Chapter 8

The Independence Problem

INTRODUCTION

The data in this chapter consist of a random sample from a bivariate population. Our basic interest here is in the statistical relationship between the two variables involved in the bivariate structure. In particular, we will discuss procedures for deciding whether or not these two variables are independent and, if not independent, for assessing both the type and degree of dependency that exists between them.

In Section 8.1, we present a distribution-free test for independence that is based on signs of appropriate products of differences. Section 8.2 presents an estimator of the measure of association τ defined by (8.2). Section 8.3 contains an asymptotically distribution-free confidence interval for τ. Section 8.4 uses Efron's bootstrap method to obtain a different asymptotically distribution-free confidence interval for τ. Section 8.5 presents a distribution-free test for independence based on ranks. Section 8.6 contains a distribution-free test of independence, which is consistent against a broader class of alternatives than those classes of alternatives that can be detected by the tests of Sections 8.1 and 8.5. Section 8.7 considers the asymptotic relative efficiencies of the procedures in this chapter with respect to their normal theory counterparts.

Data. We obtain n bivariate observations $(X_1, Y_1), \ldots, (X_n, Y_n)$, one observation on each of n subjects.

Assumptions

A The n bivariate observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ are a random sample from a continuous bivariate population. That is, the (X, Y) pairs are mutually independent and identically distributed according to some continuous bivariate population.

8.1 A DISTRIBUTION-FREE TEST FOR INDEPENDENCE BASED ON SIGNS (KENDALL)

Hypothesis

Let $F_{X,Y}$ be the joint distribution function for the common bivariate population of the (X, Y) pairs. Moreover, let $F_x(x)$ and $F_Y(y)$ be the distribution functions for the marginal

Nonparametric Statistical Methods, Third Edition. Myles Hollander, Douglas A. Wolfe, Eric Chicken.

X and Y populations, respectively. The null hypothesis of interest here is that the X and Y random variables are independent. Formally stated, this null hypothesis is

$$H_0 : [F_{X,Y}(x,y) \equiv F_X(x)F_Y(y), \text{ for all } (x,y) \text{ pairs}]. \tag{8.1}$$

The alternative hypothesis to (8.1) will be a function of the type of dependence between the X and Y variables that is of principal interest. In this section, we concentrate on a type of dependence measured by the Kendall population correlation coefficient

$$\tau = 2P\{(Y_2 - Y_1)(X_2 - X_1) > 0\} - 1. \tag{8.2}$$

We note that the event $\{(Y_2 - Y_1)(X_2 - X_1) > 0\}$ occurs if and only if either the event $\{X_2 > X_1 \text{ and } Y_2 > Y_1\}$ or the event $\{X_2 < X_1 \text{ and } Y_2 < Y_1\}$ occurs. These latter two events are mutually exclusive, therefore

$$\begin{aligned} P\{(Y_2 - Y_1)(X_2 - X_1) > 0\} &= P(X_2 > X_1, Y_2 > Y_1) \\ &\quad + P(X_2 < X_1, Y_2 < Y_1). \end{aligned} \tag{8.3}$$

If X and Y are independent, it follows that

$$P(X_2 > X_1, Y_2 > Y_1) = P(X_2 > X_1)P(Y_2 > Y_1) = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \frac{1}{4}, \tag{8.4}$$

because X_1, X_2 are independent and identically distributed variables, as are Y_1, Y_2 (although not necessarily, of course, with the same distribution as the X's). Similarly, if X and Y are independent, we also have

$$P(X_2 < X_1, Y_2 < Y_1) = \frac{1}{4}.$$

Combining this result with (8.3) and (8.4), we see that the Kendall population correlation coefficient $\tau = 2\left(\frac{1}{4} + \frac{1}{4}\right) - 1 = 0$ if X and Y are independent. (It is important to point out that this is not an if and only if statement because $\tau = 0$ does not necessarily imply that X and Y are independent. See Comment 2 for more on this relationship.)

Procedure

To compute the Kendall sample correlation statistic K, we first calculate the values of the $n(n-1)/2$ paired sign statistics $Q((X_i, Y_i), (X_j, Y_j))$, for $1 \le i < j \le n$, where

$$Q((a,b),(c,d)) = \begin{cases} 1, & \text{if } (d-b)(c-a) > 0, \\ -1, & \text{if } (d-b)(c-a) < 0. \end{cases} \tag{8.5}$$

That is, for each pair of subscripts (i,j) with $i < j$, score 1 if $(Y_j - Y_i)(X_j - X_i)$ is positive and score -1 if $(Y_j - Y_i)(X_j - X_i)$ is negative. The Kendall statistic K is then

$$K = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Q((X_i, Y_i), (X_j, Y_j)), \tag{8.6}$$

corresponding to adding up the 1's and -1's from the paired sign statistics.

a. *One-Sided Upper-Tail Test.* To test the null hypothesis of independence, namely,

$$H_0 : [F_{X,Y}(x, y) \equiv F_X(x)F_Y(y), \text{ for all}(x, y) \text{ pairs}]$$

(which implies $\tau = 0$) versus the alternative that X and Y are positively correlated (see Comment 2) corresponding to

$$H_1 : \tau > 0, \tag{8.7}$$

at the α-level of significance,

$$\text{Reject } H_0 \text{ if } \overline{K} \geq k_\alpha; \quad \text{otherwise do not reject,} \tag{8.8}$$

where the constant k_α is chosen to make the type I error probability equal to α and $\overline{K} = K/(n(n-1)/2)$, the average of the paired sign statistics Q. Values of k_α are found using the command `qKendall` (Wheeler (2009)).

b. *One-Sided Lower-Tail Test.* To test

$$H_0 : [F_{X,Y}(x, y) \equiv F_X(x)F_Y(y), \text{ for all } (x, y) \text{ pairs}]$$

versus the alternative that X and Y are negatively correlated (see Comment 2) corresponding to

$$H_2 : \tau < 0,$$

at the α-level of significance,

$$\text{Reject } H_0 \text{ if } \overline{K} \leq -k_\alpha; \quad \text{otherwise do not reject.} \tag{8.9}$$

c. *Two-Sided Test.* To test

$$H_0 : [F_{X,Y}(x, y) \equiv F_X(x)F_Y(y), \text{ for all } (x, y) \text{ pairs}]$$

versus the general alternative that X and Y are dependent variables corresponding to

$$H_3 : \tau \neq 0,$$

at the α-level of significance,

$$\text{Reject } H_0 \text{ if } |\overline{K}| \geq k_{\alpha/2}; \quad \text{otherwise do not reject.} \tag{8.10}$$

This two-sided procedure is the two-sided symmetric test with $\alpha/2$ probability in each tail of the null distribution of $\overline{K}$.

Large-Sample Approximation

The large-sample approximation is based on the asymptotic normality of K, suitably standardized. For this standardization, we need to know the expected value and variance of K when the null hypothesis of independence is true. Under H_0, the expected value and variance of K are

$$E_0(K) = 0 \tag{8.11}$$

and

$$\text{var}_0(K) = \frac{n(n-1)(2n+5)}{18}, \tag{8.12}$$

respectively. These expressions for $E_0(K)$ and $\text{var}_0(K)$ are verified by direct calculations in Comment 7 for the special case of $n = 4$. General derivations of both expressions are presented in Comment 10.

The standardized version of K is

$$K^* = \frac{K - E_0(K)}{\{\text{var}_0(K)\}^{1/2}} = \frac{K}{\{n(n-1)(2n+5)/18\}^{1/2}}. \tag{8.13}$$

When H_0 is true, K^* has, an n tends to infinity, an asymptotic $N(0,1)$ distribution (see Comment 10 for indications of the proof). The normal theory approximation for procedure (8.8) is

$$\text{Reject } H_0 \text{ if } K^* \geq z_\alpha; \quad \text{otherwise do not reject,} \tag{8.14}$$

the normal theory approximation for procedure (8.9) is

$$\text{Reject } H_0 \text{ if } K^* \leq -z_\alpha; \quad \text{otherwise do not reject,} \tag{8.15}$$

and the normal theory approximation for procedure (8.10) is

$$\text{Reject } H_0 \text{ if } |K^*| \geq z_{\alpha/2}; \quad \text{otherwise do not reject.} \tag{8.16}$$

Ties

If there are ties among the n X observations and/or separately among the n Y observations, replace the function $Q((a,b),(c,d))$ in the definition of K (8.6) by

$$Q^*((a,b),(c,d)) = \begin{cases} 1, & \text{if } (d-b)(c-a) > 0, \\ 0, & \text{if } (d-b)(c-a) = 0, \\ -1, & \text{if } (d-b)(c-a) < 0. \end{cases} \tag{8.17}$$

(Thus, in the case of tied X values and/or tied Y values, zeros are assigned to the associated paired sign statistics.) After computing K with these modified paired sign statistics, use procedure (8.8), (8.9), or (8.10). Note, however, that this test associated with tied X's and/or Y's is only approximately, and not exactly, of significance level α.

When applying the large-sample approximation, however, the loss in variability due to the tied X's and/or tied Y's must also be taken into account. While these ties do not affect the null expected value of K, its null variance is reduced to

$$\begin{aligned}\text{var}_0(K) = &\frac{\left\{n(n-1)(2n+5) - \sum_{i=1}^{g} t_i(t_i-1)(2t_i+5) - \sum_{j=1}^{h} u_j(u_j-1)(2u_j+5)\right\}}{18} \\ &+ \frac{\left\{\sum_{i=1}^{g} t_i(t_i-1)(t_i-2)\right\}\left\{\sum_{j=1}^{h} u_j(u_j-1)(u_j-2)\right\}}{9n(n-1)(n-2)} \\ &+ \frac{\left\{\sum_{i=1}^{g} t_i(t_i-1)\right\}\left\{\sum_{j=1}^{h} u_j(u_j-1)\right\}}{2n(n-1)} \end{aligned} \tag{8.18}$$

in the presence of such ties, where in (8.18) g denotes the number of tied X groups, t_i is the size of tied X group i, h is the number of tied Y groups, and u_j is the size of tied Y group j. We note that an untied $X(Y)$ observation is considered to be a tied $X(Y)$ "group" of size 1. In particular, if neither the collection of n X nor the collection of n Y observations contains tied observations, we have $g = h = n$, $t_i = u_j = 1$, $i = 1, \ldots, n$, and $j = 1, \ldots, n$. In this case of no tied X's and no tied Y's, each term involving either $(t_i - 1)$ or $(u_j - 1)$ or both reduces to zero and the variance expression in (8.18) reduces to the usual null variance of K, as given previously in (8.12).

As a consequence of the effect that ties have on the null variance of K, the following modification is needed to apply the large-sample approximation when there are tied X observations and/or tied Y observations. Compute K with the modified paired sign statistic using (8.17) and set

$$K^* = \frac{K}{\{\text{var}_0(K)\}^{1/2}}, \tag{8.19}$$

where $\text{var}_0(K)$ is now given by display (8.18). With this modified form of K^*, approximation (8.14), (8.15), or (8.16) can be applied.

EXAMPLE 8.1 ***Tuna Lightness and Quality.***

The data in Table 8.1 are a subset of the data obtained by J. Rasekh, A. Kramer, and R. Finch (1970) in a study designed to ascertain the relative importance of the various factors contributing to tuna quality and to find objective methods for determining quality parameters and consumer preference. Table 8.1 gives values of the Hunter L measure of lightness, along with panel scores for nine lots of canned tuna. The original consumer panel scores of excellent, very good, good, fair, poor, and unacceptable were converted to the numerical values of 6, 5, 4, 3, 2, and 1, respectively. The panel scores in Table 8.1 are averages of 80 such values. (The Y random variable is thus discrete, and hence, the continuity portion of Assumption A is not satisfied. Nevertheless, because each Y is an average of 80 values, we need not be nervous about this departure from Assumption A.)

It is suspected that the Hunter L value is positively associated with the panel score. Thus, we will apply procedure (8.8) to test H_0 (8.1) versus $\tau > 0$. Consider the significance level $\alpha = .10$. Using `qKendall(p=.10, N=9, lower.tail=T)` gives a value of $-.333$. By the symmetry of $\overline{K}$, the critical value is therefore .333.

Table 8.1 Hunter L Values and Consumer Panel Scores for Nine Lots of Canned Tuna

Lot	Hunter L value (X)	Panel score (Y)
1	44.4	2.6
2	45.9	3.1
3	41.9	2.5
4	53.3	5.0
5	44.7	3.6
6	44.1	4.0
7	50.7	5.2
8	45.2	2.8
9	60.1	3.8

Source: J. Rasekh, A. Kramer, and R. Finch (1970).

Table 8.2 $Q((X_i, Y_i), (X_j, Y_j))$ Values for Canned Tuna Data

$j \backslash i$	1	2	3	4	5	6	7	8
2	1							
3	1	1						
4	1	1	1					
5	1	−1	1	1				
6	−1	−1	1	1	−1			
7	1	1	1	−1	1	1		
8	1	1	1	1	−1	−1	1	
9	1	1	1	−1	1	−1	−1	1

We illustrate the computations of the paired sign statistics in (8.5) leading to the sample value of K (8.6) in Table 8.2.

Summing the $+1$ and -1 values in Table 8.2 we see that

$$K = \sum_{i=1}^{8} \sum_{j=i+1}^{9} Q((X_i, Y_i), (X_j, Y_j)) = 26 - 10 = 16,$$

and $\overline{K} = 16/36$.

This value of $\overline{K}$ is greater than the critical value .333, so we reject H_0 in favor of $\tau > 0$ at the $\alpha = .10$ level. Note that the critical value given by R results in a significance level of $\alpha = .13$, not $\alpha = .10$.

Since the one-sided P-value for these data is the lowest significance level at which we can reject H_0 in favor of $\tau > 0$ with the observed value of the test statistic $\overline{K} = 16/36$. The P-value for these data is $P_0(\overline{K} \geq 16/36) = P_0(\overline{K} \leq -16/36)$. The P-value is `pKendall(-16/36, N=9, lower.tail=T) =.060`. Thus, there is some evidence (although not overwhelming) that the Hunter L lightness values and the panel scores are positively correlated.

The R command `cor.test` will perform this test without the need to use `qKendall` or `pKendall`. The analysis above can be replicated by

```
cor.test (x, y, method="kendall", alt="greater")
```

where `x` is the Hunter L value and `y` is the panel score from Table 8.1. This results in the output

```
 Kendall's rank correlation tau

data: x and y
T = 26, p-value = 0.05972
alternative hypothesis: true tau is greater than 0
sample estimates:
tau
0.4444444.
```

The value of the test statistic is given here is $T = 26$. R sums (8.6) only over those values of Q giving a positive 1. Since the number of pairs is $n(n-1)/2$, there must be $n(n-1)/2 - T$ negative 1 values in (8.6). To convert T to K, one uses the relation $K = 2T - n(n-1)/2$. For the data in Table 8.1, $T = 26$ is equivalent to $K = 2 \cdot 26 - 9 \cdot 8/2 = 16$.

For the large-sample approximation, we find (since there are no ties in the data) from (8.13) that

$$K^* = \frac{16}{\{9(8)(23)/18\}^{1/2}} = 1.67.$$

Thus, the smallest significance level at which we can reject H_0 in favor of $\tau > 0$ using the normal theory approximation is .0475, since $z_{.0475} = 1.67$. This is in good agreement with the exact P-value of .060 found previously.

Comments

1. *Motivation for the Test.* The null hypothesis of this section is that the X and Y random variables are independent, which implies (see the discussion in Procedure) that the Kendall population correlation coefficient τ is equal to 0. However, the alternatives are stated directly in terms of $\tau (>, <, \text{or } \neq 0)$. When τ is greater than 0 (and thus $P((Y_2 - Y_1)(X_2 - X_1) > 0) > \frac{1}{2}$), there will tend to be a large number of positive paired sign statistics and fewer negative paired sign statistics. Hence, when τ is greater than 0, we would expect the sample to lead to a big, positive value for K. This suggests rejecting H_0 in favor of $\tau > 0$ for large values of K and motivates procedures (8.8) and (8.14). Similar rationales lead to procedures (8.9), (8.10), (8.15), and (8.16).
2. *Interpretation of* τ. The Kendall correlation coefficient τ can also be written as $\tau = [P((Y_2 - Y_1)(X_2 - X_1) > 0) - P((Y_2 - Y_1)(X_2 - X_1) < 0)]$. We have already noted that if X and Y are independent, then $\tau = 0$. On the other hand, if $\tau > 0$, then it is more likely that $\{X_2 > X_1 \text{ and } Y_2 > Y_1\}$ or $\{X_2 < X_1 \text{ and } Y_2 < Y_1\}$ occurs than either of the complementary events $\{X_2 > X_1 \text{ and } Y_2 < Y_1\}$ or $\{X_2 < X_1 \text{ and } Y_2 > Y_1\}$. Thus, if $\tau > 0$, it is more likely that the change from X_1 to X_2 has the same (rather than opposite) sign as that from Y_1 to Y_2. It is reasonable to interpret this type of relationship between X and Y as indicative of a positive association (as measured by τ). Similarly, $\tau < 0$ may reasonably be interpreted as indicative of a negative association (as measured by τ) between X and Y.
3. *Concordant/Discordant Pairs.* Call the $(X_i, Y_i), (X_j, Y_j)$ pairs *concordant* if $(X_i - X_j)(Y_i - Y_j) > 0$ and *discordant* if $(X_i - X_j)(Y_i - Y_j) < 0$. Thus, (X_i, Y_i) and (X_j, Y_j) are concordant if either (a) $X_i > X_j$ and $Y_i > Y_j$ or (b) $X_i < X_j$ and $Y_i < Y_j$. Similarly, (X_i, Y_i) and (X_j, Y_j) are discordant if either (c) $X_i < X_j$

and $Y_i > Y_j$ or (d) $X_i > X_j$ and $Y_i < Y_j$. Now K (8.6) can be expressed as $K = K' - K''$, where

$$K' = \text{number of concordant pairs},$$

$$K'' = \text{number of discordant pairs},$$

and the count is taken over the $n(n-1)/2$ sets of pairs $(X_i, Y_i), (X_j, Y_j)$ with $i < j$. Note that $(X_i, Y_i), (X_j, Y_j)$ are concordant if the ordering of X_i, X_j agrees with that of Y_i, Y_j. We have discordance when these orderings do not agree. Thus, $K/\{n(n-l)/2\}$ can be viewed as an average measure of agreement between the X's and the Y's, where agreement refers to order.

4. *Equivalent Expression When There Are No Ties.* Let K' = (number of concordant pairs) and K'' = (number of discordant pairs), as defined in Comment 3. If there are no ties among the X's and no ties among the Y's, then $K' + K'' = n(n-1)/2$. Thus, with no ties, we have $K = K' - K'' = K' - [n(n-1)/2 - K'] = 2K' - \{n(n-1)/2\}$. To illustrate, consider the tuna data in Example 8.1. Summing the 1's in Table 8.2 (corresponding to concordant pairs), we obtain $K' = 7 + 5 + 6 + 3 + 2 + 1 + 1 + 1 = 26$. Adding the 0's in Table 8.2 (corresponding to discordant pairs), we have $K'' = 1 + 2 + 0 + 2 + 2 + 2 + 1 + 0 = 10$. It follows that $K = K' - K'' = 26 - 10 = 2K' - \{n(n-1)/2\} = [2(26) - 9(8)/2] = 16$, in agreement with the value obtained directly in Example 8.1.

5. *Convenience Through Ordering.* It is convenient to compute the number of concordant pairs, K' by first rearranging the (X_i, Y_i) pairs so that the (new) X's are in increasing order. Then, after rearrangement, K' is equal to the number of pairs for which the corresponding Y's are in increasing order. For example, suppose our observations are

i	1	2	3	4	5
X_i	4.1	−2.4	−2.2	−5.6	5.5
Y_i	2.3	3.7	1.1	2.2	3.8

We arrange these so that the X's are in increasing order and obtain the following:

X	−5.6	−2.4	−2.2	4.1	5.5
Y	2.2	3.7	1.1	2.3	3.8

Then, proceeding from left to right, we find the Y pairs that are in increasing order to be (2.2, 3.7), (2.2, 2.3), (2.2, 3.8), (3.7, 3.8), (1.1, 2.3). (1.1, 3.8), and (2.3, 3.8). Thus, $K' = 7$ and $K = 2K' - \{5(4)/2\} = 4$.

6. *Derivation of Distribution of K under H_0 (No-Ties Case).* Let R_i be the rank X_i in the joint ranking of $X_1, \ldots, X_n$ and let S_i be the rank of Y_i in the joint ranking of $Y_1, \ldots, Y_n$. It is clear that knowledge of the R's and S's is sufficient to calculate K (8.6). (See Problem 2.) We use this fact to illustrate how the null distribution of K can be obtained. Without loss of generality, we take

$R_1 = 1, \ldots, R_n = n$; then, as under H_0 (8.1), all possible $n!$ $(S_1, S_2, \ldots, S_n)$ Y-rank configurations are equally likely, implying each has probability $(1/n!)$.

Let us consider the case $n = 4$. In the following table, we display the $4! = 24$ possible (S_1, S_2, S_3, S_4) configurations, the associated values of K, and the corresponding null probabilities.

(R_1, R_2, R_3, R_4)	(S_1, S_2, S_3, S_4)	Null probability	K
(1, 2, 3, 4)	(1, 2, 3, 4)	$\frac{1}{24}$	6
(1, 2, 3, 4)	(1, 2, 4, 3)	$\frac{1}{24}$	4
(1, 2, 3, 4)	(1, 3, 2, 4)	$\frac{1}{24}$	4
(1, 2, 3, 4)	(1, 3, 4, 2)	$\frac{1}{24}$	2
(1, 2, 3, 4)	(1, 4, 2, 3)	$\frac{1}{24}$	2
(1, 2, 3, 4)	(1, 4, 3, 2)	$\frac{1}{24}$	0
(1, 2, 3, 4)	(2, 1, 3, 4)	$\frac{1}{24}$	4
(1, 2, 3, 4)	(2, 1, 4, 3)	$\frac{1}{24}$	2
(1, 2, 3, 4)	(2, 3, 1, 4)	$\frac{1}{24}$	2
(1, 2, 3, 4)	(2, 3, 4, 1)	$\frac{1}{24}$	0
(1, 2, 3, 4)	(2, 4, 1, 3)	$\frac{1}{24}$	0
(1, 2, 3, 4)	(2, 4, 3, 1)	$\frac{1}{24}$	−2
(1, 2, 3, 4)	(3, 1, 2, 4)	$\frac{1}{24}$	2
(1, 2, 3, 4)	(3, 1, 4, 2)	$\frac{1}{24}$	0
(1, 2, 3, 4)	(3, 2, 1, 4)	$\frac{1}{24}$	0
(1, 2, 3, 4)	(3, 2, 4, 1)	$\frac{1}{24}$	−2
(1, 2, 3, 4)	(3, 4, 1, 2)	$\frac{1}{24}$	−2
(1, 2, 3, 4)	(3, 4, 2, 1)	$\frac{1}{24}$	−4
(1, 2, 3, 4)	(4, 1, 2, 3)	$\frac{1}{24}$	0
(1, 2, 3, 4)	(4, 1, 3, 2)	$\frac{1}{24}$	−2
(1, 2, 3, 4)	(4, 2, 1, 3)	$\frac{1}{24}$	−2
(1, 2, 3, 4)	(4, 2, 3, 1)	$\frac{1}{24}$	−4
(1, 2, 3, 4)	(4, 3, 1, 2)	$\frac{1}{24}$	−4
(1, 2, 3, 4)	(4, 3, 2, 1)	$\frac{1}{24}$	−6

Thus, for example, the probability is $\frac{5}{24}$ under H_0 that K is equal to 2, because $K = 2$ when any of the five outcomes $(S_1, S_2, S_3, S_4) = (1, 3, 4, 2)$, (1, 4, 2, 3), (2, 1, 4, 3), (2, 3, 1,4), or (3, 1, 2, 4) occurs and each of these outcomes has null probability $\frac{1}{24}$. Simplifying, we obtain the null distribution

Possible value of K	Probability under H_0
-6	$\frac{1}{24}$
-4	$\frac{3}{24}$
-2	$\frac{5}{24}$
0	$\frac{6}{24}$
2	$\frac{5}{24}$
4	$\frac{3}{24}$
6	$\frac{1}{24}$

The probability, under H_0, that K is greater than or equal to 2, for example, is therefore

$$\begin{aligned} P_0(K \geq 2) &= P_0(K = 2) + P_0(K = 4) + P_0(K = 6) \\ &= \frac{5}{24} + \frac{3}{24} + \frac{1}{24} = \frac{3}{8} = .375. \end{aligned}$$

This agrees with the upper-tail probability for $n = 4$ and the value $K = 2$ when using `pKendall`.

Note that we have derived the null distribution of K without specifying the form of the underlying independent X and Y populations under H_0 beyond the point of requiring that they be continuous. That is why the test procedures based on K are called *distribution-free procedures*. From the null distribution of K, we can determine the critical value k_α and control the probability α of falsely rejecting H_0 when H_0 is true, and this error probability does not depend on the specific forms of the underlying continuous and independent X and Y distributions.

7. *Calculation of the Mean and Variance of K under the Null Hypothesis.* Displays (8.11) and (8.12) present formulas for the mean and variance of K when the null hypothesis is true. In this comment, we illustrate a direct calculation of $E_0(K)$ and $\text{var}_0(K)$ in the particular case of $n = 4$, using the null distribution of K obtained in Comment 6. (Later, in Comment 10, we present general derivations of $E_0(K)$ and $\text{var}_0(K)$.) The null mean, $E_0(K)$, is obtained by multiplying each possible value of K with its probability under H_0 and summing the products. Thus,

$$\begin{aligned} E_0(K) &= -6\left(\frac{1}{24}\right) - 4\left(\frac{3}{24}\right) - 2\left(\frac{5}{24}\right) + 0\left(\frac{6}{24}\right) + 2\left(\frac{5}{24}\right) \\ &\quad + 4\left(\frac{3}{24}\right) + 6\left(\frac{1}{24}\right) \\ &= 0. \end{aligned}$$

This is in agreement with the value stated in (8.7). A check on the expression for $\text{var}_0(K)$ is also easily performed, using the well-known fact that

$$\text{var}_0(K) = E_0(K^2) - \{E_0(K)\}^2.$$

The value of $E_0(K^2)$, the second moment of the null distribution of K, is again obtained by multiplying possible values (in this case of K^2) by the corresponding probabilities under H_0 and summing. We find

$$E_0(K^2) = \left[(36+36)\left(\frac{1}{24}\right) + (16+16)\left(\frac{3}{24}\right) + (4+4)\left(\frac{5}{24}\right) + 0\left(\frac{6}{24}\right)\right]$$
$$= \frac{26}{3}.$$

Thus,

$$\text{var}_0(K) = \frac{26}{3} - (0)^2 = \frac{26}{3},$$

which agrees with what we obtain using (8.12) directly, namely,

$$\text{var}_0(K) = \frac{4(4-1)(2(4)+5)}{18} = \frac{26}{3}.$$

8. *Symmetry of the Distribution of K under the Null Hypothesis.* When H_0 is true, the distribution of K is symmetric about its mean 0 (see Comment 6 for verification of this when $n = 4$). This implies that

$$P_0(K \leq -x) = P_0(K \geq x), \tag{8.20}$$

for all x. Equation (8.20) is used directly to convert upper-tail probabilities to lower-tail probabilities. In particular, it follows from (8.20) that the lower α percentile for the null distribution of $\overline{K}$ is $-k_\alpha$, thus the use of $-k_\alpha$ as the critical value in procedure (8.9).

9. *Possible Values for K.* If $n = 4j$ or $n = 4j+1, j = 0, 1, \ldots$, the statistic K (8.6) is always an even integer. Similarly, if $n = 4j+2$ or $n = 4j+3$, $j = 0, 1, \ldots,$ K is always an odd integer. The fact that K can assume only every other integer follows from the counting procedure used to define K (see (8.5) and (8.6)). The even or odd property of K for specific sample sizes can be deduced from the relation $K = 2K' - \{n(n-1)/2\}$ and the fact that $n(n-1)/2$ is an even integer (the product of an odd and an even integer) when $n = 4j$ or $n - 4j + 1$, $j - 0, 1, \ldots$, and is an odd integer (the product of two odd integers) when $n = 4j+2$ or $n = 4j+3$, $j = 0, 1, \ldots$.

10. *Large-Sample Approximation.* From the counting representation for K in (8.5) and (8.6), we see immediately that

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

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

$$= \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} [P\{(Y_2 - Y_1)(X_2 - X_1) > 0\} - P\{(Y_2 - Y_1)(X_2 - X_1) < 0\}],$$

which, because the X and Y variables are continuous, yields

$$E(K) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} [2P\{(Y_2 - Y_1)(X_2 - X_1) > 0\} - 1]$$

$$= \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \tau = \binom{n}{2} \tau, \tag{8.21}$$

from expression (8.2) for τ. The value of τ is 0 if X and Y are independent, so it follows that the expected value of K under H_0 is 0, as noted in (8.11). For the variance of K, we can use a well-known expression for the variance of a sum of random variables to obtain

$$\text{var}(K) = \left[\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \text{var}(Q_{ij}) + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \sum_{\substack{s=1 \\ (i,j) \neq (s,t)}}^{n-1} \sum_{t=s+1}^{n} \text{cov}(Q_{ij}, Q_{st}) \right], \tag{8.22}$$

where $Q_{uv} = Q((X_u, Y_u), (X_v, Y_v))$, for $1 \leq u < v \leq n$.

After considerable tedious calculation, we can show that (8.22) simplifies to

$$\text{var}(K) = [n(n-1)] \left[\frac{1}{2}(1 - \tau^2) + 4(n-2) \left\{ \delta - \left(\frac{\tau + 1}{2} \right)^2 \right\} \right], \tag{8.23}$$

where τ is given in (8.2) and

$$\delta = P\{(Y_2 - Y_1)(X_2 - X_1) > 0 \quad \text{and} \quad (Y_3 - Y_1)(X_3 - X_1) > 0\}. \tag{8.24}$$

Using a mutually exclusive breakdown of the event in δ (8.24) similar to that in (8.2), we see that

$$\begin{aligned} \delta = [&P\{Y_2 > Y_1, X_2 > X_1, Y_3 > Y_1, X_3 > X_1\} \\ &+ P\{Y_2 > Y_1, X_2 > X_1, Y_3 < Y_1, X_3 < X_1\} \\ &+ P\{Y_2 < Y_1, X_2 < X_1, Y_3 > Y_1, X_3 > X_1\} \\ &+ P\{Y_2 < Y_1, X_2 < X_1, Y_3 < Y_1, X_3 < X_1\}] \\ = [&P\{Y_1 < \min(Y_2, Y_3), X_1 < \min(X_2, X_3)\} \\ &+ P\{Y_3 < Y_1 < Y_2, X_3 < X_1 < X_2\} \\ &+ P\{Y_2 < Y_1 < Y_3, X_2 < X_1 < X_3\} \\ &+ P\{Y_1 > \max(Y_2, Y_3), X_1 > \max(X_2, X_3)\}]. \end{aligned} \tag{8.25}$$

When X and Y are independent variables (under H_0), (8.25) simplifies to

$$\begin{aligned}\delta_0 = [&P_0\{Y_1 < \min(Y_2, Y_3)\}P_0\{X_1 < \min(X_2, X_3)\} \\ &+ P_0(Y_3 < Y_1 < Y_2)P_0(X_3 < X_1 < X_2) \\ &+ P_0(Y_2 < Y_1 < Y_3)P_0(X_2 < X_1 < X_3) \\ &+ P_0\{Y_1 > \max(Y_2, Y_3)\}P_0\{X_1 > \max(X_2, X_3)\}]. \end{aligned} \tag{8.26}$$

However, X_1, X_2, and X_3 are mutually independent, identically distributed random variables, as are Y_1, Y_2, and Y_3. Thus, we know that

$$P_0\{Y_1 < \min(Y_2, Y_3)\} = P_0\{X_1 < \min(X_2, X_3)\} = \frac{1}{3}, \tag{8.27}$$

$$P_0\{Y_1 > \max(Y_2, Y_3)\} = P_0\{X_1 > \max(X_2, X_3)\} = \frac{1}{3}, \tag{8.28}$$

and

$$\begin{aligned} P_0(X_3 < X_1 < X_2) = P_0(X_2 < X_1 < X_3) = P_0(Y_3 < Y_1 < Y_2) \\ = P_0(Y_2 < Y_1 < Y_3) = \frac{1}{6}. \end{aligned} \tag{8.29}$$

Combining (8.26), (8.27), (8.28), and (8.29), we obtain

$$\delta_0 = \left[\frac{1}{3}\left(\frac{1}{3}\right) + \frac{1}{6}\left(\frac{1}{6}\right) + \frac{1}{6}\left(\frac{1}{6}\right) + \frac{1}{3}\left(\frac{1}{3}\right)\right] = \frac{10}{36}. \tag{8.30}$$

From (8.23) and (8.30) and the fact that $\tau = 0$ when X and Y are independent, it follows that the null variance of K is given by

$$\begin{aligned}\text{var}_0(K) &= [n(n-1)]\left[\frac{1}{2}(1-0)^2 + 4(n-2)\left\{\frac{10}{36} - \left(\frac{0+1}{2}\right)^2\right\}\right] \\ &= [n(n-1)]\left[\frac{1}{2} + \frac{1}{9}(n-2)\right] = \frac{n(n-1)(2n+5)}{18}, \end{aligned}$$

as previously noted in (8.12).

The asymptotic normality under both H_0 and general alternatives of the standardized form

$$K^* = \frac{K - E_0(K)}{\{\text{var}_0(K)\}^{1/2}} = \frac{K}{\left\{\dfrac{n(n-1)(2n+5)}{18}\right\}^{1/2}}$$

follows from Hoeffding's (1948a) U-statistic theorem applied to the bivariate setting. (For additional details, see Example 3.6.12 in Randles and Wolfe (1979).)

11. *Ties within the X-Values and/or Y-Values.* We have recommended dealing with tied X observations and/or tied Y observations by counting a zero in the Q^* (8.17) counts leading to the computation of K (8.6). This approach is satisfactory as long as the number of (X, Y) pairs containing a tied X and/or tied Y observation does not represent a sizable percentage of the total number (n) of sample pairs.

    We should, however, point out that methods other than this zero assignment to the Q^* (8.17) counts have been considered for dealing with tied X and/or tied Y observations. One could use individual randomization (e.g., flipping a fair coin) to decide whether each of the tied pairs (X or Y) is to be counted as a $+1$ (i.e., as a concordant pair—see Comment 3) or as a -1 (i.e., as a discordant pair—again, see Comment 3) in the computation of K (8.6). (Although this approach maintains many of the nice properties of K that hold when there are no tied X and/or tied Y observations, it introduces extraneous randomness that could quite easily have a direct effect on the outcome of any subsequent inferences based on such a modified value of K.) A second alternative approach in the case of the one-sided test procedures in (8.8), (8.9), (8.14), and (8.15) is to be conservative about rejecting the null hypothesis H_0; that is, we could count all the tied X and/or tied Y observations as if they were in favor of not rejecting H_0. Thus, for example, in applying either procedure (8.8) or (8.14) to test H_0 against the alternative $\tau > 0$, we would treat *all* the pairs of pairs involving tied X and/or tied γ observations as if they were discordant pairs (in favor of not rejecting H_0) leading to Q (8.5) counts of -1 in the calculation of K. (In the case of procedures (8.9) and (8.15), all the pairs of pairs involving tied X and/or tied Y observations would be considered as concordant pairs—again in favor of not rejecting H_0—leading to Q (8.5) counts of $+1$ in the calculation of K.) Any rejection of H_0 with this conservative approach to dealing with tied X and/or tied Y observations could then be viewed as providing strong evidence in favor of the appropriate alternative. For more detailed discussion of methods for handling tied X and/or tied Y observations, see Sillitto (1947), Smid (1956), Burr (1960), and Kendall (1962).

12. *Some Power Results for the Kendall Test for Independence.* We consider the upper-tail α-level test of H_0 (8.1) versus $H_1 : \tau > 0$ given by procedure (8.8). The power, or probability of correctly rejecting H_0, for τ (8.2) values "near" the null hypothesis value of 0 can be approximated by

$$\text{Power} \doteq \Phi(A_F), \tag{8.31}$$

    where $\Phi(A_F)$ is the area under a standard normal density to the left of the point

$$A_F = \{[9n(n-1)/(4n+10)]^{1/2}\tau - z_\alpha\}. \tag{8.32}$$

    When $F_{X,Y}$ is the bivariate normal distribution with correlation coefficient ρ, it follows that $\tau = \frac{2}{\pi}\sin^{-1}(\rho)$ (see, for example, Gibbons and Chakraborti (2010)). Thus, when $F_{X,Y}$ is bivariate normal the approximate power depends only on the value of ρ. For purposes of illustration, suppose that the common underlying distribution is bivariate normal with $\rho = .4$. For the case of $n = 9$ and $\alpha = .060$, the test rejects H_0 if and only if $K \geq 16$, or, equivalently, $\overline{K} \geq 16/36$.

Substituting $\tau = (2/\pi)\sin^{-1}(.4) = (2/\pi)(.4115) = .2620$ in (8.32), we obtain

$$\begin{aligned} A_{\text{BIV NOR}} &= \{[9(9)(8)/2(2(9)+5)]^{1/2}.2620 - 1.555\} \\ &= \{[14.09]^{1/2}(.2620) - 1.555\} = \{.9835 - 1.555\} = -.57. \end{aligned}$$

Thus, the approximate power of this test for a bivariate normal distribution with $\rho = .4$ (and *any* means and variances) is

$$\text{Power} \doteq \Phi(-.57) = 1 - \Phi(.57) = 1 - (1 - .28) = .28.$$

This compares with the simulation estimated exact power of .35 for $n = 9$, $\rho = .4$, and $\alpha = .05$, as given in Table 8.3 of Bhattacharyya, Johnson, and Neave (1970). Additional simulation estimated exact power values for the one-sided Kendall test and sample sizes $n = 5,7,9$ and significance levels $\alpha = .01$ and .05 can be found in Bhattacharyya, Johnson, and Neave (1970) for bivariate normal and bivariate exponential distributions.

13. *Sample Size Determination.* Noether (1987) shows how to determine an approximate sample size n so that the α-level one-sided test given by procedure (8.8) will have approximate power $1 - \beta$ against an alternative value of τ (8.2) greater than zero. This approximate value of n is

$$n \doteq \frac{4(z_\alpha + z_\beta)^2}{9\tau^2}. \tag{8.33}$$

As an illustration of the use of (8.33), suppose we are testing H_0 and we desire to have an upper-tail level $\alpha = .010$ test with power $1 - \beta$ at least .90 against an alternative bivariate distribution for which $\tau = .4$. Using $z_\alpha = z_{.01} = 2.326$ and $z_\beta = z_{.10} = 1.282$, we find that the approximate required sample size for the alternative $\tau = .4$ is

$$n \doteq \frac{4(2.326 + 1.282)^2}{9(.4)^2} = 36.2.$$

To be conservative, we would take $n = 37$.

14. *Trend Test.* If we take $X_i = i,\ i = 1, \ldots, n$ and consider

$$\begin{aligned} K &= \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Q((i, Y_i), (j, Y_j)) \\ &= \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} c(Y_j - Y_i), \end{aligned}$$

where

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

then K can be used as a test for a time trend in the univariate random sample $Y_1, \ldots, Y_n$. This use of K to test for a time trend was suggested by Mann (1945).

15. *Other Uses for the K Statistic.* The Wilcoxon rank sum test (Section 4.1) and the Jonckheere–Terpstra test (Section 6.2) can be viewed as tests based on K (8.6) (or, equivalently, $\hat{\tau}$ (8.34)). For this interpretation see Jonckheere (1954a) and Kendall (1962, Sections 3.12 and 13.9). Also, Wolfe (1977) has used the K statistic to compare the correlation between variables X_2 and X_1 with that between the variables X_3 and X_1, when both X_2 and X_3 are potential predictors for X_1.
16. *Consistency of the K Test.* Under the assumption that $(X_1, Y_1), \ldots, (X_n, Y_n)$ is a random sample from a continuous bivariate population with joint distribution function $F_{X,Y}(x, y)$, the consistency of the tests based on K depends on the parameter τ (8.2). The test procedures defined by (8.8), (8.9), and (8.10) are consistent against the class of alternatives corresponding to $\tau >, <$, and $\neq 0$, respectively.
17. *Multivariate Concordance.* Joe (1990) has generalized Kendall's measure of association τ from the bivariate case where τ measures the strength of association between two variables X, Y to the multivariate case where $\mathbf{X} = (X_1, \ldots, X_m)$ is an m-dimensional random variable and one is interested in a measure of the strength of the association between the components $X_1, \ldots, X_m$ of $\mathbf{X}$. Let F denote the joint distribution function of $\mathbf{X}$,

$$F(x_1, \ldots, x_m) = P(X_1 \leq x_1 \text{ and } X_2 \leq x_2 \text{ and } \ldots \text{ and } X_m \leq x_m)$$

and denote the marginal distribution functions as $F_j(x_j) = P(X_j \leq x_j), j = 1, \ldots, m$. The null hypothesis of mutual independence of $X_1, \ldots, X_m$ is

$$H_0 : F(x_1, \ldots, x_m) = \prod_{j=1}^{m} F_j(x_j), \quad \text{for all}(x_1, \ldots, x_m).$$

That is, the joint distribution is equal to the product of the marginals.

Joe has defined a class of measures of the strength of association between $X_1, X_2, \ldots, X_m$. Let $\mathbf{X}_i = (X_{i1}, \ldots, X_{im}), i = 1, 2$ be two independent m-dimensional random variables each with joint distribution function F. One member of Joe's class reduces to $\overline{\tau}$, the average of all pairwise τ's. The measure $\overline{\tau}$ was introduced by Hays (1960) and is given by

$$\overline{\tau} = \sum_{u=1}^{m-1} \sum_{v=1}^{m} \frac{\tau_{uv}}{\left\{\dfrac{m(m-1)}{2}\right\}},$$

where

$$\begin{aligned}\tau_{uv} &= P\{(X_{1u} - X_{1v})(X_{2u} - X_{2v}) > 0\} - P\{(X_{1u} - X_{1v})(X_{2u} - X_{2v}) < 0\} \\ &= 2P\{X_{1u} - X_{1v})(X_{2u} - X_{2v}) > 0\} - 1.\end{aligned}$$

Joe has also generalized Spearman's measure (see Section 8.5) and a measure due to Blomqvist (1950).

Properties

1. *Consistency*. The tests defined by (8.8), (8.9), and (8.10) are consistent against the alternatives $\tau >, <$, and $\neq 0$, respectively.
2. *Asymptotic Normality*. See Hoeffding (1948a) or Randles and Wolfe (1979, pp. 108–109).
3. *Efficiency*. See Section 8.7.

Problems

1. The data in Table 8.3 are a subset of the data obtained by Featherston (1971). Among other things, he was interested in the relationship between the weight of tapeworms *(Taenia hydatigena)* fed to dogs and the weight of the scoleces recovered from the dogs after 20 days. (A scolex is the attachment end of a tapeworm, consisting of the head and neck.) The cysticerci used in the experiment were collected from sheep carcasses and force-fed to 10 dogs via gelatine capsules. The scoleces were recovered from each dog at autopsy, 20 days after the introduction of the tapeworms. Table 8.3 gives the mean weight of the initial cysticerci and the mean weight of the recovered worms for each of the 10 dogs in the study.

Test the hypothesis of independence versus the alternative that the mean weight of introduced cysticerci is positively correlated with the mean weight of worms recovered.

2. Let R_i be the rank of X_i in the joint ranking of $X_1, \ldots, X_n$ and let S_i be the rank of Y_i in the joint ranking of $Y_1, \ldots, Y_n$. Show that knowledge of $R_1, \ldots, R_n$ and $S_1, \ldots, S_n$ is sufficient to calculate K (8.6).

3. The data in Table 8.4 are a subset of the data obtained by Sylvester (1969) in a study concerned with the anatomical and pathological status of the corticospinal and somatosensory tracts and parietal lobes of patients who had had cerebral palsy. Among other things, he was interested in the relationship between brain weights and large fiber (>7.5 μ in diameter) counts in the medullary pyramid. Table 8.4 gives the mean brain weights (g) and medullary pyramid large fiber counts for 11 cerebral palsy subjects. Test the hypothesis of independence versus the general alternative that brain weight and large fiber count in the medullary pyramid are correlated in subjects who have had cerebral palsy.

Table 8.3 Relation Between Weight of the Cysticerci of *Taenia hydatigena* Fed to Dogs and Weight of Worms Recovered at 20 Days

	Mean weight, mg	
Dog	Cysticerci	Worms recovered
1	28.9	1.0
2	32.8	7.7
3	12.0	7.3
4	9.9	7.9
5	15.0	1.1
6	38.0	3.5
7	12.5	18.9
8	36.5	33.9
9	8.6	28.6
10	26.8	25.0

Source: D. W. Featherston (1971).

Table 8.4 Mean Brain Weights and Medullary Pyramid Large Fiber Counts for Cerebral Palsy Subjects

Subject number	Brain weight, g	Pyramidal large fiber count
1	515	32,500
2	286	26,800
3	469	11,410
4	410	14,850
5	461	23,640
6	436	23,820
7	479	29,840
8	198	21,830
9	389	24,650
10	262	22,500
11	536	26,000

Source: P. E. Sylvester (1969).

4. Let (X_1, Y_1) and (X_2, Y_2) be independent and identically distributed continuous bivariate random variables with joint probability density function

$$f_{X,Y}(x,y) = \begin{cases} e^{-y}, & 0 < x < y < \infty, \\ 0, & \text{elsewhere.} \end{cases}$$

Calculate the value of τ for this bivariate distribution.

5. Let (X_1, Y_1) and (X_2, Y_2) be independent and identically distributed discrete bivariate random variables with joint probability function

$$f_{X,Y}(x,y) = \begin{cases} \dfrac{x+y}{21}, & x = 1,2,3; \quad y = 1,2, \\ 0, & \text{elsewhere.} \end{cases}$$

Calculate the value of τ for this bivariate distribution.

6. Let (X_1, Y_1) and (X_2, Y_2) be independent and identically distributed continuous bivariate random variables with joint probability density function

$$f_{X,Y}(x,y) = \begin{cases} \dfrac{1}{2}y^2 e^{-x-y}, & 0 < x < \infty, \quad 0 < y < \infty, \\ 0, & \text{elsewhere.} \end{cases}$$

Calculate the value of τ for this bivariate distribution.

7. The data in Table 8.5 are a subset of the data considered by Clark, Vandenberg, and Proctor (1961) in a study concerned with the relationship of scores on various psychological tests and certain physical characteristics of twins. Table 8.5 gives the test scores (totals of a number of different psychological tests) of 13 dizygous (i.e., nonidentical) male twins. Test the hypothesis of independence versus the alternative that the twins' test scores are positively correlated.

8. Previously, it was shown that if X and Y are independent random variables, then τ (8.2) has a value of 0. Show that the converse is not true by constructing a joint probability distribution for the pair of random variables X and Y such that $\tau = 0$ but X and Y are not independent.

9. If we have 25 bivariate observations and $F_{X,Y}$ is bivariate normal with correlation coefficient .3, what is the approximate power of the level $\alpha = .045$ test of H_0 (8.1) versus the alternative $\tau > 0$?

Table 8.5 Psychological Test Scores of Dizygous Male Twins

Pair i	Twin X_i	Twin 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).

10. For an arbitrary number, n, of bivariate observations, what are the smallest and largest values of K? Give examples of data sets where these extremes are achieved.

11. Give an example of a data set of $n \geq 10$ bivariate observations for which K has value 0. (Consider Comment 9.)

12. Suppose $n = 20$. Compare the critical region for the exact level $\alpha = .05$ test of H_0 (8.1) versus $H_2 : \tau < 0$ based on K with the critical region for the corresponding nominal level $\alpha = .05$ based on the large-sample approximation. What is the exact significance level of this .05 nominal level test based on the large-sample approximation?

13. Consider a level $\alpha = .10$ test of H_0 (8.1) versus the alternative $\tau > 0$ based on K. How many bivariate observations (n) will we need to collect in order to have approximate power at least .95 against an alternative for which $\tau = .6$?

14. A question of significance to state legislators working with tight budgets is the spending for secondary education. The data in Table 8.6 are from the Department of Education, National Center for Education Statistics and were considered by Merline (1991) in assessing the relationship between the amount of money spent on secondary education and various performance criteria for high-school seniors. Table 8.6 gives the spending ($) per high-school senior and the percentage of those seniors who graduated for each of the 50 states in the 1987–1988 school year.

Use the large-sample approximation to test the hypothesis of independence versus the alternative that spending per high-school senior and the percentage of seniors graduating are positively correlated. (Discuss any other social or economic factors that might impact on these data and, thereby, on the conclusion from this statistical analysis.)

15. For the case of $n = 5$ untied bivariate (X, Y) observations, obtain the form of the exact null (H_0) distribution of K. (See Comment 6.)

16. Johnson (1973) studied several different managerial aspects of university associated schools of nursing. Among the data she collected were the "extent of agreement (between the dean and the faculty) on the responsibilities for decision making" and "faculty satisfaction." The ranks on the two variables for the 12 institutions that were involved in Johnson's study are presented in Table 8.7.

Test the hypothesis of independence versus the alternative that faculty/dean decision-making agreement and faculty satisfaction are negatively correlated in university schools of nursing. (*Note:* Low ranks are associated with poor faculty satisfaction and little faculty/dean decision-making agreement, respectively.)

17. Consider the test of H_0 (8.1) versus $H_1 : \tau > 0$ based on K for the following $n = 10$ (X, Y) observations: (1.5, 6), (1.9, 4), (2.3, 6), (2.7, 12), (1.5, 13), (1.8, 16), (3.6, 16), (4.2, 9), (4.7,

Table 8.6 Spending per High-School Senior and the Percentage of Those Seniors Who Graduated during the 1987–1988 School Year

State	$ per Pupil	% Graduated	State	$ per Pupil	% Graduated
Alaska	7971	65.5	Ohio	3998	79.6
New York	7151	62.3	Nebraska	3943	85.4
New Jersey	6564	77.4	Hawaii	3919	69.1
Connecticut	6230	84.9	West Virginia	3858	77.3
Massachusetts	5471	74.4	California	3840	65.9
Rhode Island	5329	69.8	Indiana	3794	76.3
Vermont	5207	78.7	Missouri	3786	74.0
Maryland	5201	74.1	Arizona	3744	61.1
Wyoming	5051	88.3	New Mexico	3691	71.9
Delaware	5017	71.7	Nevada	3623	75.8
Pennsylvania	4989	78.4	Texas	3608	65.3
Oregon	4789	73.0	North Dakota	3519	88.3
Wisconsin	4747	84.9	Georgia	3434	61.0
Michigan	4692	73.6	South Carolina	3408	64.6
Colorado	4462	74.7	North Carolina	3368	66.7
New Hampshire	4457	74.1	South Dakota	3249	79.6
Minnesota	4386	90.9	Louisiana	3138	61.4
Illinois	4369	75.6	Oklahoma	3093	71.7
Maine	4246	74.4	Tennessee	3068	69.3
Montana	4246	87.3	Kentucky	3011	69.0
Washington	4164	77.1	Arkansas	2989	77.2
Virginia	4149	71.6	Alabama	2718	74.9
Iowa	4124	85.8	Idaho	2667	75.4
Florida	4092	58.0	Mississippi	2548	66.9
Kansas	4076	80.2	Utah	2454	79.4

Source: J. W. Merline (1991).

Table 8.7 Rankings for Faculty/Dean Decision-Making Agreement and Faculty Satisfaction for Participating Schools of Nursing

School	Rank for faculty/dean decision-making agreement	Rank for faculty satisfaction
1	8	8
2	9	2
3	6	10
4	12	5
5	1	12
6	11	4
7	10	6
8	2	9
9	4	7
10	5	3
11	7	11
12	3	1

Source: B. M. Johnson (1973).

0), and (4.0, 3). Compute the P-values for the competing K-procedures based on either (a) using Q^* (8.17) counts of zero, as recommended in the Ties portion of this section, or (b) dealing with the tied X and tied Y observations in a conservative manner, as presented in Comment 11. Discuss the results.

18. In Comment 4, we noted that we have $K = 2K' - \{n(n-1)/2\}$, where $K' =$ (number of concordant pairs), when there are neither tied X nor tied Y observations. Obtain the corresponding expression for the relationship between K and K' when there are no tied X pairs and a total of t $(\neq 0)$ tied Y pairs (among the $\binom{n}{2}$ total Y pairs), and we use the Q^* (8.17) counts of zero to deal with the tied Y pairs. How does this expression change if there are no tied Y pairs and t tied X pairs? Discuss the necessary changes in the expression relating K and K' when there are s $(\neq 0)$ tied X pairs and t $(\neq 0)$ tied Y pairs.

19. Gerstein (1965) studied the long-term pollution of Lake Michigan and its effect on the water supply for the city of Chicago. One of the measurements considered by Gerstein was the annual number of "odor periods" over the period of years 1950–1964. Table 8.8 contains this information for Lake Michigan for each of these 15 years.

Test the hypothesis that the degree of pollution (as measured by the number of odor periods) had not changed with time against the alternative that there was a general increasing trend in the pollution of Lake Michigan over the period 1950–1964. (See Comment 14.)

8.2 AN ESTIMATOR ASSOCIATED WITH THE KENDALL STATISTIC (KENDALL)

Procedure

The estimator of the Kendall population correlation coefficient τ (8.2), based on the statistic K (8.6),is

$$\hat{\tau} = \frac{2K}{n(n-1)} = \overline{K}. \tag{8.34}$$

The statistic $\hat{\tau}$ is known as *Kendall's sample rank correlation coefficient* and appropriately assumes values between -1 and 1 inclusive.

Table 8.8 Annual Number of Odor Periods for Lake Michigan for the Period 1950–1964

Year	Number of odor periods
1950	10
1951	20
1952	17
1953	16
1954	12
1955	15
1956	13
1957	18
1958	17
1959	19
1960	21
1961	23
1962	23
1963	28
1964	28

Source: H. H. Gerstein (1965).

EXAMPLE 8.2 ***(Continuation of Example 8.1).***

For the canned tuna data of Table 8.1, we see from (8.34) that the sample estimate of τ is

$$\hat{\tau} = \frac{2(16)}{9(8)} = \frac{4}{9}. \tag{8.35}$$

This estimate is also found in the R output in Example 8.1.

Comments

18. *Ties.* In the presence of ties, use $\hat{\tau} = 2K/n(n-1)$, where

$$K = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Q^*((X_i, Y_i), (X_j, Y_j)) \tag{8.36}$$

and $Q^*((X_i, Y_i), (X_j, Y_j))$ is defined by (8.17).

19. *Probability Estimation.* For many problems, distribution-free test statistics are used directly to estimate basic probability parameters other than the usual distributional parameters associated with the corresponding normal theory problems. In particular, note that $\hat{\tau} = 2K/[n(n-1)]$ estimates the probability parameter τ (8.2) rather than the usual correlation coefficient for the underlying bivariate population. Estimators of such readily interpretable parameters are very helpful in data analysis. (See Crouse (1966), Wolfe and Hogg (1971), and Comment 4.18.)

Properties

1. *Standard Deviation of* $\hat{\tau}$. For the asymptotic standard deviation of $\hat{\tau}$ (8.34), see Noether (1967a, p. 78), Fligner and Rust (1983), Samara and Randles (1988), and Comment 25.
2. *Asymptotic Normality.* See Hoeffding (1948a) and Randles and Wolfe (1979, pp. 108–109).

Problems

20. Estimate τ for the tapeworm data of Table 8.3.

21. What is the maximum possible value of $\hat{\tau}$ when there are no tied X and/or tied Y observations? What is the minimum possible value of $\hat{\tau}$ when there are no tied X and/or tied Y observations? Construct three examples with $n \geq 10$ with no tied X and/or tied Y observations: one in which $\hat{\tau}$ achieves its maximum value, one in which it achieves its minimum value, and one for which $\hat{\tau} = 0$.

22. Estimate τ for the cerebral palsy data of Table 8.4.

23. Estimate τ for the twin data of Table 8.5.

24. Estimate τ for the secondary education data of Table 8.6.

25. Use Comment 17 to redo Problem 21 for the case when there are t ($\neq 0$) tied X pairs (among the total of $\binom{n}{2}$ X pairs) and no tied Y observations. How is the answer affected if there are no tied X observations and $t(\neq 0)$ tied Y pairs? if there are s ($\neq 0$) tied X pairs and t ($\neq 0$) tied Y pairs?

26. Estimate τ for the nursing faculty data of Table 8.7.

8.3 AN ASYMPTOTICALLY DISTRIBUTION–FREE CONFIDENCE INTERVAL BASED ON THE KENDALL STATISTIC (SAMARA–RANDLES, FLIGNER–RUST, NOETHER)

Procedure

For an asymptotically distribution-free symmetric two-sided confidence interval for τ, with the approximate confidence coefficient $1-\alpha$, we first compute

$$C_i = \sum_{\substack{t=1 \\ t\neq i}}^{n} Q((X_i, Y_i), (X_t, Y_t)), \quad \text{for } i = 1, \ldots, n, \tag{8.37}$$

where $Q((a,b),(c,d))$ is given by (8.5). Let $\overline{C} = (1/n)\sum_{i=1}^{n} C_i = 2K/n$ and define

$$\hat{\sigma}^2 = \frac{2}{n(n-1)}\left[\frac{2(n-2)}{n(n-1)^2}\sum_{i=1}^{n}(C_i - \overline{C})^2 + 1 - \hat{\tau}^2\right], \tag{8.38}$$

where $\hat{\tau}$ is given by (8.34). The approximate $100(1-\alpha)\%$ confidence interval (τ_L, τ_U) for τ that is associated with the point estimator $\hat{\tau}$ (8.34) is then given by

$$\tau_L = \hat{\tau} - z_{\alpha/2}\hat{\sigma}, \quad \tau_U = \hat{\tau} + z_{\alpha/2}\hat{\sigma}. \tag{8.39}$$

With τ_L and τ_U given by display (8.39), we have

$$P_\tau\{\tau_L < \tau < \tau_U\} \approx 1-\alpha \text{ for all } \tau. \tag{8.40}$$

(For approximate upper or lower confidence bounds for τ associated with $\hat{\tau}$, see Comment 23.)

EXAMPLE 8.3 ***(Continuation of Examples 8.1 and 8.2).***

Consider the canned tuna data of Table 8.1. We illustrate how to obtain an approximate 90% symmetric two-sided confidence interval for τ. From (8.37), we see that

$$\begin{aligned} C_5 &= \sum_{j\neq 5} Q((X_5, Y_5), (X_j, Y_j)) \\ &= [Q((X_5, Y_5), (X_1, Y_1)) + Q((X_5, Y_5), (X_2, Y_2)) \\ &\quad + Q((X_5, Y_5), (X_3, Y_3)) + Q((X_5, Y_5), (X_4, Y_4)) \end{aligned}$$

$$+ Q((X_5, Y_5), (X_6, Y_6)) + Q((X_5, Y_5), (X_7, Y_7))$$
$$+ Q((X_5, Y_5), (X_8, Y_8)) + Q((X_5, Y_5), (X_9, Y_9)].$$

Using the fact that $Q((X_i, Y_i), (X_j, Y_j)) = Q((X_j, Y_j), (X_i, Y_i))$ for every $i \neq j$ and the Q counts for the canned tuna data in Table 8.2, it follows that

$$C_5 = 1 - 1 + 1 + 1 - 1 + 1 - 1 + 1 = 2.$$

Note that C_5 is simply equal to the sum of the Q values in the $j = 5$ row and the $i = 5$ column in Table 8.2. In the same way, we find

$$C_1 = 7 - 1 = 6, \quad C_2 = 6 - 2 = 4, \quad C_3 = 8 - 0 = 8, \quad C_4 = 6 - 2 = 4$$
$$C_6 = 3 - 5 = -2, \quad C_7 = 6 - 2 = 4, \quad C_8 = 6 - 2 = 4, \quad C_9 = 5 - 3 = 2.$$

Thus, we have

$$\overline{C} = \frac{1}{9}\sum_{i=1}^{9} C_i = \frac{1}{9}[6 + 4 + 8 + 4 + 2 - 2 + 4 + 4 + 2] = \frac{32}{9}.$$

Thus,

$$\sum_{i=1}^{9}(C_i - \overline{C})^2 = \sum_{i=1}^{9}\left(C_i - \frac{32}{9}\right)^2$$
$$= \left[\left(\frac{22}{9}\right)^2 + \left(\frac{4}{9}\right)^2 + \left(\frac{40}{9}\right)^2 + \left(\frac{4}{9}\right)^2 + \left(-\frac{14}{9}\right)^2\right.$$
$$\left. + \left(-\frac{50}{9}\right)^2 + \left(\frac{4}{9}\right)^2 + \left(\frac{4}{9}\right)^2 + \left(-\frac{14}{9}\right)^2\right]$$
$$= \frac{484 + 4(16) + 1600 + 2(196) + 2500}{81} = \frac{560}{9}. \tag{8.41}$$

Using the values for $\hat{\tau}$ and $\Sigma_{i=1}^{9}(C_i - \overline{C})^2$ given in (8.35) and (8.41), respectively, we see from (8.38) that

$$\hat{\sigma}^2 = \frac{2}{9(8)}\left[\frac{2(7)}{9(8)^2}\left(\frac{560}{9}\right) + 1 - \left(\frac{4}{9}\right)^2\right]$$
$$= \frac{1}{36}[1.512 + 1 - .198] = .064.$$

With $1 - \alpha = .90$ (so that $\alpha = .10$), $z_{.05} = 1.65$. Hence, from (8.39), we obtain

$$\tau_L = \frac{4}{9} - 1.65(.064)^{1/2} = .444 - .417 = .027$$

and

$$\tau_U = \frac{4}{9} + 1.65(.064)^{1/2} = .444 + .417 = .861.$$

Our approximate 90% symmetric confidence interval for τ is thus $(\tau_L, \tau_U) = (.027, .861)$.

The above results may be duplicated with the command `kendall.ci` in package `NSM3`. The arguments needed are the samples X and Y, the confidence level α, and whether the interval should be the two-sided symmetric interval described above or a one-sided interval described in Comment 24. In particular,

```
kendall.ci (x, y, alpha=.1, type="t")
```

reproduces the above bounds for this example.

Comments

20. *Interpretation as a Confidence Interval for a Probability.* The confidence interval given by (8.39) is an approximate $1 - \alpha$ confidence interval for a parameter that is a linear function of the probability $P\{(X_1 - X_2)(Y_1 - Y_2) > 0\}$. This is common practice in the field of nonparametric statistics, where probabilities are often natural and easily interpretable parameters. Recall the relation of the Wilcoxon two-sample test of Section 4.1 to the parameter $P(X < Y)$. (See Comments 4.7, 4.10, 4.14, and 4.18.)
21. *Ties.* In the presence of ties, use $Q^*((X_i, Y_i), (X_j, Y_j))$ defined by (8.17) instead of $Q((X_i, Y_i), (X_j, Y_j))$ given by (8.5) in the computation of $C_1, \ldots, C_n$ and $\hat{\tau}$.
22. *Alternative Method of Calculation.* The following equivalent formula for the term $\Sigma_{i=1}^n (C_i - \overline{C})^2$ in the definition of $\hat{\sigma}^2$ (8.38), namely,

$$\sum_{i=1}^{n} (C_i - \overline{C})^2 = \sum_{i=1}^{n} C_i^2 - \frac{4K^2}{n}, \tag{8.42}$$

    is often computationally more convenient.
23. *Concordant/Discordant Pairs Representation for the C_i's.* Let C_i' and C_i'' be the numbers of pairs (X_j, Y_j), $j \neq i$, that are concordant and discordant, respectively, with (X_i, Y_i), for $i = 1, \ldots, n$. Then, the C_i (8.37) counts can be expressed as $C_i = C_i' - C_i''$, for $i = 1, \ldots, n$.
24. *Confidence Bounds.* In many settings, we are interested only in making one-sided confidence statements about the parameter τ; that is, we wish to assert with specified confidence that τ is no larger (or, in other settings, no smaller) than some upper (lower) confidence bound based on the sample data. To obtain such one-sided confidence bounds for τ, we proceed as follows. For a specified confidence coefficient $1 - \alpha$, find z_a (not $z_{\alpha/2}$, as for the confidence interval). An approximate $100(1 - \alpha)\%$ *upper* confidence bound τ_U^* for τ is then given by

$$[-1, \tau_U^*) = [-1, \hat{\tau} + z_\alpha \hat{\sigma}), \tag{8.43}$$

    where $\hat{\tau}$ and $\hat{\sigma}^2$ are given by (8.34) and (8.38), respectively. With τ_U^* given by display (8.43), we have

$$P_\tau\{-1 \leq \tau < \tau_U^*\} \approx 1 - \alpha \text{ for all } \tau. \tag{8.44}$$

The corresponding approximate $100(1-\alpha)\%$ *lower* confidence bound τ_{L}^* for τ is given by

$$(\tau_{\mathrm{L}}^*, 1) = (\hat{\tau} - z_\alpha \hat{\sigma}, 1), \tag{8.45}$$

with

$$P_\tau\{\tau_{\mathrm{L}}^* < \tau \leq 1\} \approx 1 - \alpha \text{ for all } \tau. \tag{8.46}$$

25. *Alternative Approximate Confidence Limits.* Samara and Randles (1988) showed that the statistic $\hat{\tau}/\hat{\sigma}$ is itself distribution-free under the null hypothesis (H_0) of independence, and they tabled the upper αth percentile of its null distribution, k_α^*, for $\alpha = .005, .01, .025, .05$, and .10 and sample sizes $n = 6(1)20$. Slightly improved confidence intervals and confidence bounds can be obtained by replacing the normal percentiles $z_{\alpha/2}$ and z_α by $k_{\alpha/2}^*$ and k_α^*, respectively, in (8.39), (8.43), and (8.45).

26. *Estimating the Asymptotic Standard Deviation of* $\hat{\tau}$. The statistic $\hat{\sigma}$ (8.38) is chosen to be a consistent estimator for the asymptotic standard deviation of the point estimator $\hat{\tau}$ (8.34). It is not necessary to use all the sample observations in calculating $\hat{\sigma}$. In fact, any fixed percentage subset of the n sample observations can be employed to find the C_i values used in (8.38). For example, 25% of a random sample of n paired observations (namely, $n/4$ observations) could be used to obtain $\hat{\sigma}$.

27. *Asymptotic Coverage Probability.* Asymptotically, the true coverage probability of the interval defined by (8.39) and the bounds in (8.43) and (8.45) will agree with the nominal confidence coefficient $1 - \alpha$. Subject to Assumption A, this asymptotic (n infinitely large) result does not depend on the distribution of the underlying bivariate population. Thus, the interval given by (8.39) and the bounds in (8.43) and (8.45) have been constructed to have the asymptotically distribution-free property.

28. *Historical Development.* The initial effort at constructing asymptotically distribution-free confidence intervals and bounds for τ was due to Noether (1967a). The approximate $100(1-\alpha)\%$ confidence interval proposed by Noether is $\hat{\tau} \pm z_{\alpha/2}\hat{\sigma}_N$, where $\hat{\sigma}_N^2$ is a consistent estimator (based on U-statistics theory) of the variance of $\hat{\tau}$. However, it was later pointed out that $\hat{\sigma}_N^2$ can assume negative values, even though it is estimating the nonnegative quantity $\mathrm{var}(\hat{\tau})$. Although this distressing possibility is more likely to occur in small samples, it can be negative for sample sizes as large as $n = 30$, To avoid this problem, Fligner and Rust (1983) proposed the use of $\hat{\tau} \pm z_{\alpha/2}\hat{\sigma}_{\mathrm{FR}}$ as an asymptotically distribution-free $100(1-\alpha)\%$ confidence interval for $\hat{\tau}$, where $\hat{\sigma}_{\mathrm{FR}}^2$ is a jackknife estimator (different from $\hat{\sigma}_{\mathrm{N}}^2$) of $\mathrm{var}(\hat{\tau})$ that is consistent and cannot assume negative values. A few years later, Samara and Randles (1988) noted that although the Fligner–Rust jackknife estimator $\hat{\sigma}_{\mathrm{FR}}^2$ can never be negative, it can be zero for a variety of rank configurations, including some nonextreme cases. They suggested a final modification leading to the asymptotically distribution-free $100(1-\alpha)\%$ confidence interval in display (8.39), where $\hat{\sigma}^2$ (8.38) $= \hat{\sigma}_{\mathrm{SR}}^2$ is a third consistent estimator of $\mathrm{var}(\hat{\tau})$. The estimator $\hat{\sigma}_{\mathrm{SR}}^2 = \hat{\sigma}^2$ (8.38) is also based on U-statistics methodology (as is $\hat{\sigma}_{\mathrm{N}}^2$), but it can never be negative and is zero only in the two extreme cases where $\hat{\tau} = \pm 1$. For the approximate confidence

interval $\hat{\tau} \pm z_{\alpha/2}\hat{\sigma}$ to be simply the singleton point $\hat{\tau}(= +1 \text{ or } -1)$ in such extreme cases is not ideal, but it is also not unreasonable.

29. *Competitor Tests for Independence.* In Section 8.1, we discussed tests of independence (8.1) based on Kendall's statistic K (8.6). Noether (1967a), Fligner and Rust (1983), and Samara and Randles (1988) also proposed distribution-free tests of H_0 (8.1) based on the statistics $\hat{\tau}/\hat{\sigma}_{\mathrm{N}}$, $\hat{\tau}/\hat{\sigma}_{\mathrm{FR}}$, and $\hat{\tau}/\hat{\sigma}_{\mathrm{SR}}$, respectively, where $\hat{\tau}$ is given by (8.34) and $\hat{\sigma}^2_{\mathrm{N}}$, $\hat{\sigma}^2_{\mathrm{FR}}$, and $\hat{\sigma}^2_{SR}$ are the various consistent estimators of $\mathrm{var}(\hat{\tau})$ discussed in Comment 28. Although not generally as powerful as the procedures based on K for testing H_0 (8.1), the tests based on $\hat{\tau}/\hat{\sigma}_{\mathrm{N}}$, $\hat{\tau}/\hat{\sigma}_{\mathrm{FR}}$, and $\hat{\tau}/\hat{\sigma}_{\mathrm{SR}}$ all have the advantage (not possessed by the tests based on K) that they are also asymptotically distribution-free procedures for testing the more general null hypothesis H_0^*: $\tau = 0$.

30. *Partial Correlation Coefficients.* Let $(X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n)$ be a random sample from a continuous trivariate distribution. It is often of interest to assess the association between the X and Y variables, controlled for the third variable Z. Gripenberg (1992) proposed measuring this "partial correlation" by the parameter

$$\begin{aligned} \tau_{XY/Z} &= 2P\{(Y_2 - Y_1)(X_2 - X_1) > 0 | Z_1 = Z_2\} - 1 \\ &= E[Q((X_1, Y_1), (X_2, Y_2)) | Z_1 = Z_2], \end{aligned} \tag{8.47}$$

where $Q((X_1, Y_1), (X_2, Y_2))$ is defined by (8.5). To estimate $\tau_{XY/Z}$, Gripenberg arranged the (X_i, Y_i, Z_i) triples in an increasing order with respect to the values of the Z variable. Letting $Z_{N(1)} \leq \cdots \leq Z_{N(n)}$ denote the order statistics for $Z_1, \ldots, Z_n$, the ordered triples correspond to $(X_{N(1)}, Y_{N(1)}, Z_{N(1)}), \ldots, (X_{N(n)}, Y_{N(n)}, Z_{N(n)})$ (with respect to increasing Z values). Gripenberg's estimator for $\tau_{XY/z}$ is then given by

$$T_{XY/Z} = \frac{1}{n-1} \sum_{i=1}^{n-1} Q((X_{N(i)}, Y_{N(i)}), (X_{N(i+1)}, Y_{N(i+1)})), \tag{8.48}$$

where, once again, $Q((X_{N(i)}, Y_{N(i)}), (X_{N(i+1)}, Y_{N(i+1)})$ is given by (8.5). The approximate $100(1-\alpha)\%$ confidence interval for $\tau_{XY/Z}$ (8.47) proposed by Gripenberg is then

$$\frac{T_{XY/Z}}{\left[1 + \dfrac{bz^2_{\alpha/2}}{n}\right]} \pm \frac{\left\{\dfrac{bz^2_{\alpha/2}}{n}\left(1 - T^2_{XY/Z} + \dfrac{bz^2_{\alpha/2}}{n}\right)\right\}^{1/2}}{\left[1 + \dfrac{bz^2_{\alpha/2}}{n}\right]}, \tag{8.49}$$

where b is an arbitrary consistent estimator of $\beta = \frac{\sigma^{*2}}{1-\tau^2_{XY/Z}}$, with σ^{*2} representing the asymptotic variance of $n^{1/2}T_{XY/Z}$. Two competing estimators b are considered by Gripenberg.

Properties

1. *Asymptotic Distribution-Freeness.* For populations satisfying Assumption A, (8.40) holds. Hence, we can control the coverage probability to be approximately $1-\alpha$ for large sample size n without having more specific knowledge about the form of the underlying bivariate (X, Y) distribution. Thus, (τ_L, τ_U) is an asymptotically distribution-free confidence interval for τ over the class of all continuous bivariate distributions.

Problems

27. For the tapeworm data of Table 8.3, find a confidence interval for τ with the approximate confidence coefficient .95.

28. For the cerebral palsy data of Table 8.4, find a confidence interval for τ with the approximate confidence coefficient .90.

29. Use only six (X, Y) pairs (those corresponding to the first six lot numbers) in Table 8.1 and compute a new estimator $\hat{\sigma}^2$ (8.38) for the asymptotic variance of $\hat{\tau}$ (see Comment 26). Compare it with the estimator based on all nine observations obtained in Example 8.3.

30. For the twins data in Table 8.5, find a lower confidence bound for τ with the approximate confidence coefficient .95. (See Comment 24.)

31. For the educational expense data in Table 8.6, find a lower confidence bound for τ with the approximate confidence coefficient .95. (See Comment 24.)

32. For the nursing data of Table 8.7, find an upper confidence bound for τ with the approximate confidence coefficient .95.

33. Suppose that $(X_1, Y_1, Z_1) = (7.0, 2.5, 1.9), (X_2, Y_2, Z_2) = (6.3, 9.6, 4.1)$, $(X_3, Y_3, Z_3) = (6.9, 3.7, 12.4)$, $(X_4, Y_4, Z_4) = (3.6, 12.1, 6.5)$, $(X_5, Y_5, Z_5) = (9.0, 6.4, 11.2)$, $(X_6, Y_6, Z_6) = (3.0, 6.2, 7.7)$, and $(X_7, Y_7, Z_7) = (4.2, 0.4, 8.2)$ represent a random sample of size $n = 7$ from a trivariate probability distribution. Estimate the partial correlation $\tau_{XY/Z}$ (8.47) between X and Y conditional on Z (See Comment 30).

8.4 AN ASYMPTOTICALLY DISTRIBUTION-FREE CONFIDENCE INTERVAL BASED ON EFRON'S BOOTSTRAP

The asymptotically distribution-free confidence for the parameter τ described in Section 8.3 is based on obtaining a mathematical expression for σ^2, the variance of $\hat{\tau}$. Such an expression depends on the unknown underlying bivariate distribution and so σ^2 must be estimated from the data. The estimate given by (8.38) is consistent, and it is used to form the confidence interval of Section 8.3. In many problems, however, it will be difficult or impossible to obtain a tractable mathematical expression for the variance of the statistic of interest. Efron's bootstrap is a general method for obtaining estimated standard deviations of estimators $\hat{\theta}$ and confidence intervals for parameters θ without requiring a tractable mathematical expression for the asymptotic variance of $\hat{\theta}$. Efron's technique eliminates the mathematical intractability obstacle by relying on computing power and is known as a *computer-intensive method*. It is applicable in a great variety of problems (see Efron (1979), Efron and Gong (1983), Efron and Tibshirani (1993), Davison and Hinkley (1997), DiCiccio and Efron (1996), and Manly (2007)). In this section, we apply Efron's bootstrap method to obtain an asymptotically distribution-free

confidence interval for the parameter τ (the population measure of association defined by (8.2)) using the estimator $\hat{\tau}$ given by (8.34), where Q (8.5) is replaced by Q^* (8.17) in the definition of K.

Procedure

Denote the observed bivariate sample values as

$$Z_1 = (X_1, Y_1), \quad Z_2 = (X_2, Y_2), \ldots, Z_n = (X_n, Y_n).$$

1. Make n random draws with replacement from the bivariate sample $Z_1, Z_2, \ldots, Z_n$. This is equivalent to doing independent random sampling from the bivariate empirical distribution function $\hat{F}$, which puts probability $1/n$ on each of the data points $Z_i, i = 1, \ldots, n$.

   For the canned tuna data of Table 8.1, $n = 9$ and

   $$Z_1 = (44.4, 2.6),\ Z_2 = (45.9, 3.1),\ Z_3 = (41.9, 2.5),\ Z_4 = (53.3, 5.0),$$
   $$Z_5 = (44.7, 3.6),\ Z_6 = (44.1, 4.0),\ Z_7 = (50.7, 5.2),\ Z_8 = (45.2, 2.8),$$
   $$Z_9 = (60.1, 3.8).$$

   A possible bootstrap sample of these data is, for example, 1 copy of Z_1, 2 copies of Z_2, 0 copies of Z_3, 0 copies of Z_4, 1 copy of Z_5, 3 copies of Z_6, 0 copies of Z_7, 1 copy of Z_8, and 1 copy of Z_9.

2. Perform step 1 a large number, say, B, of times. For each draw, compute $\hat{\tau}$. Note that in computing $\hat{\tau}$, it will be necessary to use Q^* (8.17) rather than Q (8.5) in the definition of K. This is because ties will occur in most bootstrap samples because we sample with replacement. Denote the B values of $\hat{\tau}$ as $\hat{\tau}^{*1}, \hat{\tau}^{*2}, \ldots, \hat{\tau}^{*B}$. These are called the *bootstrap replications* of $\hat{\tau}$. Let $\hat{\tau}^{*(1)} \le \hat{\tau}^{*(2)} \le \cdots \le \hat{\tau}^{*(B)}$ denote the ordered values of the bootstrap replications.

   An asymptotically distribution-free confidence interval for τ, with the approximate confidence coefficient $100(1-\alpha)\%$, is $(\tau'_{\mathrm{L}}, \tau'_{\mathrm{U}})$, where

   $$\tau'_{\mathrm{L}} = \hat{\tau}^{*(k)}, \quad \tau'_{\mathrm{U}} = \hat{\tau}^{*(B+1-k)} \tag{8.50}$$

   and

   $$k = B\left(\frac{\alpha}{2}\right). \tag{8.51}$$

   If $k = B(\alpha/2)$ is an integer, then τ'_{L} is the kth-largest bootstrap replication and τ'_{U} is the $(B+1-k)$th-largest replication. For example, if $\alpha = .10$ and $B = 1,000$, $k = 1,000(.05) = 50$, τ'_{L} is the bootstrap replication occupying position 50 in the ordered list, and τ'_{U} is the bootstrap replication occupying position 951 in the ordered list. If $B(\alpha/2)$ is not an integer, we follow the convention of Effon and Tibshirani (1993, p. 160) and set $k = \langle (B+1)(\alpha/2) \rangle$, the largest integer that is less than or equal to $(B+1)(\alpha/2)$. With this value of k, τ'_{L} is the bootstrap replication occupying position k in the ordered list and τ'_{U} is the bootstrap replication occupying position $B+1-k$ in the ordered list.

EXAMPLE 8.4 ***(Continuation of Examples 8.1, 8.2, and 8.3).***

We obtained 1000 bootstrap replications of $\hat{\tau}$. Figure 8.1 is a histogram of the 1000 bootstrap replications. The 1000 bootstrap replications are as follows:

$$\hat{\tau}^{*(1)} = -.556, \hat{\tau}^{*(2)} = -.500, \ \hat{\tau}^{*(3)} = \cdots = \hat{\tau}^{*(5)} = -.444,$$

$$\hat{\tau}^{*(6)} = -.417, \ \hat{\tau}^{*(7)} = \cdots \hat{\tau}^{*(9)} = -.361,$$

$$\hat{\tau}^{*(10)} = \hat{\tau}^{*(11)} = -.306, \hat{\tau}^{*(12)} = -.250,$$

$$\hat{\tau}^{*(13)} = \hat{\tau}^{*(14)} = -.222, \ \hat{\tau}^{*(15)} = \hat{\tau}^{*(16)} = -.194,$$

$$\hat{\tau}^{*(17)} = \hat{\tau}^{*(18)} = -.167, \ \hat{\tau}^{*(19)} = -.139,$$

$$\hat{\tau}^{*(20)} = \hat{\tau}^{*(21)} = -.111, \hat{\tau}^{*(22)} = \cdots = \hat{\tau}^{*(32)} = -.083,$$

$$\hat{\tau}^{*(33)} = \cdots = \hat{\tau}^{*(40)} = -.056, \ \hat{\tau}^{*(41)} = \cdots = \hat{\tau}^{*(50)} = -.028,$$

$$\hat{\tau}^{*(51)} = \cdots = \hat{\tau}^{*(70)} = .000, \ \hat{\tau}^{*(71)} = \cdots = \hat{\tau}^{*(85)} = .028,$$

$$\hat{\tau}^{*(86)} = \cdots = \hat{\tau}^{*(91)} = .056, \ \hat{\tau}^{*(92)} = \cdots = \hat{\tau}^{*(113)} = .083,$$

$$\hat{\tau}^{*(114)} = \cdots = \hat{\tau}^{*(133)} = -.111, \ \hat{\tau}^{*(134)} = \cdots = \hat{\tau}^{*(148)} = .139,$$

$$\hat{\tau}^{*(149)} = \cdots = \hat{\tau}^{*(173)} = .167, \ \hat{\tau}^{*(174)} = \cdots = \hat{\tau}^{*(194)} = .194,$$

$$\hat{\tau}^{*(195)} = \cdots = \hat{\tau}^{*(227)} = .222, \ \hat{\tau}^{*(228)} = \cdots = \hat{\tau}^{*(264)} = .250,$$

$$\hat{\tau}^{*(265)} = \cdots = \hat{\tau}^{*(316)} = .278, \ \hat{\tau}^{*(317)} = \cdots = \hat{\tau}^{*(351)} = .306,$$

$$\hat{\tau}^{*(352)} = \cdots = \hat{\tau}^{*(393)} = .333, \ \hat{\tau}^{*(394)} = \cdots = \hat{\tau}^{*(433)} = .361,$$

$$\hat{\tau}^{*(434)} = \cdots = \hat{\tau}^{*(478)} = .389, \ \hat{\tau}^{*(479)} = \cdots = \hat{\tau}^{*(530)} = .417,$$

$$\hat{\tau}^{*(531)} = \cdots = \hat{\tau}^{*(592)} = .444, \ \hat{\tau}^{*(593)} = \cdots = \hat{\tau}^{*(640)} = .472,$$

$$\hat{\tau}^{*(641)} = \cdots = \hat{\tau}^{*(687)} = .500, \ \hat{\tau}^{*(688)} = \cdots = \hat{\tau}^{*(730)} = .528,$$

$$\hat{\tau}^{*(731)} = \cdots = \hat{\tau}^{*(769)} = .556, \ \hat{\tau}^{*(770)} = \cdots = \hat{\tau}^{*(808)} = .583,$$

$$\hat{\tau}^{*(809)} = \cdots = \hat{\tau}^{*(847)} = .611, \ \hat{\tau}^{*(848)} = \cdots = \hat{\tau}^{*(880)} = .639,$$

$$\hat{\tau}^{*(881)} = \cdots = \hat{\tau}^{*(912)} = .667, \ \hat{\tau}^{*(913)} = \cdots = \hat{\tau}^{*(935)} = .694,$$

$$\hat{\tau}^{*(936)} = \cdots = \hat{\tau}^{*(955)} = .722, \ \hat{\tau}^{*(956)} = \cdots = \hat{\tau}^{*(969)} = .750,$$

$$\hat{\tau}^{*(970)} = \cdots = \hat{\tau}^{*(986)} = .778, \ \hat{\tau}^{*(987)} = \cdots = \hat{\tau}^{*(992)} = .806,$$

$$\hat{\tau}^{*(993)} = \hat{\tau}^{*(994)} = .833, \ \hat{\tau}^{*(995)} = \cdots = \hat{\tau}^{*(999)} = .861, \hat{\tau}^{*(1000)} = .889.$$

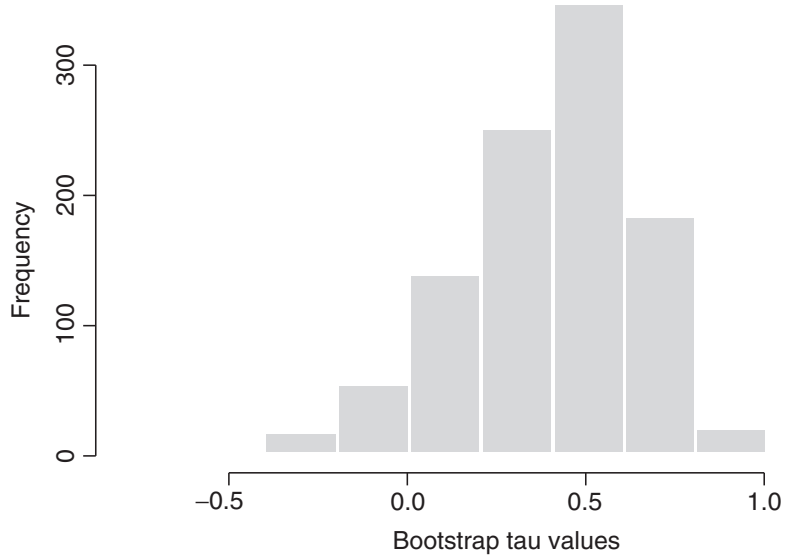

Figure 8.1 Histogram of 1000 bootstrap replications of Kendall's sample correlation coefficient for the canned tuna data of Table 8.1.

For an approximate 90% confidence interval $\alpha = .1, (\alpha/2) = 0.5$, and from (8.51), we find $k = 1000(.05) = 50$. Then from (8.50) and the ordered list of 1000 bootstrap replications,

$$\tau'_{\mathrm{L}} = \hat{\tau}^{*(50)} = -.028, \quad \tau'_{\mathrm{U}} = \hat{\tau}^{*(951)} = .722.$$

The command `kendall.ci` will provide a bootstrap confidence interval. In addition to the arguments specified in Example 8.3, one here must set `bootstrap=T` and a value for the number of replicates B. For example,

```
kendall.ci(x, y, alpha=.1, type="t", bootstrap=T, B=1000)
```

will find an interval similar, but almost certainly not identical, to the interval above or the interval found in Example 8.3. Running the above command five times results in the following intervals: $(-.063, .818)$, $(-.067, .857)$, $(-.063, .824)$. $(-.030, .824)$, and $(-.030, .871)$. Recall that the method from Section 8.3 gave the interval (.027, .861).

Comments

31. *The Bootstrap Estimated Standard Error.* For Kendall's sample correlation coefficient $\hat{\tau}$, the standard deviation of $\hat{\tau}$, which we have denoted thus far as σ, depends on the bivariate distribution function $F_{X,Y}$. We now denote $F_{X,Y}$ as F, dropping the subscripts. We could exhibit the dependence of σ on F by writing σ as $\sigma(F)$. Although F is unknown, it can be estimated by the bivariate empirical distribution function $\hat{F}$, which puts probability $1/n$ on each of the observed data points $Z_i = (X_i, Y_i), i = 1, \ldots, n$. The bootstrap estimate of $\sigma(F)$ is $\sigma(\hat{F})$, where $\sigma(\hat{F})$ is the standard deviation of $\hat{\tau}$ when the true underlying distribution is $\hat{F}$ rather than F. A tractable mathematical expression for $\sigma(\hat{F})$ is very difficult to obtain. However, $\sigma(\hat{F})$ can be estimated using the B bootstrap replications, by

$$\hat{\sigma}_B = \left\{ \frac{\sum_{i=1}^{B} (\hat{\tau}^{*i} - \hat{\tau}^{*\cdot})^2}{(B-1)} \right\}^{1/2}, \tag{8.52}$$

where

$$\hat{\tau}^{*\cdot} = \frac{\sum_{i=1}^{B} \hat{\tau}^{*i}}{B}. \tag{8.53}$$

As B tends to ∞, $\hat{\sigma}_B$ tends to $\sigma(\hat{F})$. As n tends to ∞, $\sigma(\hat{F})$ tends to $\sigma(F)$. Thus, $\hat{\sigma}_B$ can be used as an estimate of the standard deviation of $\hat{\tau}$.

32. *The Bootstrap in the One-Sample Nonparametric Framework.* In this section, we applied the bootstrap in a bivariate situation, where the data are bivariate observations $Z_i = (X_i, Y_i), i = 1, \ldots, n$, and the parameter of interest is τ. The bootstrap can be used in a wide variety of situations, including the one-sample problem, the k-sample problem, censored data problems, and complicated multivariate frameworks. (See Efron and Gong (1983), Efron and Tibshirani (1993), Davison and Hinkley (1997), and DiCiccio and Efron (1996).) In this comment, we describe the approach in the context of the one-sample nonparametric framework.

Suppose we are interested in estimating a parameter $\theta = \theta(F)$, when $X_1, \ldots, X_n$ are a random sample from an unknown distribution F. The nonparametric maximum likelihood estimate of F is the statistic $\hat{\theta} = \theta(F_n)$, where F_n is the sample distribution function. For example, if we are interested in estimating the rth moment of the F distribution, $\theta(F) = E(X^r)$, then $\theta(F_n) = (\sum_{i=1}^{n} X_i^r)/n$.

The bootstrap procedure in the one-sample problem is analogous to the procedure we described for the bivariate situation. The steps are as follows:

1. Make n random draws with replacement from the sample $X_1, \ldots, X_n$.
2. Perform step 1 a large number, say B, of times. For each draw, compute $\hat{\theta}$. Denote the B values of $\hat{\theta}$ as $\hat{\theta}^{*1}, \hat{\theta}^{*2}, \ldots, \hat{\theta}^{*B}$. These are the bootstrap replications of $\hat{\theta}$.

The bootstrap estimate of the standard deviation of $\hat{\theta}$ is

$$\hat{\sigma}_B = \left\{ \frac{\sum_{i=1}^{B} (\hat{\theta}^{*i} - \hat{\theta}^{*\cdot})^2}{(B-1)} \right\}^{1/2}, \tag{8.54}$$

where

$$\hat{\theta}^{*\cdot} = \frac{\sum_{i=1}^{B} \hat{\theta}^{*i}}{B}. \tag{8.55}$$

An asymptotically distribution-free confidence interval for θ, with approximate confidence coefficient $100(1-\alpha)\%$, is $(\theta'_{\mathrm{U}}, \theta'_{\mathrm{L}})$, where

$$\theta'_{\mathrm{L}} = \hat{\theta}^{*(k)}, \quad \theta'_{\mathrm{U}} = \hat{\theta}^{*(B+1-k)}, \tag{8.56}$$

where $\hat{\theta}^{*(1)} \le \hat{\theta}^{*(2)} \le \cdots \le \hat{\theta}^{*(B)}$ are the ordered values for the bootstrap replications and $k = \langle (B+1)(\alpha/2) \rangle$, the largest integer that is less than or equal to $(B+1)(\alpha/2)$.

The confidence interval defined by (8.56) is called the *percentile interval*. Let $\hat{G}$ denote the cumulative distribution function of $\hat{\theta}^*$:

$$\hat{G}(t) = \frac{\#\{\hat{\theta}^{*i} < t\}}{B}. \tag{8.57}$$

The end points $\theta'_{\mathrm{L}}, \theta'_{\mathrm{U}}$ are, respectively, the α and $1-\alpha$ percentiles of $\hat{G}$.

The percentile confidence interval is *transformation-respecting*. If $\eta = m(\theta)$ is a monotone transformation, then a confidence interval (η_L, η_U) for the parameter η is obtained directly from the confidence interval (8.56) for θ via $\eta_L = m(\theta'_L), \eta_U = m(\theta'_U)$. For example, a confidence interval for θ^2 is obtained by squaring the end points of the confidence interval for θ.

The percentile confidence interval is *range-preserving*. For example, consider the percentile interval based on bootstrapping $\hat{\tau}$. The values of the parameter τ, Kendall's population correlation coefficient, are always between -1 and 1. The possible values of the estimator $\hat{\tau}$ also are in the interval $[-1, 1]$. Thus, the bootstrap replications of $\hat{\tau}$ must be in the interval $[-1, 1]$, as must the confidence interval end points, because the end points are particular bootstrap replications. More generally, the estimator $\hat{\theta}$ of the form $\hat{\theta} = \theta(F_n)$ satisfies the same range restrictions as $\theta = \theta(F)$, and thus, the percentile interval based on bootstrapping $\hat{\theta}$ also satisfies the same range restrictions as θ.

33. *The BC_a Confidence Interval.* Efron and Tibshirani (1993, Chapter 14) (see also DiCiccio and Efron, 1996) present a method, called the BC_a method, that gives more accurate confidence limits than does the percentile method of Comment 32. The acronym BC_a means "bias-corrected and accelerated." The BC_a method depends on a bias-correction z_0 and an acceleration a. In the one-sample nonparametric framework, z_0 can be estimated by

$$\hat{z}_0 = \Phi^{-1}\left\{\frac{\#\{\hat{\theta}^{*i} < \hat{\theta}\}}{B}\right\}, \tag{8.58}$$

where Φ denotes the standard normal cumulative distribution function. Thus, $\hat{z}_0$ is Φ^{-1} of the proportion of the bootstrap replications less than $\hat{\theta}$.

The estimate of a is

$$\hat{a} = \frac{\sum_{i=1}^{n}(\hat{\theta}_\cdot - \hat{\theta}_{-i})^3}{6\left\{\sum_{i=1}^{n}(\hat{\theta}_\cdot - \hat{\theta}_{-i})^2\right\}^{3/2}}, \tag{8.59}$$

where $\hat{\theta}_{-i}$ is the estimate of θ obtained by deleting X_i, and computing $\hat{\theta}$ for the reduced sample $X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$ and

$$\hat{\theta}_\cdot = \frac{\sum_{i=1}^{n}\hat{\theta}_{-i}}{n}. \tag{8.60}$$

The lower and upper end points of the $100(1-\alpha)\%$ confidence interval are

$$\theta''_L = \hat{G}^{-1}\Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha/2)}}{1 - \hat{a}(\hat{z}_0 + z^{(\alpha)})}\right), \tag{8.61}$$

$$\theta''_U = \hat{G}^{-1}\Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha/2)}}{1 - \hat{a}(\hat{z}_0 + z^{(1-\alpha/2)})}\right), \tag{8.62}$$

where in (8.61) and (8.62), Φ is the standard normal distribution function, $z^{(\alpha/2)} = \Phi^{-1}(\alpha/2), z^{(1-\alpha/2)} = \Phi^{-1}(1-\alpha/2)$ (if, for example, $a = .10$, then

$z^{(\alpha/2)} = -1.65$, and $z^{(1-\alpha/2)} = 1.65$), $\hat{G}$ is given by (8.57), $\hat{z}_0$ is given by (8.58), and $\hat{a}$ is given by (8.59).

The end points θ''_{L} and θ''_{U} of the BC_a interval are also percentiles of the bootstrap distribution $\hat{G}$ but not necessarily the same ones as given by the percentile interval. If $\hat{a} = \hat{z}_0 = 0$, the BC_a and percentile intervals are the same.

The BC_a interval also enjoys the transformation-respecting and range-preserving properties that hold for the percentile interval. The BC_a interval, however, has an accuracy advantage. The BC_a interval has a second-order accuracy property, whereas the percentile interval is only first-order accurate. See Section 14.3 of Efron and Tibshirani (1993).

The appendix of Efron and Tibshirani (1993) describes some available bootstrap software and contains some programs in the S language, including a program for computing $\mathrm{BC_a}$ intervals.

34. *The Choice of B, the Number of Bootstrap Replications.* The choice of B depends to some extent on the particular statistic that is being bootstrapped and the complexity of the situation. Efron and Tibshirani (1993, p. 52) give some rules of thumb based on their extensive experience with the bootstrap. Roughly speaking, $B = 200$ replications are usually sufficient for estimating a standard error but much larger values of B, such as 1000 or 2000, are required for bootstrap confidence intervals.

35. *An Example Where the Bootstrap Fails.* Let $X_1, \ldots, X_n$ be a random sample from the uniform distribution on $(0, \theta)$. The maximum likelihood estimator of θ, the upper end point of the interval, is $\hat{\theta} = \text{maximum}(X_1, \ldots, X_n) = X_{(n)}$. Efron and Tibshirani (1993, p. 81) point out that the bootstrap does not do well in this situation. Miller (1964) showed that the jackknife estimator of θ also fails in this situation, because it depends not only on $X_{(n)}$ but also on $X_{(n-1)}$, the second largest observation, and the latter contains no additional information about θ when the value of $X_{(n)}$ is available.

36. *Jackknife versus Bootstrap.* For a linear statistic of the form $\hat{\theta} = \mu + \left\{\sum_{i=1}^{n} h(X_i)/n\right\}$, where μ is a constant and h is a function, there is no loss of information in using the jackknife rather than the bootstrap. For nonlinear statistics, there is a loss of information and the bootstrap should be preferred. See Efron and Tibshirani (1993) for a detailed discussion of the relationship between the jackknife and the bootstrap.

    One disadvantage of the bootstrap is that two different people bootstrapping the same data will not in general get the same bootstrap estimate of the standard deviation or the same confidence interval. This violates what Gleser (1996) calls "the first law of applied statistics," namely: "Two individuals using the same statistical method on the same data should arrive at the same conclusion." See Gleser (1996) for other disadvantages of the bootstrap.

37. *Development of the Bootstrap.* The bootstrap was formally introduced by Efron (1979). Efron and Tibshirani (1993, p. 56), however, credit many authors for similar ideas, and they designate as "particularly notable" the contributions of Hartigan's typical value theory (1969, 1971, 1975). Hartigan recognized the wide applicability of subsample methods (a subsample of $X_1, \ldots, X_n$ is any subset of the whole sample) as a tool for assessing variability. See Efron and Tibshirani (1993, p. 56) for references to other papers that contain ideas related to bootstrapping.

Properties

See Bickel and Freedman (1981) for asymptotic consistency. See Efron and Tibshirani (1993) and Davison and Hinkley (1997) for various properties including accuracy, transformation-respecting, range-preserving, and the relationship of the bootstrap to the jackknife. See Hall (1992) for a high-level mathematical treatment of the bootstrap. See Manly (2007) for bootstrap methods in biology.

Problems

34. For the cerebral palsy data of Table 8.4, use the bootstrap method to find a confidence interval for τ with approximate confidence coefficient .90. Compare your results with those of Problem 28.

35. For the psychological test scores data of Table 8.5, use the bootstrap method to find a confidence interval for τ with approximate confidence coefficient .95.

36. Consider the case $n = 3$ where you have three bivariate observations Z_1, Z_2, and Z_3. List the possible bootstrap samples and give the corresponding probability of each being selected on a given bootstrap replication.

37. Consider the case where you have four bivariate observations Z_1, Z_2, Z_3, and Z_4. List the possible bootstrap samples and give the corresponding probability of each being selected on a given bootstrap replication.

38. Illustrate by means of an example or show directly that with n observations the number of possible bootstrap samples is $\binom{2n-1}{n}$.

39. Show that if $\hat{a} = \hat{z}_0 = 0$, the BC_a interval given by (8.61) and (8.62) reduces to the percentile interval.

8.5 A DISTRIBUTION-FREE TEST FOR INDEPENDENCE BASED ON RANKS (SPEARMAN)

Hypothesis

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from a continuous bivariate population (i.e., Assumption A is satisfied) with joint distribution function $F_{X,Y}$ and marginal distribution functions F_X and F_Y. In this section, we return to the problem of testing for independence between the X and Y variables corresponding to the null hypothesis H_0 (8.1). Here, however, alternatives to H_0 will no longer be stated in terms of the correlation coefficient τ (8.2). Instead, the alternatives of interest in this section are less specifically interpretable, corresponding to the quite general (but vague; see Comment 47) concepts of positive or negative association between the X and Y variables.

Procedure

To compute the Spearman rank correlation coefficient r_s, we first order the n X observations from least to greatest and let R_i denote the rank of $X_i, i = 1, \ldots, n$, in this ordering. Similarly, we separately order the n Y observations from least to greatest and let S_i denote the rank of $Y_i, i = 1, \ldots, n$, in this ordering. The Spearman (1904) rank correlation coefficient is defined as the Pearson product moment sample correlation of

the R_i and the S_i. (See Comment 40.) When no ties within a sample are present, this is equivalent to two computationally efficient formulae:

$$r_s = \frac{12 \sum_{i=1}^{n} \left\{ \left[R_i - \frac{n+1}{2} \right] \left[S_i - \frac{n+1}{2} \right] \right\}}{n(n^2 - 1)} \tag{8.63}$$

$$= 1 - \frac{6 \sum_{i=1}^{n} D_i^2}{n(n^2 - 1)}, \tag{8.64}$$

where $D_i = S_i - R_i, i = 1, \ldots, n$.

a. *One-Sided Upper-Tail Test.* To test the null hypothesis of independence, H_0 (8.1), versus the directional alternative

$$H_1 : [X \text{ and } Y \text{ are positively associated}] \tag{8.65}$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } r_s \geq r_{s,\alpha}; \quad \text{otherwise do not reject,} \tag{8.66}$$

where the constant $r_{s,\alpha}$ is chosen to make the type I error probability equal to α. Values of $r_{s,\alpha}$ are found with the command `qSpearman`.

b. *One-Sided Lower-Tail Test.* To test independence, H_0 (8.1), versus the directional alternative

$$H_2 : [X \text{ and } Y \text{ are negatively associated}] \tag{8.67}$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } r_s \leq -r_{s,\alpha}; \quad \text{otherwise do not reject.} \tag{8.68}$$

c. *Two-Sided Test.* To test independence, H_0 (8.1), versus the general dependency alternative

$$H_3 : [X \text{ and } Y \text{ are not independent variables}] \tag{8.69}$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } |r_s| \geq r_{s,\alpha/2}; \quad \text{otherwise do not reject.} \tag{8.70}$$

This two-sided procedure is the two-sided symmetric test with $\alpha/2$ probability in each tail of the null distribution of r_s.

Large-Sample Approximation

The large-sample approximation is based on the asymptotic normality of r_s, suitably standardized. For this standardization, we need to know the expected value and variance of r_s when the null hypothesis of independence is true. Under H_0, the expected value and variance of r_s are

$$E_0(r_s) = 0 \tag{8.71}$$

and

$$\text{var}_0(r_s) = \frac{1}{n-1}, \tag{8.72}$$

respectively. These expressions for $E_0(r_s)$ and $\text{var}_0(r_s)$ are verified by direct calculations in Comment 42 for the special case of $n = 4$. General derivations of both expressions are discussed in Comment 45.

The standardized version of r_s is

$$r_s^* = \frac{r_s - E_0(r_s)}{\{\text{var}_0(r_s)\}^{1/2}} = (n-1)^{1/2} r_s. \tag{8.73}$$

When H_0 is true, r_s^* has, as n tends to infinity, an asymptotic $N(0,1)$ distribution. (See Comment 45 for indications of the proof.) The normal theory approximation for procedure (8.66) is

$$\text{Reject } H_0 \text{ if } r_s^* \geq z_\alpha; \quad \text{otherwise do not reject,} \tag{8.74}$$

the normal theory approximation for procedure (8.68) is

$$\text{Reject } H_0 \text{ if } r_s^* \leq -z_\alpha; \quad \text{otherwise do not reject,} \tag{8.75}$$

and the normal theory approximation for procedure (8.70) is

$$\text{Reject } H_0 \text{ if } |r_s^*| \geq z_{\alpha/2}; \quad \text{otherwise do not reject.} \tag{8.76}$$

Ties

If there are ties among the n X observations and/or separately among the n Y observations, assign each of the observations in a tied (either X or Y) group the average of the integer ranks that are associated with the tied group. After computing r_s with these average ranks for tied observations, use procedure (8.66), (8.68), or (8.70). Note, however, that this test associated with tied X's and/or tied Y's is only approximately, and not exactly, of significance level α. (To get an exact level α test even in this tied setting, see Comment 46.)

If there are tied X's and/or tied Y's, Spearman's rank correlation coefficient calculated with Pearson's correlation does not require modification. If using the computationally efficient version of r_s at (8.64), some changes to the statistic are necessary. The statistic r_s in this case becomes

$$r_s = \frac{n(n^2-1) - 6\sum_{s=1}^{n} D_s^2 - \frac{1}{2}\left\{\sum_{i=1}^{g}\left[t_i\left(t_i^2-1\right)\right] + \sum_{j=1}^{h}\left[u_j\left(u_j^2-1\right)\right]\right\}}{\left\{\left[n(n^2-1) - \sum_{i=1}^{g} t_i\left(t_i^2-1\right)\right]\left[n(n^2-1) - \sum_{j=1}^{h} u_j\left(u_j^2-1\right)\right]\right\}^{1/2}}, \tag{8.77}$$

where in (8.77) g denotes the number of tied X groups, t_i is the size of tied X group i, h is the number of tied Y groups, and u_j is the size of tied Y group j. We note that an untied $X(Y)$ observation is considered to be a tied $X(Y)$ group of size 1. In particular, if neither the collection of X nor the collection of Y observations contains tied values, we have $g = h = n, t_j = u_j = 1, i = 1, \ldots, n$, and $j = 1, \ldots, n$. In this case of no tied X's

and no tied Y's, each term involving either $(t_i^2 - 1)$ or $(u_j^2 - 1)$ reduces to zero and the "ties" expression for r_s in (8.77) reduces to the "no-ties" form for r_s, as given in (8.64).

As a consequence of this effect that ties have on the null distribution of r_s, in order to use the large-sample approximation when there are tied X observations and/or tied Y observations, we first compute r_s^* (8.73) using average ranks and the ties-corrected version of r_s (8.77). Approximation (8.74), (8.75), or (8.76) can then be applied, as appropriate for the problem, with this value of r_s^*.

EXAMPLE 8.5 ***Proline and Collagen in Liver Cirrhosis.***

Kershenobich, Fierro, and Rojkind (1970) have studied the relation between the free pool of proline and collagen content in human liver cirrhosis. The data in Table 8.9 are based on an analysis of cirrhotic livers from seven patients, each having a histological diagnosis of portal cirrhosis.

We are interested in assessing whether there is a positive relationship between the total collagen and the free proline in cirrhotic livers. Thus, we wish to apply procedure (8.66) to test the hypothesis of independence, H_0 (8.1), versus the alternative, H_1 (8.65), of positive association. For purposes of illustration, we consider the significance level $\alpha = .01$. The statistic r_s is symmetric about 0 (see Comment 43), so $P(r_s \geq r_{s,.01}) = P(r_s \leq -r_{s,.01})$. In R, this is found with

```
qSpearman(.01, r=7),
```

where `r` is the number of samples. The result is $-.786$, so procedure (8.66) becomes

$$\text{Reject } H_0 \text{ if } r_s \geq .786.$$

Ranking the X (total collagen) values from least to greatest, using average ranks for the tied pair, we obtain $R_1 = (1+2)/2 = 1.5, R_2 = (1+2)/2 = 1.5, R_3 = 3, R_4 = 4, R_5 = 5, R_6 = 6$, and $R_7 = 7$. Similarly, ranking the Y (free proline) values from least to greatest, again using average ranks for the tied pairs, we find $S_1 = (2+3)/2 = 2.5, S_2 = 4, S_3 = (2+3)/2 = 2.5, S_4 = 1, S_5 = 5, S_6 = 6$, and $S_7 = 7$. Taking differences, we see that

$$D_1 = 2.5 - 1.5 = 1, \quad D_2 = 4 - 1.5 = 2.5, \quad D_3 = 2.5 - 3 = -.5$$
$$D_4 = 1 - 4 = -3, \quad D_5 = 5 - 5 = 0, \quad D_6 = 6 - 6 = 0$$
$$D_7 = 7 - 7 = 0.$$

Table 8.9 Free Proline and Total Collagen Contents of Cirrhotic Patients

Patient	Total collagen, X_i (mg/g dry weight of liver)	Free proline, Y_i, (μ mole/g dry weight of liver)
1	7.1	2.8
2	7.1	2.9
3	7.2	2.8
4	8.3	2.6
5	9.4	3.5
6	10.5	4.6
7	11.4	5.0

Source: D. Kershenobich, F. J. Fierro, and M. Rojkind (1970).

There are ties, so we need to use the ties-corrected version of the statistic r_s if using (8.64). If using Pearson's correlation of ranks, no modifications are necessary.

For this purpose, we note that there are $g = 6$ tied X groups, with $t_1 = 2, t_2 = t_3 = t_4 = t_5 = t_6 = 1$, and $h = 6$ tied Y groups, with $u_2 = 2, u_1 = u_3 = u_4 = u_5 = u_6 = 1$. Thus, for these tied data, the modified value of r_s (8.77) is calculated to be

$$r_s = \frac{7(7^2-1) - 6[(1)^2 + (2.5)^2 + (-.5)^2 + (-3)^2 + 3(0)^2] - \frac{1}{2}(2)(2)(2^2-1)}{\{[7(7^2-1) - 2(2^2-1)][7(7^2-1) - 2(2^2-1)]\}^{1/2}}$$

$$= \frac{7(48) - 6(16.5) - 6}{\{[7(48) - 6][7(48) - 6]\}^{1/2}} = \frac{231}{330} = .700.$$

This value of r_s is also obtained through the R command

```
cor(x, y, method="spearman")
```

where x and y are the data from Table 8.9. This value of r_s is not greater than the critical value .786, so we do not reject the null at the $\alpha = .01$ level. Note that the critical value given by R results in a significance level of $\alpha = .024$, not $\alpha = .01$.

The one-sided P-value for these data is the smallest significance level at which we can reject H_0 in favor of a positive association between total collagen and free proline in cirrhotic patients with the observed value of the test statistic $r_s = .700$. We see that the P-value is $P_0(r_s \geq .700)$. By symmetry, this is the same as $P_0(r_s \leq -.700)$. The R command `pSpearman` will provide the following:

```
pSpearman(-.700, r=7)=.044.
```

Thus, there is some marginal evidence that total collagen and free proline are positively associated in subjects with liver cirrhosis.

For the large-sample approximation, we use $r_s = .700$ to compute r_s^* (8.73) and obtain

$$r_s^* = (6)^{1/2}(.700) = 1.71.$$

Thus, the smallest significance level at which we can reject H_0 in favor of positive association between total collagen and free proline in subjects with liver cirrhosis using the normal theory approximation is .0436 ($z_{.0436} = 1.71$).

The R function `cor.test` reproduces the above analysis.

```
cor.test (x, y, method="spearman", alternative="greater")
```

produces this output:

```
 Spearman's rank correlation rho

data: x and y
S = 16.8, p-value = 0.03996
alternative hypothesis: true rho is greater than 0
sample estimates:
rho
0.7

  Warning message:
In cor.test.default(x8.9, y8.9, method = "s", alt = "g") :
Cannot compute exact p-values with ties
```

The statistic reported as S is the sum of D_i^2 from (8.64). However, in the presence of ties, this value is not accurate. Rather, it is found using Pearson's correlation of the ranks for r_s and then solving for D_i^2 in (8.64). R provides a warning about inexact P-values in the presence of ties. In this case, one should use `pSpearman` to obtain a P-value.

Comments

38. *Motivation for the Test.* The null hypothesis of this section is that the X and Y variables are independent, which, in the case of no ties, implies that any permutation of the X ranks $(R_1, \ldots, R_n)$ is equally likely to occur with any permutation of the Y ranks $(S_1, \ldots, S_n)$. As a result, under the null hypothesis H_0 (8.1) of independence, the Spearman rank correlation coefficient r_s (8.64) will have a tendency to assume values near zero. However, when the alternative H_1 : [X and Y are positively associated] is true, the rank vectors $(R_1, \ldots, R_n)$ and $(S_1, \ldots, S_n)$ will tend to agree, resulting in small differences $D_i = S_i - R_i, i = 1, \ldots, n$. Thus, when H_1 (8.65) is true, we would expect the value of $\sum_{j=1}^n D_j^2$ to be small and the resulting value of r_s (8.64) to be large and positive. This suggests rejecting H_0 in favor of positive association H_1 (8.65) for large positive values of r_s and motivates procedures (8.66) and (8.74). Similar rationales apply to procedures (8.68), (8.70), (8.75), and (8.76).

39. *Computation of* r_s. The value of r_s (8.63) can also be obtained in R using

```
cor(rank(x), rank(y), method=''pearson'').
```

The command rank provides the ranks of a sample. The default method of dealing with ties in this command is to average the ranks within a tie group.

40. *Pearson's Product Moment Sample Correlation Coefficient.* The classical Pearson product moment sample correlation coefficient for the pair of vectors $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ is given by

$$r_p = \frac{\sum_{k=1}^n (X_k - \overline{X})(Y_k - \overline{Y})}{\left[\sum_{i=1}^n (X_i - \overline{X})^2 \ \sum_{j=1}^n (Y_j - \overline{Y})^2\right]^{1/2}}, \tag{8.78}$$

where $\overline{X} = \sum_{s=1}^n X_s/n$ and $\overline{Y} = \sum_{t=1}^n Y_t/n$. We note that the Spearman rank correlation coefficient r_s is simply the classical correlation coefficient applied to the rank vectors $(R_1, \ldots, R_n)$ and $(S_1, \ldots, S_n)$ instead of the actual X and Y observations, respectively. (See Problem 49.)

41. *Derivation of the Distribution of* r_s *under* H_0 *(No-Ties Case).* Without loss of generality, we take $R_1 = 1, \ldots, R_n = n$; under H_0 (8.61) all possible $n!(S_1, S_2, \ldots, S_n)$ Y-rank configurations are equally likely, therefore each has null probability $1/n!$.

Let us consider the case $n = 4$. In the following table, we display the $4! = 24$ possible (S_1, S_2, S_3, S_4) configurations, the associated values of r_s, and the corresponding null probabilities.

(R_1, R_2, R_3, R_4)	(S_1, S_2, S_3, S_4)	Null probability	r_s
(1, 2, 3, 4)	(1, 2, 3, 4)	$\frac{1}{24}$	1
(1, 2, 3, 4)	(1, 2, 4, 3)	$\frac{1}{24}$	.8
(1, 2, 3, 4)	(1, 3, 2, 4)	$\frac{1}{24}$	.8
(1, 2, 3, 4)	(1, 3, 4, 2)	$\frac{1}{24}$	.4
(1, 2, 3, 4)	(1, 4, 2, 3)	$\frac{1}{24}$	.4
(1, 2, 3, 4)	(1, 4, 3, 2)	$\frac{1}{24}$	.2
(1, 2, 3, 4)	(2, 1, 3, 4)	$\frac{1}{24}$	.8
(1, 2, 3, 4)	(2, 1, 4, 3)	$\frac{1}{24}$	.6
(1, 2, 3, 4)	(2, 3, 1, 4)	$\frac{1}{24}$	.4
(1, 2, 3, 4)	(2, 3, 4, 1)	$\frac{1}{24}$	−.2
(1, 2, 3, 4)	(2, 4, 1, 3)	$\frac{1}{24}$	0
(1, 2, 3, 4)	(2, 4, 3, 1)	$\frac{1}{24}$	−.4
(1, 2, 3, 4)	(3, 1, 2, 4)	$\frac{1}{24}$	.4
(1, 2, 3, 4)	(3, 1, 4, 2)	$\frac{1}{24}$	0
(1, 2, 3, 4)	(3, 2, 1, 4)	$\frac{1}{24}$	.2
(1, 2, 3, 4)	(3, 2, 4, 1)	$\frac{1}{24}$	−.4
(1, 2, 3, 4)	(3, 4, 1, 2)	$\frac{1}{24}$	−.6
(1, 2, 3, 4)	(3, 4, 2, 1)	$\frac{1}{24}$	−.8
(1, 2, 3, 4)	(4, 1, 2, 3)	$\frac{1}{24}$	−.2
(1, 2, 3, 4)	(4, 1, 3, 2)	$\frac{1}{24}$	−.4
(1, 2, 3, 4)	(4, 2, 1, 3)	$\frac{1}{24}$	−.4
(1, 2, 3, 4)	(4, 2, 3, 1)	$\frac{1}{24}$	−.8
(1, 2, 3, 4)	(4, 3, 1, 2)	$\frac{1}{24}$	−.8
(1, 2, 3, 4)	(4, 3, 2, 1)	$\frac{1}{24}$	−1

Thus, for example, the probability is $\frac{3}{24}$ under H_0 that r_s is equal to .8, because $r_s = .8$ when any of the three outcomes $(S_1, S_2, S_3, S_4) = (1, 2, 4, 3)$, (1, 3, 2, 4), or (2, 1, 3, 4) occurs and each of these outcomes has null probability $\frac{1}{24}$. Simplifying, we obtain the null distribution

Possible value of r_s	Probability under H_0
-1.0	$\frac{1}{24}$
-0.8	$\frac{3}{24}$
-0.6	$\frac{1}{24}$
-0.4	$\frac{4}{24}$
-0.2	$\frac{2}{24}$
0.0	$\frac{2}{24}$
0.2	$\frac{2}{24}$
0.4	$\frac{4}{24}$
0.6	$\frac{1}{24}$
0.8	$\frac{3}{24}$
1.0	$\frac{1}{24}$

The probability, under H_0, that r_s is greater than or equal to .6, for example, is therefore

$$P_0(r_s \geq .6) = P_0(r_s = 1.0) + P_0(r_s = .8) + P_0(r_s = .6)$$
$$= \frac{1}{24} + \frac{3}{24} + \frac{1}{24} = \frac{5}{24}.$$

Note that we have obtained the null distribution of r_s without specifying the form of the underlying independent X and Y populations under H_0, beyond the point of requiring that they be continuous. This is why the test procedures based on r_s are called *distribution-free procedures*. From the null distribution of r_s, we can determine the critical value $r_{s,\alpha}$ and control the probability α of falsely rejecting H_0 when H_0 is true, and this error probability does not depend on the specific forms of the underlying continuous and independent X and Y distributions.

42. *Calculation of the Mean and Variance of r_s under the Null Hypothesis.* In displays (8.71) and (8.72), we presented formulas for the mean and variance of r_s when the null hypothesis is true. In this comment, we illustrate a direct calculation of $E_0(r_s)$ and $\text{var}_0(r_s)$ in the particular case of $n = 4$, using the null distribution of r_s obtained in Comment 41. (Later, in Comment 45, we present general derivations of $E_0(r_s)$ and $\text{var}_0(r_s)$.) The null mean, $E_0(r_s)$, is obtained by multiplying each possible value of r_s with its probability under H_0 and summing the products. Thus,

$$E_0(r_s) = -1\left(\frac{1}{24}\right) - .8\left(\frac{3}{24}\right) - .6\left(\frac{1}{24}\right) - .4\left(\frac{4}{24}\right) - .2\left(\frac{2}{24}\right)$$
$$+ 0\left(\frac{2}{24}\right) + .2\left(\frac{2}{24}\right) + .4\left(\frac{4}{24}\right) + .6\left(\frac{1}{24}\right) + .8\left(\frac{3}{24}\right) + 1\left(\frac{1}{24}\right)$$
$$= 0.$$

This is in agreement with the value stated in (8.71). A check on the expression for $\text{var}_0(r_s)$ is also easily performed, using the well-known fact that

$$\text{var}_0(r_s) = E_0(r_s^2) - \{E_0(r_s)\}^2.$$

The value of $E_0\left(r_s^2\right)$, the second moment of the null distribution of r_s, is again obtained by multiplying possible values (in this case, of r_s^2 by the corresponding probabilities under H_0 and summing). We find

$$\begin{aligned} E_0(r_s^2) = \Bigg[&(1+1)\left(\frac{1}{24}\right) + (.64+.64)\left(\frac{3}{24}\right) + (.36+.36)\left(\frac{1}{24}\right) \\ &+ (.16+.16)\left(\frac{4}{24}\right) + (.04+.04)\left(\frac{2}{24}\right) + 0\left(\frac{2}{24}\right)\Bigg] \\ &= \frac{1}{3}. \end{aligned}$$

Thus,

$$\text{var}_0(r_s) = \frac{1}{3} - 0^2 = \frac{1}{3},$$

which agrees with what we obtain using (8.72) directly, namely,

$$\text{var}_0(r_s) = \frac{1}{4-1} = \frac{1}{3}.$$

43. *Symmetry of the Distribution of r_s under the Null Hypothesis.* When H_0 is true, the distribution of r_s is symmetric about its mean 0. (See Comment 41 for verification of this when $n = 4$.) This implies that

$$P_0(r_s \le -x) = P_0(r_s \ge x), \tag{8.79}$$

for all x. Equation (8.79) is used directly to convert upper-tail probabilities to lower-tail probabilities. In particular, it follows from (8.79) that the lower αth percentile for the null distribution of r_s is $-r_{s,\alpha}$; thus, the use of $-r_{s,\alpha}$ as the critical value in procedure (8.68).

44. *Equivalent Form.* Let $(R_1, \ldots, R_n)$ and $(S_1, \ldots, S_n)$ be the vectors of separate ranks for the X and Y observations, respectively. We note that

$$\begin{aligned} \sum_{i=1}^{n}\left(R_i - \frac{n+1}{2}\right)\left(S_i - \frac{n+1}{2}\right) = &\sum_{i=1}^{n} R_i S_i - \frac{n+1}{2}\sum_{i=1}^{n} R_i \\ &- \frac{n+1}{2}\sum_{i=1}^{n} S_i + \frac{n(n+1)^2}{4}. \end{aligned}$$

However, $\sum_{i=1}^{n} R_i = \sum_{i=1}^{n} S_i = \sum_{i=1}^{n} i = n(n+1)/2$. Thus, we have

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

Combining this fact with the definition of r_s in display (8.63), we obtain an alternative computational expression for r_s, namely,

$$r_s = \frac{12\sum_{i=1}^{n} R_i S_i}{n(n^2-1)} - 3\left(\frac{n+1}{n-1}\right). \tag{8.80}$$

Thus, r_s is a linear function of the statistic $\sum_{i=1}^{n} R_i S_i$. Therefore, the various tests of independence discussed in this section can be as easily based on $\sum_{i=1}^{n} R_i S_i$ as on the more complicated formula for r_s given in (8.63) (or its counterpart in (8.64)).

45. *Large-Sample Approximation.* Under the null hypothesis H_0 (8.1), the rank vectors $(R_1, \ldots, R_n)$ and $(S_1, \ldots, S_n)$ are independent and each is uniformly distributed over the set of $n!$ permutations of $(1, 2, \ldots, n)$. It follows that the random variables $\sum_{i=1}^{n} R_i S_i$ and $\sum_{j=1}^{n} jS_j$ have the same null distribution. Combining this fact with the representation for r_s given in (8.80), it follows that

$$\begin{aligned} E_0(r_s) &= E_0\left[\frac{12\sum_{j=1}^{n} jS_j}{n(n^2-1)} - 3\left(\frac{n+1}{n-1}\right)\right] \\ &= \frac{12\sum_{j=1}^{n} jE_0(S_j)}{n(n^2-1)} - 3\left(\frac{n+1}{n-1}\right). \end{aligned}$$

Each $S_j, j = l, \ldots, n$, has a probability distribution that is uniform over the set of integers $\{1, 2, \ldots, n\}$. It follows that $E_0(S_j) = \sum_{k=1}^{n} k(1/n) = (n+1)/2$, for $j = 1, \ldots, n$. Thus, we have that

$$\begin{aligned} E_0(r_s) &= \frac{12\sum_{j=1}^{n} j\left(\frac{n+1}{2}\right)}{n(n^2-1)} - 3\left(\frac{n+1}{n-1}\right) \\ &= \frac{12\frac{n(n+1)}{2}\left(\frac{n+1}{2}\right)}{n(n^2-1)} - 3\left(\frac{n+1}{n-1}\right) = 0, \end{aligned}$$

as previously noted in (8.71). For the null variance of r_s, we first note that

$$\mathrm{var}_0(r_s) = \mathrm{var}_0\left[\frac{12\sum_{j=1}^{n} jS_j}{n(n^2-1)} - 3\left(\frac{n+1}{n-1}\right)\right] = \frac{144}{n^2(n^2-1)^2}\mathrm{var}_0\left(\sum_{j=1}^{n} jS_j\right). \tag{8.81}$$

Using a well-known expression for the variance of a sum of random variables, we have that

$$\begin{aligned} \mathrm{var}_0\left(\sum_{j=1}^{n} jS_j\right) &= \sum_{j=1}^{n}\mathrm{var}_0(jS_j) + \sum_{j=1}^{n}\sum_{k=1,k\neq j}^{n}\mathrm{cov}_0(jS_j, kS_k) \\ &= \sum_{j=1}^{n} j^2\mathrm{var}_0(S_j) + \sum_{j=1}^{n}\sum_{k=1,k\neq j}^{n} jk\ \mathrm{cov}_0(S_j, S_k). \end{aligned}$$

The joint distribution of (S_j, S_k) is the same for every $j \neq k = 1, \ldots, n$ and the marginal distribution of S_j is the same for each $j = 1, \ldots, n$. It follows that

$$\text{var}_0\left(\sum_{j=1}^{n} jS_j\right) = \text{var}_0(S_1)\sum_{j=1}^{n} j^2 + \text{cov}_0(S_1, S_2)\sum_{j=1}^{n}\sum_{k=1, k\neq j}^{n} jk.$$

Using the facts that

$$\sum_{j=1}^{n} j^2 = \frac{n(n+1)(2n+1)}{6}$$

and

$$\begin{aligned}\sum_{j=1}^{n}\sum_{k=1,k\neq j}^{n} jk &= \left(\sum_{j=1}^{n} j\right)\left(\sum_{k=1}^{n} k\right) - \sum_{j=1}^{n} j^2 \\ &= \left[\frac{n(n+1)}{2}\right]^2 - \frac{n(n+1)(2n+1)}{6} \\ &= \frac{n(n^2-1)(3n+2)}{12},\end{aligned}$$

we obtain

$$\text{var}_0\left(\sum_{j=1}^{n} jS_j\right) = \left[\frac{n(n+1)(2n+1)}{6}\text{var}_0(S_1) + \frac{n(n^2-1)(3n+2)}{12}\right.$$
$$\left.\text{cov}_0(S_1, S_2)\right].$$

Moreover, under H_0 (8.1), it can be shown (see Problems 53 and 54) that $\text{var}_0(S_1) = (n^2-1)/12$ and $\text{cov}_0(S_1, S_2) = -(n+1)/12$. Thus, we have

$$\begin{aligned}\text{var}_0\left(\sum_{j=1}^{n} jS_j\right) &= \left[\frac{n(n+1)(2n+1)(n^2-1)}{72}\right] \\ &\quad - \frac{n(n^2-1)(3n+2)(n+1)}{144} \\ &= \frac{n(n+1)(n^2-1)}{144}[2(2n+1) - (3n+2)] \\ &= \frac{n^2(n+1)(n^2-1)}{144}. \qquad (8.82)\end{aligned}$$

Combining (8.81) and (8.82) yields

$$\text{var}_0(r_s) = \frac{144}{n^2(n^2-1)^2}\left[\frac{n^2(n+1)(n^2-1)}{144}\right] = \frac{1}{n-1},$$

as noted in (8.72).

The asymptotic normality under H_0 of the standardized form

$$r_s^* = \frac{r_s - E_0(r_s)}{\{\text{var}_0(r_s)\}^{1/2}} = (n-1)^{1/2} r_s$$

follows from the fact that r_s has the same null distribution as

$$\frac{12\sum_{j=1}^{n} jS_j}{n(n^2-1)} - 3\left(\frac{n+1}{n-1}\right)$$

and standard techniques for establishing the asymptotic nonnality of a linear combination $\left(\Sigma_{j=1}^{n} \quad jS_j\right)$ of random variables. (For additional details, see Sections 8.4 and 12.3 in Randles and Wolfe (1979).)

46. *Exact Conditional Null Distribution of r_s with Ties among the X- and/or Y-Values.* To have a test with exact significance level even in the presence of tied X and/or Y observations, we must consider all the possible values of r_s corresponding to the fixed observed rank vector $(R_1, \ldots, R_n) = (r_1, \ldots, r_n)$ and every one of the $n!$ permutations of the observed rank vector $(S_1, \ldots, S_n) = (s_1, \ldots, s_n)$, where average ranks have been used to break ties in both of the rank vectors. As in Comment 41, it still follows that under H_0 each of the $n!$ possible outcomes for the ordered configurations $(s_1, \ldots, s_n)$, in conjunction with a fixed value of $(r_1, \ldots, r_n)$, both based on using average ranks to break ties, occurs with probability $1/n!$. For each such $(s_1, \ldots, s_n)$ configuration and fixed $(r_1, \ldots, r_n)$, the value of r_s is computed and the results are tabulated. We illustrate this construction for $n = 4$ and the data $(X_1, Y_1) = (2, 3.1), (X_2, Y_2) = (3.9, 4), (X_3, Y_3) = (2, 5.1)$, and $(X_4, Y_4) = (3.6, 4)$. Using average ranks to break ties, the associated X and Y rank vectors are $(R_1, R_2, R_3, R_4) = (1.5, 4, 1.5, 3)$ and $(S_1, S_2, S_3, S_4) = (1, 2.5, 4, 2.5)$, respectively. Thus, we have $D_1 = -.5, D_2 = -1.5, D_3 = 2.5, D_4 = -.5$, and an obtained value of $r_s = .1$. To assess the significance of r_s, we obtain its conditional distribution by considering the $4! = 24$ equally likely (under H_0) possible values of r_s for the fixed rank vector $(r_1, r_2, r_3, r_4) = (1.5, 4, 1.5, 3)$ in conjunction with each of the 24 permutations of the rank vector $(s_1, s_2, s_3, s_4) = (1, 2.5, 4, 2.5)$. These 24 permutations of (1, 2.5, 4, 2.5) and associated values of r_s are as follows:

(s_1, s_2, s_3, s_4)	Probability under H_0	Value or r_s
(1, 2.5, 4, 2.5)	$\frac{1}{24}$	.1
(1, 2.5, 2.5, 4)	$\frac{1}{24}$	.55
(1, 2.5, 2.5, 4)	$\frac{1}{24}$	.55
(2.5, 1, 2.5, 4)	$\frac{1}{24}$	−.2
(1, 4, 2.5, 2.5)	$\frac{1}{24}$	.85
(1, 4, 2.5, 2.5)	$\frac{1}{24}$	.85
(1, 2.5, 4, 2.5)	$\frac{1}{24}$	.1
(2.5, 1, 4, 2.5)	$\frac{1}{24}$	−.65

(s_1, s_2, s_3, s_4)	Probability under H_0	Value or r_s
(4, 1, 2.5, 2.5)	$\frac{1}{24}$	−.65
(4, 1, 2.5, 2.5)	$\frac{1}{24}$	−.65
(4, 2.5, 1, 2.5)	$\frac{1}{24}$	.1
(2.5, 4, 1, 2.5)	$\frac{1}{24}$	.85
(4, 2.5, 1, 2.5)	$\frac{1}{24}$	.1
(4, 2.5, 2.5, 1)	$\frac{1}{24}$	−.35
(4, 2.5, 2.5, 1)	$\frac{1}{24}$	−.35
(2.5, 4, 2.5, 1)	$\frac{1}{24}$	.4
(2.5, 1, 4, 2.5)	$\frac{1}{24}$	−.65
(2.5, 1, 2.5, 4)	$\frac{1}{24}$	−.2
(2.5, 2.5, 1, 4)	$\frac{1}{24}$	.55
(2.5, 2.5, 1, 4)	$\frac{1}{24}$	.55
(2.5, 4, 1, 2.5)	$\frac{1}{24}$	.85
(2.5, 4, 2.5, 1)	$\frac{1}{24}$	.4
(2.5, 2.5, 4, 1)	$\frac{1}{24}$	−.35
(2.5, 2.5, 4, 1)	$\frac{1}{24}$	−.35

This yields the null-tail probabilities

$$P_0(r_s \geq .85) = \frac{4}{24} \quad P_0(r_s \geq -.2) = \frac{16}{24}$$
$$P_0(r_s \geq .55) = \frac{8}{24} \quad P_0(r_s \geq -.35) = \frac{20}{24}$$
$$P_0(r_s \geq .4) = \frac{10}{24} \quad P_0(r_s \geq -.65) = 1$$
$$P_0(r_s \geq .1) = \frac{14}{24}.$$

This distribution is called the *conditional null distribution* or the *permutation null distribution* of r_s, given the observed sets of tied ranks $(r_1, r_2, r_3, r_4) = (1.5, 4, 1.5, 3)$ and $(s_1, s_2, s_3, s_4) = (1, 2.5, 4, 2.5)$. For the particular observed value $r_s = .1$, we have $P_0(r_s \geq .1) = \frac{14}{24}$, so that such a value does not indicate a deviation from H_0 in the direction of positive association between the X and Y variables. (We note that *both* the null expected value and null variance for r_s are different in this case of tied ranks (see Problem 51) than the corresponding expressions given in (8.71) and (8.72) for the no ties setting.)

47. *Point Estimation and Confidence Intervals Associated with* r_s. The Kendall statistic K (8.6) is directly associated with the population correlation coefficient

τ (8.2). This leads naturally to point estimators and approximate confidence intervals for τ based on K. Such is not the case for the Spearman statistic r_s (8.63). The measure of association linked with the independence tests based on r_s is

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

where τ is given by (8.2) and

$$\phi = 2P\{(Y_3 - Y_1)\}(X_2 - X_1) > 0\} - 1.$$

This measure of association η has several undesirable properties, including the facts that it is dependent on the sample size n and it is asymmetric in the X and Y labels. (For more discussion along these lines, see Fligner and Rust (1983).) As a result, point estimators and confidence intervals for η based on r_s are of little practical interest.

48. *Trend Test*. If we take $X_i = i, i = 1, \ldots, n$, and compute r_s, then the procedures based on r_s can be used as tests for a time trend in the univariate random sample $Y_1, \ldots, Y_n$.
49. *Other Uses for the r_s Statistic*. Spearman's rank correlation coefficient r_s also finds use in other settings where association is a primary issue. One such instance is in connection with Page's test for ordered alternatives in a two-way layout (see Section 7.2). Page's L statistic (7.10) is directly related to r_s. For more details, see Comment 7.22.

Properties

1. *Asymptotic Normality*. See Randles and Wolfe (1979, pp. 405–407).
2. *Efficiency*. See Section 8.7.

Problems

40. In order to study the effects of pharmaceutical and chemical agents on mucociliary clearance, doctors often use the ciliary beat frequency (CBF) as an index of ciliary activity. One accepted way to measure CBF in a subject is through the collection and analysis of an endobronchial forceps biopsy specimen. However, this technique is a rather invasive method for measuring CBF. In a study designed to assess the effectiveness of less invasive procedures for measuring CBF, Low et al. (1984) considered the alternative technique of nasal brushing. The data in Table 8.10 are a subset of the data collected by Low et al. during their investigation.

The subjects in the study were all men undergoing bronchoscopies for diagnoses of a variety of pulmonary problems. The CBF values reported in Table 8.10 are averages of 10 consecutive measurements on each subject.

Test the hypothesis of independence versus the alternative that the CBF measurements via nasal brushing and endobronchial forceps biopsy are positively associated (and, therefore, that nasal brushing is an acceptable alternative to the more invasive endobronchial forceps biopsy technique for measuring CBF).

41. Test the hypothesis of independence versus the alternative that the mean weight of introduced cysticerci is positively correlated with the mean weight of worms recovered for the tapeworm data in Table 8.3.

Table 8.10 Relation between Ciliary Beat Frequency (CBF) Values Obtained through Nasal Brushing and Endobronchial Forceps Biopsy

	CBF (hertz)	
Subject	Nasal brushing	Endobronchial forceps biopsy
1	15.4	16.5
2	13.5	13.2
3	13.3	13.6
4	12.4	13.6
5	12.8	14.0
6	13.5	14.0
7	14.5	16.0
8	13.9	14.1
9	11.0	11.5
10	15.0	14.4
11	17.0	16.0
12	13.8	13.2
13	17.4	16.6
14	16.5	18.5
15	14.4	14.5

Source: P. P. Low, C. K. Luk, M. J. Dulfano, and P. J. P. Finch (1984).

42. Test the hypothesis of independence versus the alternative that spending per high-school senior and percentage seniors graduating are positively correlated for the secondary education data in Table 8.6.

43. Show that the two expressions for r_s in displays (8.63) and (8.64) are equivalent.

44. For arbitrary number of observations, what are the smallest and largest possible values of r_s? Justify your answers.

45. Suppose $n = 5$ and we observe the data $(X_1, Y_1) = (3.7, 9.2), (X_2, Y_2) = (4.3, 9.4), (X_3, Y_3) = (5.0, 9.2), (X_4, Y_4) = (6.2, 10.4)$, and $X_5, Y_5) = (5.3, 9.2)$. What is the conditional probability distribution of r_s under H_0 (8.1) when average ranks are used to break ties among the Y's? How extreme is the observed value of r_s in this conditional null distribution? Compare this fact with that obtained by taking the observed value of r_s to the (incorrect) unconditional null distribution of r_s. (See also Problem 48.)

46. Give an example of a data set of $n \geq 10$ bivariate observations for which r_s has value 0.

47. Suppose $n = 25$. Compare the critical region for the level $\alpha = .05$ test of H_0 (8.1) versus H_2 (8.67) based on r_s with the critical region for the corresponding nominal level $\alpha = .05$ test based on the large-sample approximation.

48. For the case of $n = 5$ untied bivariate (X, Y) observations, obtain the form of the exact null (H_0) distribution of r_s. (See Comment 41.)

49. Let r_p be the Pearson product moment correlation coefficient defined in (8.78). Show that r_s (8.63) is simply this Pearson product moment correlation coefficient applied to the rank vectors $(R_1, \ldots, R_n)$ and $(S_1, \ldots, S_n)$ instead of the original $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ vectors.

50. Use the computer software R obtain the value of r_s for the secondary education data in Table 8.6, using average ranks to break the ties in the X and Y values.

51. Obtain the values of $E_0(r_s)$ and $\text{var}_0(r_s)$ corresponding to the exact conditional null distribution of r_s for the case of $n = 5$ and the tied data considered in Comment 46. Compare

these values with the corresponding values for $E_0(r_s)$ and $\text{var}_0(r_s)$ given in expressions (8.71) and (8.72) for the no ties setting. Discuss a possible reason for the difference in these null variances.

52. Use the Lake Michigan pollution data in Table 8.8 to test the hypothesis that the degree of pollution (as measured by the number of odor periods) had not changed with time against the alternative that there was a general increasing trend in the pollution of Lake Michigan over the period of 1950–1964. (See Comment 48.)

53. Let $(S_1, \ldots, S_n)$ be a vector of ranks that is uniformly distributed over the set of all $n!$ permutations of $(1, 2, \ldots, n)$. Show that the marginal probability distribution of each S_i, for $i = 1, \ldots, n$, is uniform over the set $\{1, 2, \ldots, n\}$. Use this fact to show that $E(S_i) = (n+1)/2$ and $\text{var}(S_i) = (n^2 - 1)/12$, for $i = 1, \ldots, n$.

54. Let $(S_1, \ldots, S_n)$ be a vector of ranks that is uniformly distributed over the set of all $n!$ permutations of $(1, 2, \ldots, n)$. Show that the joint marginal probability distribution of (S_i, S_j), for $i \neq j = 1, \ldots, n$, is given by

$$P(S_i = s, S_j = t) = \begin{cases} \dfrac{1}{n(n-1)}, & s \neq t = 1, \ldots, n \\ 0, & \text{otherwise.} \end{cases}$$

Use this fact to show that $\text{cov}(S_i, S_j) = -(n+1)/12$, for $i \neq j = 1, \ldots, n$.

55. The data in Table 8.11 were considered by Gentry and Pike (1970) in their study of the relationship between the mean rate of return over the period 1956 through 1969 and the 1969 value of common stock portfolios for 32 life insurance companies.

Test the hypothesis of independence versus the general alternative that the 1956–1969 mean rate of return for a stock portfolio is correlated in some fashion with its 1969 value.

8.6 A DISTRIBUTION-FREE TEST FOR INDEPENDENCE AGAINST BROAD ALTERNATIVES (HOEFFDING)

Hoeffding (1948b) proposed a test of independence that is able to detect a much broader class of alternatives to independence than the classes of alternatives that can be detected by the tests of Sections 8.1 and 8.5 based on sample correlation coefficients.

Procedure

To test the hypothesis that the X and Y random variables are independent, namely, H_0 given by (8.1), we first rank $X_1, \ldots, X_n$ jointly and let R_i denote the rank of X_i in this joint ranking, $i = 1, \ldots, n$. Then rank $Y_1, \ldots, Y_n$ jointly, and let S_i denote the rank of Y_i in this joint ranking, $i = 1, \ldots, n$. We let c_i denote the number of sample pairs (X_α, Y_α) for which both $X_\alpha < X_i$ and $Y - \alpha < Y_i$; that is,

$$c_i = \sum_{\alpha=1}^{n} \phi(X_\alpha, X_i)\phi(Y_\alpha, Y_i), \quad i = 1, \ldots, n, \tag{8.83}$$

where $\phi(a, b) = 1$ if $a < b, = 0$, otherwise.

Table 8.11 Mean Rate of Return of Common Stock Portfolios over the Period 1956–1969 and the 1969 Value of Each Equity Portfolio for 32 Life Insurance Companies

Company	Mean rate (%) of return, 1956–1969	Value of common stock portfolio, December 31, 1969 (millions of dollars)
1	18.83	96.0
2	16.98	54.6
3	15.36	84.4
4	14.65	251.5
5	14.21	131.8
6	13.68	37.3
7	13.65	109.9
8	13.07	13.5
9	12.99	76.3
10	12.81	72.6
11	11.60	42.1
12	11.51	41.5
13	11.50	56.2
14	11.41	59.3
15	11.26	1184.0
16	10.67	144.0
17	10.44	111.9
18	10.44	179.8
19	10.33	29.2
20	10.30	279.5
21	10.22	166.6
22	10.05	194.3
23	10.04	40.8
24	9.57	428.4
25	9.50	7.0
26	9.48	485.6
27	9.29	165.3
28	9.21	343.8
29	9.04	35.4
30	8.82	24.7
31	8.78	2.7
32	7.26	8.9

Source: J. Gentry and J. Pike (1970).

We set

$$Q = \sum_{i=1}^{n} (R_i - 1)(R_i - 2)(S_i - 1)(S_i - 2), \tag{8.84}$$

$$R = \sum_{i=1}^{n} (R_i - 2)(S_i - 2)c_i, \tag{8.85}$$

and

$$S = \sum_{i=1}^{n} c_i(c_i - 1), \tag{8.86}$$

and compute

$$D = \frac{Q - 2(n-2)R + (n-2)(n-3)S}{n(n-1)(n-2)(n-3)(n-4)}. \tag{8.87}$$

For a (two-sided) test of H_0 versus the alternative that X and Y are dependent (see Comment 52), at the α level of significance,

$$\text{Reject } H_0 \text{ if } D \geq d_\alpha; \quad \text{otherwise do not reject,} \tag{8.88}$$

where the constant d_α satisfies the equation $P_0(D \geq d_\alpha) = \alpha$.

Large-Sample Approximation

For the large-sample approximation, we use a statistic B, proposed by Blum, Kiefer, and Rosenblatt (1961), that is slightly different than Hoeffding's D statistic. (The tests based on B and D are, however, asymptotically equivalent, because the statistics $nD + \left(\frac{1}{36}\right)$ and nB have the same asymptotic distribution under H_0. See Comment 53.) Let

$$B = n^{-5} \sum_{i=1}^{n} [N_1(i)N_4(i) - N_2(i)N_3(i)]^2, \tag{8.89}$$

where

$$\begin{aligned}
N_1(i) &= \text{number of sample pairs } (X_\alpha, Y_\alpha) \text{ lying in the region} \\
&\quad T_1(i) = \{(x, y) : x \leq X_i \text{ and } y \leq Y_i\}, \\
N_2(i) &= \text{number of sample pairs } (X_\alpha, Y_\alpha) \text{ lying in the region} \\
&\quad T_2(i) = \{(x, y) : x > X_i \text{ and } y \leq Y_i\}, \\
N_3(i) &= \text{number of sample pairs } (X_\alpha, Y_\alpha) \text{ lying in the region} \\
&\quad T_3(i) = \{(x, y) : x \leq X_i \text{ and } y > Y_i\}, \\
N_4(i) &= \text{number of sample pairs } (X_\alpha, Y_\alpha) \text{ lying in the region} \\
&\quad T_4(i) = \{(x, y) : x > X_i \text{ and } y > Y_i\}.
\end{aligned} \tag{8.90}$$

That is, for each i, determine the number of sample pairs (X_α, Y_α) lying in each of the regions determined by the horizontal and vertical lines through the point (X_i, Y_i).

A large-sample approximation to procedure (8.88) is

$$\text{Reject } H_0 \text{ if } \frac{1}{2}\pi^4 nB \geq b_\alpha; \text{ otherwise do not reject,} \tag{8.91}$$

where the constant b_α satisfies the equation $P_0(\frac{1}{2}\pi^4 nB \geq b_\alpha) = \alpha$. Blum, Kiefer, and Rosenblatt suggested that when n is small, the error introduced when utilizing the large-sample approximation may be reduced by substituting $(n-1)B$ for nB in the left-hand side of (8.91). For a different large-sample approximation, see Comment 53.

Ties

Use average ranks and replace (8.83) by

$$c_i = \sum_{\substack{\alpha=1 \\ a \neq i}}^{n} \phi^*(X_\alpha, X_i)\phi^*(Y_\alpha, Y_i), \quad i = 1, \ldots, n, \tag{8.92}$$

where

$$\phi^*(a,b) = \begin{cases} 1, & \text{if } a < b, \\ \frac{1}{2}, & \text{if } a = b, \\ 0, & \text{otherwise.} \end{cases} \tag{8.93}$$

EXAMPLE 8.6 ***[Continuation of Example 8.5].***

We return to the data of Table 8.9 and consider the relation between the free pool of proline and collagen content in human liver cirrhosis. We apply Hoeffding's test of independence. From (8.92), we find

$$\begin{aligned}
c_1 &= \phi^*(X_2,X_1)\phi^*(Y_2,Y_1) + \phi^*(X_3,X_1)\phi^*(Y_3,Y_1) + \phi^*(X_4,X_1)\phi^*(Y_4,Y_1) \\
&\quad + \phi^*(X_5,X_1)\phi^*(Y_5,Y_1) + \phi^*(X_6,X_1)\phi^*(Y_6,Y_1) + \phi^*(X_7,X_1)\phi^*(Y_7,Y_1) \\
&= \frac{1}{2}(0) + 0\left(\frac{1}{2}\right) + 0(1) + 0(0) + 0(0) + 0(0) = 0, \\
c_2 &= \phi^*(X_1,X_2)\phi^*(Y_1,Y_2) + \phi^*(X_3,X_2)\phi^*(Y_3,Y_2) + \phi^*(X_4,X_2)\phi^*(Y_4,Y_2) \\
&\quad + \phi^*(X_5,X_2)\phi^*(Y_5,Y_2) + \phi^*(X_6,X_2)\phi^*(Y_6,Y_2) + \phi^*(X_7,X_2)\phi^*(Y_7,Y_2) \\
&= \frac{1}{2}(1) + 0(1) + 0(1) + 0(0) + 0(0) + 0(0) = \frac{1}{2}, \\
c_3 &= \phi^*(X_1,X_3)\phi^*(Y_1,Y_3) + \phi^*(X_2,X_3)\phi^*(Y_2,Y_3) + \phi^*(X_4,X_3)\phi^*(Y_4,Y_3) \\
&\quad + \phi^*(X_5,X_3)\phi^*(Y_5,Y_3) + \phi^*(X_6,X_3)\phi^*(Y_6,Y_3) + \phi^*(X_7,X_3)\phi^*(Y_7,Y_3) \\
&= 1\left(\frac{1}{2}\right) + 1(0) + 0(1) + 0(0) + 0(0) + 0(0) = \frac{1}{2}, \\
c_4 &= \phi^*(X_1,X_4)\phi^*(Y_1,Y_4) + \phi^*(X_2,X_4)\phi^*(Y_2,Y_4) + \phi^*(X_3,X_4)\phi^*(Y_3,Y_4) \\
&\quad + \phi^*(X_5,X_4)\phi^*(Y_5,Y_4) + \phi^*(X_6,X_4)\phi^*(Y_6,Y_4) + \phi^*(X_7,X_4)\phi^*(Y_7,Y_4) \\
&= 1(0) + 1(0) + 1(0) + 0(0) + 0(0) + 0(0) = 0, \\
c_5 &= \phi^*(X_1,X_5)\phi^*(Y_1,Y_5) + \phi^*(X_2,X_5)\phi^*(Y_2,Y_5) + \phi^*(X_3,X_5)\phi^*(Y_3,Y_5) \\
&\quad + \phi^*(X_4,X_5)\phi^*(Y_4,Y_5) + \phi^*(X_6,X_5)\phi^*(Y_6,Y_5) + \phi^*(X_7,X_5)\phi^*(Y_7,Y_5) \\
&= 1(1) + 1(1) + 1(1) + 1(1) + 0(0) + 0(0) = 4, \\
c_6 &= \phi^*(X_1,X_6)\phi^*(Y_1,Y_6) + \phi^*(X_2,X_6)\phi^*(Y_2,Y_6) + \phi^*(X_3,X_6)\phi^*(Y_3,Y_6) \\
&\quad + \phi^*(X_4,X_6)\phi^*(Y_4,Y_6) + \phi^*(X_5,X_6)\phi^*(Y_5,Y_6) + \phi^*(X_7,X_6)\phi^*(Y_7,Y_6) \\
&= 1(1) + 1(1) + 1(1) + 1(1) + 1(1) + 0(0) = 5, \\
c_7 &= \phi^*(X_1,X_7)\phi^*(Y_1,Y_7) + \phi^*(X_2,X_7)\phi^*(Y_2,Y_7) + \phi^*(X_3,X_7)\phi^*(Y_3,Y_7) \\
&\quad + \phi^*(X_4,X_7)\phi^*(Y_4,Y_7) + \phi^*(X_5,X_7)\phi^*(Y_5,Y_7) + \phi^*(X_6,X_7)\phi^*(Y_6,Y_7) \\
&= 1(1) + 1(1) + 1(1) + 1(1) + 1(1) + 1(1) = 6.
\end{aligned}$$

We next compute the values of Q, R, S, and D. Using (8.84) and (8.87) and the R_i's and S_i's found in Example 8.5, we obtain

$$\begin{aligned} Q &= .5(-.5)(1.5)(.5) + .5(-.5)(3)(2) + 2(1)(1.5)(.5) \\ &\quad + 3(2)(0)(-1) + 4(3)(4)(3) + 5(4)(5)(4) + 6(5)(6)(5) \\ &= 1443.81, \\ R &= -.5(.5)(0) + (-.5)(2)\left(\frac{1}{2}\right) + 1(.5)\left(\frac{1}{2}\right) \\ &\quad + 2(-1)(0) + 3(3)(4) + 4(4)(5) + 5(5)(6) \\ &= 265.75, \\ S &= 0(-1) + \frac{1}{2}\left(-\frac{1}{2}\right) + \frac{1}{2}\left(-\frac{1}{2}\right) + 0(-1) + 4(3) + 5(4) + 6(5) \\ &= 61.5 \end{aligned}$$

and

$$\begin{aligned} D &= \frac{1443.81 - 2(5)(265.75) + 5(4)(61.5)}{7(6)(5)(4)(3)} \\ &= \frac{16.31}{2520}. \end{aligned}$$

We now use these data to illustrate the computations needed to perform the large-sample approximation. For example, for the pair $(X_4, Y_4) = (8.3, 2.6)$, dividing the plane into the four regions defined by (8.90) and counting the number of sample pairs in these regions yields the $N_1(4), N_2(4), N_3(4)$, and $N_4(4)$ values defined by (8.90), namely,

$$N_1(4) = 1, \quad N_2(4) = 0, \quad N_3(4) = 3, \quad N_4(4) = 3.$$

Performing similar subdivisions and counts corresponding to the other six sample pairs, we find

$$\begin{aligned} &N_1(1) = 1, \quad N_2(1) = 2, \quad N_3(1) = 1, \quad N_4(1) = 3, \\ &N_1(2) = 2, \quad N_2(2) = 2, \quad N_3(2) = 0, \quad N_4(2) = 3, \\ &N_1(3) = 3, \quad N_2(3) = 1, \quad N_3(3) = 1, \quad N_4(3) = 3, \\ &N_1(5) = 5, \quad N_2(5) = 0, \quad N_3(5) = 0, \quad N_4(5) = 2, \\ &N_1(6) = 6, \quad N_2(6) = 0, \quad N_3(6) = 0, \quad N_4(6) = 1, \\ &N_1(7) = 7, \quad N_2(7) = 0, \quad N_3(7) = 0, \quad N_4(7) = 0. \end{aligned}$$

From (8.89), we then obtain

$$B = (7)^{-5}\{[3-2]^2 + [6-0]^2 + [6-1]^2 + [3-0]^2 + [10-0]^2 + [6-0]^2 + [0-0]^2\}$$
$$= 7^{-5}(207).$$

The sample size $n = 7$ is relatively small, so we calculate the left-hand side of (8.91) with $(7-1)B$ replacing $7B$. We find

$$\frac{1}{2}\pi^4(n-1)B = \frac{1}{2}(3.14)^4(6)(207)(7)^{-5} = 3.60.$$

The value of D given above may be reproduced using the `R` command `HoeffD`. The arguments are the two samples X and Y. Estimates of P-values and critical values d_α may be obtained from `pHoeff`. These are approximate values based on Monte Carlo simulation. For this value of D, the P-value is approximately .077. This is approximate not only due to the simulation of the distribution, but also because ties exist in the data.

For the large-sample approximation, $nD + 1/36 = .073$ and $nB = .086$. For larger n, we would see closer agreement. The command `hoeffd` in package `Hmisc` (Harrell (2012)) will perform this test and give asymptotic P-values based on B. The following is the relevant `R` output from the call `hoeffd (x, y)` where `x` is the collagen data from Table 8.9 and `y` is the proline data:

```
D

     x     y
x   1.00  0.19
y   0.19  1.00

n= 7

P
     x    y
x         0.0215
y    0.0215
```

The test statistic D has an upper bound of 1/30 for all n (Wilding and Mudholkar (2008)). `R` reports the statistic D scaled to an upper bound of 1. So, the value of D previously calculated as 16.31/2520 is scaled to $D' = (16.31/2520) \cdot 30 = .194$. The asymptotic P-value given is based on B (using n, not $n-1$ despite the low sample size). For this data, the P-value is .0215. Thus, we would reject the null hypothesis for any specified significance level α greater or equal to .0215. Note that this test is approximate due to the presence of ties. Additionally, the sample size may be inappropriate for the use of asymptotic P-values. A combination of the two factors (ties, sample size) may explain the discrepancy between the P-value based on D (0.077 and the value found here. 0.0215).

For comparison, recall that in Section 8.5, we applied the Spearman's test to the data of Table 8.9 and found the one-sided P-value to be between .05 and .10. Thus, the two-sided P-value for the test is between .10 and .20.

Comments

50. *Motivation for Hoeffding's Test*. Define

$$D^*(x,y) = P(X \le x \text{ and } Y \le y) - P(X \le x)P(Y \le y). \tag{8.94}$$

We note that $D^*(x, y) = 0$ for all (x, y) if and only if H_0 is true. This fact was used by Hoeffding in devising the test based on D. The statistic D estimates the parameter

$$\Delta_1(F) = E_F\{D^*(X', Y')\}^2, \tag{8.95}$$

where (X', Y') is a random member from the underlying bivariate population with distribution F. In other words, we may think of $D^*(x, y)$ as a measure of the deviation from H_0 at the point (x, y), and $\Delta_1(F)$ as the average value of the square of this deviation.

51. *Null Distribution of D*. In determining the null distribution of D, we can, without loss of generality, take $R_1 = i$ and obtain the associated values of D for the $n!$ possible Y rank configurations of the form $(S_1, \ldots, S_n)$. Each of these configurations has probability $[1/(n!)]$ under H_0.

52. *Consistency of D against a Broad Class of Alternatives*. The D test was designed by Hoeffding to detect a broad class of alternatives to the hypothesis of independence, and in this sense its character differs from that of the tests of independence of Sections 8.1 and 8.5 based on sample correlation coefficients. Although Hoeffding (1948b) showed that the D test is not sensitive to all alternatives to H_0, he demonstrated that under mild restrictions on the nature of the underlying bivariate population F, the test is consistent when H_0 is false. Thus, the D test detects alternatives where the X's and Y's are positively associated and alternatives where the X's and Y's are negatively associated. Furthermore, there exist populations F where X, Y are dependent and D is consistent, but the tests based on the sample correlation coefficients are not consistent.

53. *Relationship of D and B*. The statistics $nD + (\frac{1}{36})$ and nB have the same asymptotic distribution under H_0. (See Hoeffding (1948b) and Blum, Kiefer, and Rosenblatt (1961).) Thus, another large-sample approximation to procedure (8.88) is

$$\text{Reject } H_0 \text{ if } \left(\frac{1}{2}\right)\pi^4\left\{nD + \left(\frac{1}{36}\right)\right\} \geq b_\alpha; \quad \text{otherwise do not reject.}$$

54. *Development of D Test*. The test based on D was introduced by Hoeffding (1948b). The related test based on B was considered by Blum, Kiefer, and Rosenblatt (1961), who extended the approach to testing for the independence of k $(k \geq 2)$ variables. A one-sided test, similar in character to the two-sided B test, was proposed by Crouse (1966). Skaug and Tjøstheim (1993) considered the Blum–Kiefer–Rosenblatt statistic in a time-series setting and established (under mild conditions) consistency against lag one dependent alternatives. Zheng (1997) used smoothing methods to develop a nonparametric test of independence between two variables. His test is consistent against any form of dependence.

55. *Finite Sample Size Distribution of D*. Wilding and Mudholkar (2008) proposed improved methods for estimating the distribution of Hoeffding's D under the null distribution for small sample sizes. Their approximations are based on the Weibull family of distributions. The command `pHoeff` provides an approximation of this distribution using Monte Carlo simulation. For the discrete values d

of D, it provides the probability that d occurs $P(D = d)$, the lower-tail probabilities $P(D \leq d)$, and upper-tail probabilities $P(D \geq d)$. For example, if $n = 5$ and 20,000 Monte Carlo runs are used, the output is

```
          d  P(D = d)  P(D <= d)  P(D >= d)
-0.01667    0.13280    0.13280     1.00000
0.000000    0.79995    0.93275     0.86720
0.033333    0.06725    1.00000     0.06725.
```

Properties

1. *Consistency*. The test defined by (8.88) is consistent against populations for which the parameter $\Delta_1(F)$ defined by (8.95) is positive. For conditions on F that ensure that $\Delta_1(F)$ will be positive, see Hoeffding (1948b) and Yanagimoto (1970).
2. *Asymptotic Distribution*. For the asymptotic distribution of $\{nD + (\frac{1}{36})\}$, see Hoeffding (1948b) and Blum, Kiefer, and Rosenblatt (1961).

Problems

56. The data in Table 8.12 are a subset of the data obtained by Shen et al. (1970) in an experiment concerned with the hypothesis that diabetes mellitus is not simply a function of insulin deficiency and that perhaps insulin insensitivity could play an important role in the hyperglycemia of diabetes. One of the purposes of the study was to investigate the relation between the response to a glucose tolerance test and glucose impedance, a quantity describing the body tissues' resistance to glucose and expected to be constant for a given individual throughout the experimental range of glucose uptake rate in the author's study. The seven subjects represented in Table 8.12 were volunteers recently released from a minimum security prison and characterized by low plasma glucose response to oral glucose. Table 8.12 gives the weighted glucose response to an oral glucose tolerance test (X) and the glucose impedance reading (Y) for each of the seven subjects.

Use procedure (8.88) to test for impedance of weighted glucose response and glucose impedance. (Recall that procedure (8.88) is designed to detect all alternatives to the hypothesis of independence. However, if one has prior reasons or evidence to suspect that the weighted glucose response is positively correlated with glucose impedance, it would be more appropriate to focus on alternatives of positive association by using the one-sided procedure, based on Kendall's K, given by (8.8).)

Table 8.12 Weighted Glucose Response and Glucose Impedance

Subject	Weighted glucose response, X	Glucose impedance, Y
1	130	26.1
2	116	19.7
3	122	26.8
4	117	23.7
5	108	23.4
6	115	24.4
7	107	16.5

Source: S. Shen, G. M. Reaven, J. W. Farquhar, and R. H. Nakanishi (1970).

57. Apply Kendall's two-sided test based on (8.10) to the data of Table 8.12. Compare your result with the result of Problem 56.

58. Apply the large-sample approximation given in Comment 53 to the data in Table 8.9. Compare this approximation with the approximation based on (8.91) that was used in Example 8.6.

8.7 EFFICIENCIES OF INDEPENDENCE PROCEDURES

Investigation of the asymptotic relative efficiencies of tests for independence is made more difficult by our inability to define natural classes of alternatives to the hypothesis of independence. The asymptotic relative efficiencies of the test procedure (one- or two-sided) based on Kendall's statistic K (8.6) with respect to the corresponding normal theory test based on Pearson's product moment correlation coefficient r_{p} (8.78) have been found by Stuart (1954) and Konijn (1956) for a class of dependence alternatives "close" to the hypothesis of independence. Values of this asymptotic relative efficiency $e(K, r_{\mathrm{P}})$, for selected bivariate $F_{X,Y}$, are as follows:

$F_{X,Y}$:	Normal	Uniform	Double Exponential
$e(K, r_{\mathrm{p}})$:	.912	1.000	1.266

In the normal setting, natural alternatives to independence correspond to bivariate normal distributions with nonzero correlation. In this case, the asymptotic relative efficiency of the test procedure (one- or two-sided) based on Spearman's statistic r_{s} (8.63) with respect to the corresponding test procedure based on Kendall's K is 1. Moreover, the common asymptotic relative efficiency of either the test procedure based on r_{s} or the test procedure based on K with respect to the corresponding normal theory test based on r_{p} is $(3/\pi)^2 = .912$.

The point estimator and confidence interval associated with normality assumptions for the independence problem are concerned with the underlying correlation coefficient, whereas the estimator and confidence intervals based on Kendall's K relate to the parameter τ. In view of this, the estimator $\hat{\tau}$ (8.34) and the approximate confidence intervals given by (8.39) and (8.50) are not easily compared with the normal theory procedures; hence, their asymptotic efficiencies are not presented here.

We do not know of any results for the asymptotic efficiency of Hoeffding's independence test (Section 8.6).