OPTIONAL TOPIC—SPEARMAN RANK CORRELATION COEFFICIENT

Chapter 13 introduced a measure of the strength of the linear relationship between two variables called the *Pearson correlation coefficient*. Recall the correlation coefficient ranges between ±1.0, with a value of 0 indicating no linear relationship between the two variables. Recall also that Chapter 13 introduced a test for the statistical significance of the correlation coefficient. This test assumes that the data for the two variables are at least interval scaled and the joint distribution of the variables is bivariate normal. However, in cases where the level of data measurement is ordinal or when the bivariate normal distribution assumption seems not to be satisfied, an alternative nonparametric measure of correlation called *Spearman's Rank Correlation Coefficient*, or *Spearman's rho*, can be used. Like many other nonparametric statistics methods, Spearman's rho is based on the ranks of the data and does not have the data or distribution requirements of the Pearson correlation coefficient.

Coastal Distribution—Consider, for example, Coastal Distribution, which provides emergency one-day delivery service to small manufacturing companies from Baltimore to Boston. This service is important to small companies who are making the move to *just-in-time* manufacturing but do not have the well-established distribution system of larger companies. Recently, the company's sales department undertook a study of its clients to determine whether there is a significant correlation between the number of deliveries made per month and the number of employees working for the company. They hoped this information would be useful in trying to correlate future equipment needs with an expanding client list.

Table 16-1-1 presents data collected from 12 clients selected at random. The individual in charge of the study wishes to determine whether there is significant correlation. However, the data consists of counts. Of course, counts only assume integer values and, therefore, are discrete random variables. He, therefore, does not feel justified in making the bivariate normal distribution assumption required for the Pearson correlation coefficient. Thus, he has decided to calculate the nonparametric Spearman's correlation coefficient. The first step is to convert the data in Table 16-1-1 to ranks, as shown in Table 16-1-2. Note that the ranks are done separately for each variable, again ranking from lowest to highest.

The Spearman's correlation coefficient is computed using Equation 16-A.

Spearman's Rank Correlation Coefficient

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n} \qquad \textbf{16-A}$$

where:

$d_i = y_i - x_i$ (difference in ranks)
n = Sample size

TABLE 16-1-1
Sample Data for Coastal Distribution

NO. OF DELIVERIES NEEDED (*Y*)	EMPLOYEES (*X*)
23	140
11	101
10	43
4	55
20	79
15	134
7	75
42	211
3	78
2	36
15	45
6	11

TABLE 16-1-2

Data in Ranked Form for Coastal Distribution

NO. OF DELIVERIES NEEDED (*Y*)	EMPLOYEES (*X*)
11	11
7	9
6	3
3	5
10	8
8	10
5	6
12	12
2	7
1	2
9	4
4	1

Table 16-1-3 shows the calculations for the Spearman's correlation coefficient.

As shown in Table 16-1-3, the correlation is positive 0.699 for these sample data. Now the question remains whether the true population correlation is 0. If the sample size exceeds 10, the test statistic is approximated by a t-statistic with $n - 2$ degrees of freedom, as shown in Equation 16-B.

t-Statistic for $n > 10$

$$t = r_s \sqrt{\frac{n-2}{1-r_s^2}} \tag{16-B}$$

where:

r_s = Spearman's correlation coefficient
n = Sample size

The null and alternative hypotheses are

$$H_0: \rho_s = 0.0$$
$$H_A: \rho_s \neq 0.0$$

For Coastal Distribution, we compute the test statistic as follows.

$$t = 0.699 \sqrt{\frac{12-2}{1-0.699^2}}$$
$$= 3.092$$

To test the null hypothesis using the t-statistic, we go to the t-distribution table with $n - 2 = 10$ degrees of freedom for the appropriate significance level. Using a significance level of 0.05, we get critical t-values equal to ±2.228. The decision rule becomes:

If $t > 2.228$, reject H_0.
If $t < -2.228$, reject H_0.
Otherwise, do not reject H_0.

Since $t = 3.092 > 2.228$, we reject H_0 and conclude a significant correlation exists between the number of emergency deliveries needed and the number of employees working at the company.

As with the other nonparametric tests introduced in this chapter, ties are handled by giving each tied value the mean of the rank positions for which it is tied. As an example, if the ranked observations were 10, 12, 12, 13, the rankings would be 1, 2.5, 2.5, 4.

Neither Minitab nor Excel has a procedure for directly computing the Spearman's rho. However, since the formula for Spearman's rho is the same as for the Pearson correlation coefficient (included in both Minitab and Excel), we can use it providing that we have first converted the x and y variables to rankings. Excel and Minitab have functions that will transform the data into ranks.[1]

TABLE 15-1-3

Spearman's Correlation Coefficient Computation for Coastal Distribution

Y	*X*	*D*	*D*²
11	11	0	0
7	9	−2	4
6	3	3	9
3	5	−2	4
10	8	2	4
8	10	−2	4
5	6	−1	1
12	12	0	0
2	7	−5	25
1	2	−1	1
9	4	5	25
4	1	3	9

Sum of squared differences = 86

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n}$$
$$= 1 - \frac{6\,(86)}{1728 - 12}$$
$$= 1 - 0.301$$
$$= 0.699$$

[1] In Minitab, use the **Manip** and **Rank** options. In Excel, use **Tools—Data Analysis—Rank and Percentiles**.