textbook (on page 430),

$$(R_1,\ldots,R_7)=(1.5,1.5,3,4,5,6,7) \text{ and } (S_1,\ldots,S_7)=(2.5,4,2.5,1,5,6,7)$$

Hence $(D_1,\ldots,D_7)=(1,2.5,-0.5,-3,0,0,0)$ and

$$g=h=1 \text{ with } t_1=u_1=2 \Rightarrow A=B=2(2^2-1)=6 \text{ by (7.53).}$$

It then follows from (7.54) that

$$r_s=\frac{1}{7(49-1)-6}\left[7(49-1)-6\left(1+2.5^2+0.5^2+3^2\right)-\frac{1}{2}(6+6)\right]=\frac{231}{330}=0.7$$

To test H_0 against $H_1: r>0$, the p-value by R is $\Pr(r_s\geq 0.7)=0.044$.

By (7.52), $r_s^*=r_s\sqrt{n-1}=0.7\sqrt{7-1}=1.71$. So the large-sample approximation gives p-value $=\Pr\left(r_s^*\geq 1.71\right)\approx\Pr(Z\geq 1.71)=0.0436$, where $Z\sim N(0,1)$.

Both p-values rejects H_0 at the 5% level. Thus there is sufficient evidence that $(X_1,\ldots,X_n)$ and $(Y_1,\ldots,Y_n)$ are positively correlated.