Nonparametric Independence Tests

Nathaniel E. Helwig

Assistant Professor of Psychology and Statistics
University of Minnesota (Twin Cities)

Updated 04-Jan-2017

Copyright

Outline of Notes

1) Kendall's Rank Correlation:

- Overview
- Procedure
- Large Samples & Ties
- Example

2) Spearman's Rank Correlation:

- Overview
- Procedure
- Large Samples & Ties
- Example (revisited)

Kendall’s Rank Correlation

Problem(s) of Interest

Suppose we have n bivariate observations

- $(X_1, Y_1), \ldots, (X_n, Y_n)$ where (X_i, Y_i) is i-th subject's data

We want to make inferences about association between X and Y

- Let $F_{X,Y}$ denote joint distribution of X and Y
- Let F_X and F_Y denote marginal distributions of X and Y
- Null hypothesis is statistical independence:
  $F_{X,Y}(x, y) = F_X(x)F_Y(y)$ for all (x, y)

Assumptions

Independence assumption:

- $\{(X_i, Y_i)\}_{i=1}^n$ are iid from some bivariate population

Continuity assumption:

- $F_{X,Y}$ is a continuous distribution

Parameter of Interest and Hypothesis

Parameter of interest is Kendall's Population correlation coefficient:

$$\tau = 2P[(Y_2 - Y_1)(X_2 - X_1) > 0] - 1$$

and note that

- $P[(Y_2 - Y_1)(X_2 - X_1) > 0] = P(X_2 > X_1, Y_2 > Y_1) + P(X_2 < X_1, Y_2 < Y_1)$
- If X and Y are independent, then we have:
  $P(X_2 > X_1, Y_2 > Y_1) = P(X_2 > X_1)P(Y_2 > Y_1) = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \frac{1}{4}$
  $P(X_2 < X_1, Y_2 < Y_1) = P(X_2 < X_1)P(Y_2 < Y_1) = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \frac{1}{4}$

The null hypothesis about τ is independence

$$H_0 : \tau = 0$$

and we could have one of three alternative hypotheses:

- One-Sided Upper-Tail: $H_1 : \tau > 0$
- One-Sided Lower-Tail: $H_1 : \tau < 0$
- Two-Sided: $H_1 : \tau \neq 0$

Test Statistic

For all $n(n-1)/2$ paris of observations (X_i, Y_i) and (X_j, Y_j) with $1 \leq i < j \leq n$, calculate paired sign statistic $Q[(X_i, Y_i), (X_j, Y_j)]$ where

$$Q[(a,b),(c,d)] = \begin{cases} 1 & \text{if } (d-b)(c-a) > 0 \\ -1 & \text{if } (d-b)(c-a) < 0 \end{cases}$$

The Kendall test statistic K is defined as

$$K = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Q[(X_i, Y_i), (X_j, Y_j)]$$

which is simply the sum of the paired sign statistic for all pairs.

Concordant and Discordant Pairs

Given pairs of observations (X_i, Y_i) and (X_j, Y_j), we say the pairs are

- Concordant if $(X_i - X_j)(Y_i - Y_j) > 0$
- Discordant if $(X_i - X_j)(Y_i - Y_j) < 0$

Concordant: either (a) $X_i > X_j$ and $Y_i > Y_j$, or (b) $X_i < X_j$ and $Y_i < Y_j$
Discordant: either (a) $X_i < X_j$ and $Y_i > Y_j$, or (b) $X_i > X_j$ and $Y_i < Y_j$

$K = \{\# \text{ of concordant pairs}\} - \{\# \text{ of discordant pairs}\}$

Distribution of Test Statistic under H_0

WLOG suppose data are ordered according to X_i; then, under H_0 all $n!$ arrangements of Y-ranks occur with equal probability

- Given n, calculate K for all $n!$ possible outcomes
- Each outcome has probability $1/n!$ under H_0

Example null distribution with $n = 3$:

X-ranks	Y-ranks	K	Probability under H_0
1,2,3	1,2,3	3	1/6
1,2,3	1,3,2	1	1/6
1,2,3	2,1,3	1	1/6
1,2,3	2,3,1	-1	1/6
1,2,3	3,1,2	-1	1/6
1,2,3	3,2,1	-3	1/6

Note: there are $3! = 6$ possibilities

Hypothesis Testing

One-Sided Upper Tail Test:

- $H_0 : \tau = 0$ versus $H_1 : \tau > 0$
- Reject H_0 if $\bar{K} \geq k_\alpha$ where $P(K > k_\alpha) = \alpha$ and $\bar{K} = \frac{K}{n(n-1)/2}$

One-Sided Lower Tail Test:

- $H_0 : \tau = 0$ versus $H_1 : \tau < 0$
- Reject H_0 if $\bar{K} \leq -k_\alpha$

Two-Sided Test:

- $H_0 : \tau = 0$ versus $H_1 : \tau \neq 0$
- Reject H_0 if $\bar{K} \geq k_{\alpha/2}$ or $\bar{K} \leq -k_{\alpha/2}$

Estimating Kendall's τ

Can estimate population τ using sample estimate

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

given that $-\frac{n(n-1)}{2} \leq K \leq \frac{n(n-1)}{2}$.

$\hat{\tau}$ is sometimes referred to as Kendall's τ rank correlation coefficient.

Confidence Intervals for Kendall's τ

To form a confidence interval based on $\hat{\tau}$, first calculate

$$C_i = \sum_{j=1}^{n} I_{\{i \neq j\}} Q[(X_i, Y_i), (X_j, Y_j)] \quad \text{for } i = 1, \ldots, n$$

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

where $I_{\{\cdot\}}$ is an indicator function and $\bar{C} = \frac{1}{n} \sum_{i=1}^{n} C_i = 2K/n$.

Can form a symmetric confidence interval using

$$[\hat{\tau} - z_{\alpha/2}\hat{\sigma}; \;\; \hat{\tau} + z_{\alpha/2}\hat{\sigma}]$$

where $z_{\alpha/2}$ is critical value from standard normal.

Large Sample Approximation

Under H_0, the expected value and variance of K are

- $E(K) = 0$
- $V(K) = \frac{n(n-1)(2n+5)}{18}$

We can create a standardized test statistic K^* of the form

$$K^* = \frac{K - E(K)}{\sqrt{V(K)}}$$

which asymptotically follows a $\mathrm{N}(0, 1)$ distribution.

Derivation of Large Sample Expectation

For the expectation of K, note that

$$
\begin{aligned}
E(K) &= E\left\{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} Q[(X_i, Y_i), (X_j, Y_j)]\right\} \\
&= \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} E\{Q[(X_i, Y_i), (X_j, Y_j)]\} \\
&= \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} P\{(Y_2 - Y_1)(X_2 - X_1) > 0\} - P\{(Y_2 - Y_1)(X_2 - X_1) < 0\} \\
&= \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} [2P\{(Y_2 - Y_1)(X_2 - X_1) > 0\} - 1] \\
&= \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \tau = \binom{n}{2}\tau
\end{aligned}
$$

and note that $\tau = 0$ under H_0.

Derivation of Large Sample Variance

For the variance of K, note that

$$V(K) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} V(Q_{ij}) + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \sum_{s=1}^{n-1} \sum_{t=s+1}^{n} I_{\{(i,j)\neq(s,t)\}} \mathrm{cov}(Q_{ij}, Q_{st})$$

where $Q_{ij} = Q[(X_i, Y_i), (X_j, Y_j)]$ and $I_{\{\cdot\}}$ is an indicator function.

After some tedious manipulation, it can be shown that $V(K)$ reduces to

$$V(K) = \frac{n(n-1)(2n+5)}{18}$$

see Hollander et al. (2014) for more information.

Handling Ties

If there are ties among the n X values and/or among the n Y values, then define the modified paired sign statistic $Q^*[(X_i, Y_i), (X_j, Y_j)]$ as

$$Q^*[(a,b),(c,d)] = \begin{cases} 1 & \text{if } (d-b)(c-a) > 0 \\ 0 & \text{if } (d-b)(c-a) = 0 \\ -1 & \text{if } (d-b)(c-a) < 0 \end{cases}$$

- $K = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Q^*_{ij}$ is calculated in same fashion
- Using Q^* gives an approximate level α test
- Can still obtain an exact level α test via individual randomization

Large sample approximation variance formula also needs to be reduced when there are ties (see Hollander et al., 2014).

Kendall's τ-b

The $\hat{\tau}$ that we used before is called τ-a, and is an ideal estimate of τ when we have no ties.

When ties are present, we use τ-b, which has the form:

$$\hat{\tau}_b = \frac{K}{\sqrt{([n(n-1)/2] - n_x)}\sqrt{([n(n-1)/2] - n_y)}}$$

where

- $n_x = \sum_i t_i(t_i - 1)/2$ with t_i denoting size of i-th group of ties on X
- $n_y = \sum_i u_i(u_i - 1)/2$ with u_i denoting size of i-th group of ties on Y

Example: Data

Nonparametric Statistical Methods, 3rd Ed. (Hollander et al., 2014)

Table 8.5 Psychological Test Scores of Dizygous Male Twins

Pair i	X_i	Y_i
1	277	256
2	169	118
3	157	137
4	139	144
5	108	146
6	213	221
7	232	184
8	229	188
9	114	97
10	232	231
11	161	114
12	149	187
13	128	230

Source: P. J. Clark, S. G. Vandenberg, and C. H. Proctor (1961).

Example: By Hand

$Q[(X_i, Y_i), (X_j, Y_j)]$ for all $n(n-1)/2 = 78$ pairs of interest

$j \backslash i$	1	2	3	4	5	6	7	8	9	10	11	12
2	1	0	0	0	0	0	0	0	0	0	0	0
3	1	−1	0	0	0	0	0	0	0	0	0	0
4	1	−1	−1	0	0	0	0	0	0	0	0	0
5	1	−1	−1	−1	0	0	0	0	0	0	0	0
6	1	1	1	1	1	0	0	0	0	0	0	0
7	1	1	1	1	1	−1	0	0	0	0	0	0
8	1	1	1	1	1	−1	−1	0	0	0	0	0
9	1	1	1	1	−1	1	1	1	0	0	0	0
10	1	1	1	1	1	1	0	1	1	0	0	0
11	1	1	−1	−1	−1	1	1	1	1	1	0	0
12	1	−1	−1	1	1	1	−1	1	1	1	−1	0
13	1	−1	−1	−1	1	−1	−1	−1	1	1	−1	−1

$K = \sum_{i=1}^{12} \sum_{j=i+1}^{13} = 52 - 25 = 27$

$\hat{\tau} = \frac{K}{n(n-1)/2} = 27/78 = 0.3461538$ and

$\hat{\tau}_b = \frac{K}{\sqrt{\{78-1\}\{78-0\}}} = 27/\sqrt{(77 * 78)} = 0.3483943$

Example: Using R (Hard Way, part 1)

```
> x = c(277,169,157,139,108,213,232,229,114,232,161,149,128)
> y = c(256,118,137,144,146,221,184,188,97,231,114,187,230)
> n = 13
> Qmat = matrix(0,n-1,n-1)
> colnames(Qmat) = 1:(n-1)
> rownames(Qmat) = 2:n
> for(i in 1:(n-1)){
+     for(j in (i+1):n){
+         qval = (y[j]-y[i])*(x[j]-x[i])
+         if(qval>0){
+             Qmat[j-1,i] = 1
+         } else if(qval<0){
+             Qmat[j-1,i] = -1
+         }
+     }
+ }
```

Example: Using R (Hard Way, part 2)

```
> Qmat
   1  2  3  4  5  6  7  8 9 10 11 12
2  1  0  0  0  0  0  0  0 0  0  0  0
3  1 -1  0  0  0  0  0  0 0  0  0  0
4  1 -1 -1  0  0  0  0  0 0  0  0  0
5  1 -1 -1 -1  0  0  0  0 0  0  0  0
6  1  1  1  1  1  0  0  0 0  0  0  0
7  1  1  1  1  1 -1  0  0 0  0  0  0
8  1  1  1  1  1 -1 -1  0 0  0  0  0
9  1  1  1  1 -1  1  1  1 0  0  0  0
10 1  1  1  1  1  1  0  1 1  0  0  0
11 1  1 -1 -1 -1  1  1  1 1  1  0  0
12 1 -1 -1  1  1  1 -1  1 1  1 -1  0
13 1 -1 -1 -1  1 -1 -1 -1 1  1 -1 -1
> K = sum(Qmat)    # or #    K = sum(Qmat[Qmat>0]) + sum(Qmat[Qmat<0])
> tauhat = K/(n*(n-1)/2)
> taub = K/sqrt(77*78)    # Kendall's tau-b
> K
[1] 27
> tauhat
[1] 0.3461538
> taub
[1] 0.3483943
```

Example: Using R (Hard Way, part 3)

```
> Qfun = function(i,j){
+     qval = (y[j]-y[i])*(x[j]-x[i])
+     q = 0
+     if(qval>0) { q = 1 } else if(qval<0) { q = -1 }
+     return(q)
+ }
> Cvec = rep(0,n)
> idx = 1:n
> for(i in idx){
+     for(j in idx[idx!=i]){
+         Cvec[i] = Cvec[i] + Qfun(i,j)
+     }
+ }
> Cbar = mean(Cvec)
> const = 2/(n*(n-1))
> sigsq = const*( const*((n-2)/(n-1))*sum((Cvec-Cbar)^2) + 1 - taub^2 )
> c(taub-qnorm(.975)*sqrt(sigsq), taub+qnorm(.975)*sqrt(sigsq))
[1] -0.09026324  0.78705193
> c(taub-qnorm(.95)*sqrt(sigsq), 1)
[1] -0.0197387  1.0000000
```

Example: Using R (Easy Way)

```
> x = c(277,169,157,139,108,213,232,229,114,232,161,149,128)
> y = c(256,118,137,144,146,221,184,188,97,231,114,187,230)
> cor.test(x,y,method="kendall",alternative="greater")

 Kendall's rank correlation tau

data:  x and y
z = 1.6503, p-value = 0.04944
alternative hypothesis: true tau is greater than 0
sample estimates:
      tau
0.3483943

Warning message:
In cor.test.default(x, y, method = "kendall", alternative = "greater") :
  Cannot compute exact p-value with ties
> require(NSM3)
> kendall.ci(x,y)

1 - alpha = 0.95 two-sided CI for tau:
-0.09, 0.787
```

Spearman’s Rank Correlation

Same Problem of Interest

Suppose we have n bivariate observations

- $(X_1, Y_1), \ldots, (X_n, Y_n)$ where (X_i, Y_i) is i-th subject's data

We want to make inferences about association between X and Y

- Let $F_{X,Y}$ denote joint distribution of X and Y
- Let F_X and F_Y denote marginal distributions of X and Y
- Null hypothesis is statistical independence: $F_{X,Y}(x, y) = F_X(x)F_Y(y)$ for all (x, y)

Assumptions

Independence assumption:

- $\{(X_i, Y_i)\}_{i=1}^n$ are iid from some bivariate population

Continuity assumption:

- $F_{X,Y}$ is a continuous distribution

Parameter of Interest and Hypothesis

Parameter of interest is an awkward measure of association:

$$\eta = \frac{3[\tau + (n-2)\phi]}{n+1}$$

where $\phi = 2P[(Y_3 - Y_1)(X_2 - X_1) > 0] - 1$.

- Note that $\tau = \phi = 0$ if X and Y are independent

The null hypothesis about η is independence

$$H_0 : \eta = 0$$

and we could have one of three alternative hypotheses:

- One-Sided Upper-Tail: $H_1 : \eta > 0$
- One-Sided Lower-Tail: $H_1 : \eta < 0$
- Two-Sided: $H_1 : \eta \neq 0$

Test Statistic

Letting R_i and S_i denote the (separate) ranks of the X and Y values, Spearman's rank correlation is

$$r_s = \frac{\sum_{i=1}^n (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^n (R_i - \bar{R})^2}\sqrt{\sum_{i=1}^n (S_i - \bar{S})^2}}$$

where

- $\bar{R} = (1/n)\sum_{i=1}^n R_i$ is the mean X rank
- $\bar{S} = (1/n)\sum_{i=1}^n S_i$ is the mean Y rank

Note that Spearman's rank correlation r_s is the Pearson product moment correlation of the ranks R_i and S_i.

Distribution of Test Statistic under H_0

WLOG suppose data are ordered according to X_i; then, under H_0 all $n!$ arrangements of Y-ranks occur with equal probability

- Given n, calculate r_s for all $n!$ possible outcomes
- Each outcome has probability $1/n!$ under H_0

Example null distribution with $n = 3$:

X-ranks	Y-ranks	r_s	Probability under H_0
1,2,3	1,2,3	1.0	1/6
1,2,3	1,3,2	0.5	1/6
1,2,3	2,1,3	0.5	1/6
1,2,3	2,3,1	-0.5	1/6
1,2,3	3,1,2	-0.5	1/6
1,2,3	3,2,1	-1.0	1/6

Note: there are $3! = 6$ possibilities

Hypothesis Testing

One-Sided Upper Tail Test:

- $H_0 : \eta = 0$ versus $H_1 : \eta > 0$
- Reject H_0 if $r_s \geq r_{s;\alpha}$ where $P(r_s > r_{s;\alpha}) = \alpha$

One-Sided Lower Tail Test:

- $H_0 : \eta = 0$ versus $H_1 : \eta < 0$
- Reject H_0 if $r_s \leq -r_{s;\alpha}$

Two-Sided Test:

- $H_0 : \eta = 0$ versus $H_1 : \eta \neq 0$
- Reject H_0 if $r_s \geq r_{s;\alpha/2}$ or $r_s \leq -r_{s;\alpha/2}$

Large Sample Approximation

Under H_0, the expected value and variance of r_s are

- $E(r_s) = 0$
- $V(r_s) = \frac{1}{n-1}$

We can create a standardized test statistic r_s^* of the form

$$r_s^* = \frac{r_s - E(r_s)}{\sqrt{V(r_s)}}$$

which asymptotically follows a $\mathrm{N}(0, 1)$ distribution.

Derivation of Large Sample Expectation

Assuming there are no ties and H_0 is true, we have that

- $r_s = \frac{12\sum_{i=1}^n (R_i - \frac{n+1}{2})(S_i - \frac{n+1}{2})}{n(n^2-1)} = 1 - \frac{6\sum_{i=1}^n D_i^2}{n(n^2-1)}$ where $D_i = S_i - R_i$
- $\sum_{i=1}^n R_i S_i$ has the same distribution as $\sum_{i=1}^n iS_i$

Therefore, under H_0, the expectation of r_s is

$$
\begin{aligned}
E(r_s) &= E\left\{ \frac{12\sum_{i=1}^n iS_i}{n(n^2-1)} - 3\frac{n+1}{n-1} \right\} \\
&= \frac{12\sum_{i=1}^n iE(S_i)}{n(n^2-1)} - 3\frac{n+1}{n-1} \\
&= \frac{12\sum_{i=1}^n i\frac{n+1}{2}}{n(n^2-1)} - 3\frac{n+1}{n-1} \\
&= \frac{12\frac{n(n+1)}{2}\frac{n+1}{2}}{n(n^2-1)} - 3\frac{n+1}{n-1} = 0
\end{aligned}
$$

Derivation of Large Sample Variance

Similar to the previous argument, under H_0, the variance of r_s is

$$\begin{aligned} V(r_s) &= V\left\{ \frac{12\sum_{i=1}^{n} iS_i}{n(n^2-1)} - 3\frac{n+1}{n-1} \right\} \\ &= \frac{144}{n^2(n^2-1)^2} V\left(\sum_{i=1}^{n} iS_i \right) \end{aligned}$$

Furthermore, under H_0, we can show that

$$V\left(\sum_{i=1}^{n} iS_i \right) = \frac{n^2(n+1)(n^2-1)}{144}$$

which implies the large sample variance has the form $V(r_s) = \frac{1}{n-1}$; see Hollander et al. (2014) for more information.

Handling Ties

If there are ties within the X or Y values, then use the average ranking procedure to handle the ties.

- Just calculate r_s using Pearson formula with averaged ranks
- No longer an exact level α test
- Can obtain an exact level α test using *conditional null distribution*

Example: Data Revisited

Nonparametric Statistical Methods, 3rd Ed. (Hollander et al., 2014)

Table 8.5 Psychological Test Scores of Dizygous Male Twins

Pair i	X_i	R_i	Y_i	S_i
1	277	13.0	256	13
2	169	8.0	118	3
3	157	6.0	137	4
4	139	4.0	144	5
5	108	1.0	146	6
6	213	9.0	221	10
7	232	11.5	184	7
8	229	10.0	188	9
9	114	2.0	97	1
10	232	11.5	231	12
11	161	7.0	114	2
12	149	5.0	187	8
13	128	3.0	230	11

Source: P. J. Clark, S. G. Vandenberg, and C. H. Proctor (1961).

Example: By Hand

Pair i	X_i	R_i	Y_i	S_i
1	277	13.0	256	13
2	169	8.0	118	3
3	157	6.0	137	4
4	139	4.0	144	5
5	108	1.0	146	6
6	213	9.0	221	10
7	232	11.5	184	7
8	229	10.0	188	9
9	114	2.0	97	1
10	232	11.5	231	12
11	161	7.0	114	2
12	149	5.0	187	8
13	128	3.0	230	11
$\sum_{i=1}^{13}$	2308	91.0	2253	91

$$r_s = \frac{\sum_{i=1}^{13}(R_i - 7)(S_i - 7)}{\sqrt{\sum_{i=1}^{13}(R_i - 7)^2}\sqrt{\sum_{i=1}^{13}(S_i - 7)^2}} = 0.5144434$$

Example: Using R (Hard Way)

```
> x = c(277,169,157,139,108,213,232,229,114,232,161,149,128)
> y = c(256,118,137,144,146,221,184,188,97,231,114,187,230)
> rx = rank(x)
> ry = rank(y)
> mx = mean(rx)
> my = mean(ry)
> sum((rx-mx)*(ry-my))/sqrt(sum((rx-mx)^2)*sum((ry-my)^2))
[1] 0.5144434
> cor(rx,ry)
[1] 0.5144434
```

Example: Using R (Easy Way)

```
> x = c(277,169,157,139,108,213,232,229,114,232,161,149,128)
> y = c(256,118,137,144,146,221,184,188,97,231,114,187,230)
> cor(x,y,method="spearman")
[1] 0.5144434
> cor.test(x,y,method="spearman")

 Spearman's rank correlation rho

data:  x and y
S = 176.7426, p-value = 0.07206
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.5144434

Warning message:
In cor.test.default(x, y, method = "spearman") :
  Cannot compute exact p-value with ties
```