Chapter fourteen

Measuring Association for Interval-Level Data: Pearson's Correlation Coefficient

The linear correlation coefficient

What Does a Correlation Coefficient Describe?

What are the Characteristics of Pearson's *r*?

When Might Pearson's *r* Provide Misleading Results?

What are the Characteristics of Spearman's *r*?

Testing for statistical significance

What is the Test of Statistical Significance for Pearson's *r*?

What is the Test of Statistical Significance for Spearman's *r*?

D. Weisburd and C. Britt, *Statistics in Criminal Justice*, DOI 10.1007/978-1-4614-9170-5_14,

THIS CHAPTER INTRODUCES the linear correlation coefficient, a widely used descriptive statistic that enables the researcher to describe the relationship between two interval-level measures. This situation is encountered often in criminal justice research. For example, researchers may want to establish whether number of prior arrests is related to age, education, or monthly income. Similarly, it is common in criminal justice research to ask whether the severity of a sanction measured on an interval scale (e.g., number of years sentenced to imprisonment or amount of a fine) is related to such variables as the amount stolen in an offense or the number of prior arrests or convictions of a defendant. We also examine an alternative rank-order measure of association that may be used when the linear correlation coefficient will lead to misleading results.

Measuring Association Between Two Interval-Level Variables

It may not be intuitively obvious why we need to go to the trouble of examining a new statistic to describe the relationship between two interval-level measures. Why can't we just use the means, as we did when we examined interval-level measures in Chapters 11 and 12? Suppose, for example, that we are presented with the data in Table 14.1. Can we find a simple way of expressing the relationship between these two variables?

For each of the 15 young offenders in our sample, we have information regarding age and number of arrests over the last year. The mean age of the sample overall is 17.1 years. The mean number of arrests is 4.9. These statistics describe the characteristics of our sample overall, but, importantly, they do not help us to understand the relationship between age and arrests in the study.

One way to understand this relationship is to change one of these measures into a categorical variable. For example, we might divide the

Table 14.1 Age and Number of Arrests over the Last Year for 15 Young Offenders

SUBJECT	NUMBER OF ARRESTS	AGE
1	0	14
2	1	13
3	1	15
4	2	13
5	2	14
6	3	14
7	3	17
8	4	19
9	4	21
10	6	19
11	8	16
12	9	18
13	9	20
14	10	21
15	11	22
	$\bar{X} = 4.8667$	$\bar{X} = 17.0667$

offenders into two groups—one consisting of offenders under age 18 and the other of offenders 18 and older. Then we could use the same approach taken in earlier chapters and simply compare the means for the younger and older groups, as shown in Table 14.2. On average, the older offenders appear to have more arrests than the younger offenders ($\bar{X} = 7.571$ versus $\bar{X} = 2.500$).

Similarly, we could divide arrests into categories and compare the mean age of offenders in each category. For example, Table 14.3 divides arrests into three categories: low number of arrests (less than 3), moderate number of arrests (3–8), and high number of arrests (9 and above). This table again shows that, on average, older offenders have more arrests than younger ones. In this case, the mean age for the high-arrest

Table 14.2 Mean Numbers of Arrests for Offenders Under Age 18 versus Those Age 18 and Older

NUMBER OF ARRESTS (UNDER AGE 18)	NUMBER OF ARRESTS (AGE 18 AND OLDER)
0	4
1	4
1	6
2	9
2	9
3	10
3	11
8	
$\bar{X} = 2.5000$	$\bar{X} = 7.5714$

Table 14.3 Mean Ages for Offenders with Low, Moderate, and High Numbers of Arrests

LOW NUMBER OF ARRESTS (0–2)	MODERATE NUMBER OF ARRESTS (3–8)	HIGH NUMBER OF ARRESTS (9+)
14	14	18
13	17	20
15	19	21
13	21	22
14	19	
	16	
$\bar{X} = 13.8000$	$\bar{X} = 17.6667$	$\bar{X} = 20.2500$

group was 20.3 and those for the moderate- and low-arrest groups were 17.7 and 13.8, respectively.

Although this approach allowed us to come to a general conclusion regarding the relationship between age and arrests in our sample, it forced us to convert one measure from an interval- to a nominal-level variable. In each example, we had to take a step down the ladder of measurement, which means that we did not use all of the information provided by our data. This, of course, violates one of the general principles stated earlier in the text: Statistics based on more information are generally preferred over those based on less information.

But how can we describe the relationship between two interval-level variables without converting one to a nominal scale? A logical solution to this dilemma is provided by a coefficient named after Karl Pearson, a noted British statistician who died in 1936. **Pearson's *r*** estimates the correlation, or relationship, between two measures by comparing how specific individuals stand relative to the mean of each measure. **Pearson's correlation coefficient** (*r*) has become one of the most widely used measures of association in the social sciences.

Pearson's Correlation Coefficient

Pearson's *r* is based on a very simple idea. If we use the mean of each distribution as a starting point, we can then see how specific individuals in the sample stand on each measure relative to its mean. If, in general, people who are above average on one trait are also above average on another, we can say that there is a generally positive relationship between the two traits. That is, being high, on average, on one trait is related to being high, on average, on the other. If, in contrast, people who are higher, on average, on one trait tend to be low, on average, on the

other, then we conclude that there is a negative relationship between those traits.

To illustrate these relationships, let's use the data presented in Table 14.1. If we put a plus next to each subject whose average age or number of arrests is above the mean for the sample overall and a minus next to those whose average is below the mean, a pattern begins to emerge (see Table 14.4). When a subject is above average in number of arrests, the subject is also generally above average in age. This is true for five of the six subjects above average in number of arrests (subjects 10, 12, 13, 14, and 15). Conversely, when a subject is below average in number of arrests, the subject is generally below the mean age for the sample. This is true for seven of the nine subjects below average in number of arrests (subjects 1 through 7).

Accordingly, for this sample, subjects generally tend to stand in the same relative position to the mean for both age and arrests. When individuals in the sample have a relatively high number of arrests, they also tend to be relatively older. When they have fewer arrests, they tend to be younger than average for the sample. A simple mathematical way to express this relationship is to take the product of the signs. By doing this, we find that for 12 of the 15 subjects, the result is a positive value (see Table 14.4). Put simply, 12 of the cases move in the same direction relative to the mean. The relationship observed in this case is generally positive.

Table 14.4 A Positive Relationship Between Age and Number of Arrests for 15 Young Offenders Relative to the Means

SUBJECT	NUMBER OF ARRESTS	ABOVE OR BELOW THE MEAN?	AGE	ABOVE OR BELOW THE MEAN?	PRODUCT OF THE SIGNS
1	0	−	14	−	+
2	1	−	13	−	+
3	1	−	15	−	+
4	2	−	13	−	+
5	2	−	14	−	+
6	3	−	14	−	+
7	3	−	17	−	+
8	4	−	19	+	−
9	4	−	21	+	−
10	6	+	19	+	+
11	8	+	16	−	−
12	9	+	18	+	+
13	9	+	20	+	+
14	10	+	21	+	+
15	11	+	22	+	+
	$\bar{X} = 4.8667$		$\bar{X} = 17.0667$		

Table 14.5 A Negative Relationship Between Age and Number of Arrests for 15 Young Offenders Relative to the Means

SUBJECT	NUMBER OF ARRESTS	ABOVE OR BELOW THE MEAN?	AGE	ABOVE OR BELOW THE MEAN?	PRODUCT OF THE SIGNS
1	11	+	14	−	−
2	10	+	13	−	−
3	9	+	15	−	−
4	9	+	13	−	−
5	8	+	14	−	−
6	6	+	14	−	−
7	4	−	17	−	+
8	4	−	19	+	−
9	3	−	21	+	−
10	3	−	19	+	−
11	2	−	16	−	+
12	2	−	18	+	−
13	1	−	20	+	−
14	1	−	21	+	−
15	0	−	22	+	−
	$\bar{X} = 4.8667$		$\bar{X} = 17.0667$		

A generally negative relationship can be illustrated by reversing the scores for arrests in Table 14.4. That is, the first subject does not have 0 arrests, but 11; the second does not have 1 arrest, but 10; and so forth. If we now indicate each subject's placement relative to the mean, we obtain the set of relationships listed in Table 14.5. In this table, subjects who are above average in number of arrests are generally below average in age, and subjects who are below average in number of arrests are generally above average in age. The products of these signs are mostly negative. Put differently, the scores generally move in opposite directions relative to the mean. There is still a relationship between age and number of arrests, but in this case the relationship is negative.

This is the basic logic that underlies Pearson's *r*. However, we need to take into account two additional pieces of information to develop this correlation coefficient. The first is the values of scores. Using plus (+) and minus (−) divides the scores into categories and thus does not take full advantage of the information provided by interval-level measures. Accordingly, instead of taking the product of the signs, we take the product of the difference between the actual scores and the sample means. This measure is termed the **covariation** of scores and is expressed mathematically in Equation 14.1.

$$\text{Covariation of scores} = \sum_{i=1}^{N}(X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2) \qquad \textbf{Equation 14.1}$$

Table 14.6 illustrates what we gain by including the values of the scores. We now have not only a measure of the subjects' placement on both variables relative to the mean—the sign of the relationship— but also an estimate of how strongly the scores vary from the mean. In general, for this distribution, the stronger the deviation from the mean on one variable, the stronger the deviation on the second variable. For example, if we look at the scores most distant in value from the mean in terms of number of arrests over the last year, we also find the scores most distant in terms of age. Those subjects with either zero or one arrest are not just younger, on average, than other subjects; they are among the youngest offenders overall in the sample. Similarly, those with the most arrests (10 or 11) are also the oldest members of the sample (ages 21 and 22).

The covariation of scores provides an important piece of information for defining Pearson's *r*. However, the size of the covariation between two measures depends on the units of measurement used. To permit comparison of covariation across variables with different units of measurement, we must standardize the covariation between the two

Table 14.6 Covariation of Number of Arrests (X_1) and Age (X_2) for 15 Young Offenders

	NUMBER OF ARRESTS		AGE		
SUBJECT	X_1	$X_{1i} - \bar{X}_1$	X_2	$X_{2i} - \bar{X}_2$	$(X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)$
1	0	−4.8667	14	−3.0667	14.9247
2	1	−3,8667	13	−4.0667	15.7247
3	1	−3.8667	15	−2.0667	7.9913
4	2	−2.8667	13	−4.0667	11.6580
5	2	−2.8667	14	−3.0667	8.7913
6	3	−1.8667	14	−3.0667	5.7246
7	3	−1.8667	17	−0.0667	0.1245
8	4	−0.8667	19	1.9333	−1.6756
9	4	−0.8667	21	3.9333	−3.4090
10	6	1.1333	19	1.9333	2.1910
11	8	3.1333	16	−1.0667	−3.3423
12	9	4.1333	18	0.9333	3.8576
13	9	4.1333	20	2.9333	12.1242
14	10	5.1333	21	3.9333	20.1908
15	11	6.1333	22	4.9333	30.2574
	$\bar{X}_1 = 4.8667$		$\bar{X}_2 = 17.0667$		$\sum_{i=1}^{N}(X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2) = 125.1332$

variables according to the variability within each. This is done by taking the square root of the product of the sums of the squared deviations from the mean for the two variables. Pearson's r is then the ratio between the covariation of scores and this value (see Equation 14.2). The numerator of the equation is the covariation of the two variables. The denominator of the equation standardizes this outcome according to the square root of the product of the variability found in each of the two distributions, again summed across all subjects.

$$\text{Pearson's } r = \frac{\sum_{i=1}^{N}(X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{\sqrt{\left(\sum_{i=1}^{N}(X_{1i} - \bar{X}_1)^2\right)\left(\sum_{i=1}^{N}(X_{2i} - \bar{X}_2)^2\right)}} \qquad \textbf{Equation 14.2}$$

This ratio will be positive when the covariation between the variables is positive (i.e., when subjects' scores vary in the same direction relative to the mean). It will be negative when the covariation between the variables is negative (i.e., when subjects' scores vary in opposite directions relative to the mean). The ratio will be largest when there is a good deal of covariation of the variables and when the variability of scores around each mean is small. The ratio will be smallest when there is little covariation and a good deal of variability in the measures. The range of possible values of r is between -1 and $+1$.

The Calculation

Calculating Pearson's r by hand takes a good deal of work. For that reason, in the future you will probably enter the data into a computer and then use a packaged statistical program to calculate correlation coefficients. But it will help you to understand r better if we take the time to calculate an actual example. We will use the data on number of arrests and age presented in Table 14.1. The calculations needed for Pearson's r are shown in Table 14.7.

To calculate the numerator of Equation 14.2, we must first take the simple deviation of each subject's score from the mean number of arrests (Table 14.7, column 3) and multiply it by the deviation of the subject's age from the mean age of the sample (column 6). The result, the covariation between the measures, is presented in column 8. So, for the first subject, the product of the deviations from the means is 14.9247; for the second, it is 15.7247; and so on. The covariation for our problem, 125.1332, is gained by summing these 15 products.

To obtain the denominator of the equation, we again begin with the deviations of subjects' scores from the mean. However, in this case we

Table 14.7 Calculations for the Correlation of Number of Arrests (X_1) and Age (X_2) for 15 Young Offenders

	NUMBER OF ARRESTS			AGE			
SUBJECT	X_1	$X_{1i} - \bar{X}_1$	$(X_{1i} - \bar{X}_1)^2$	X_2	$X_{2i} - \bar{X}_2$	$(X_{2i} - \bar{X}_2)^2$	$(X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	0	−4.8667	23.6848	14	−3.0667	9.4046	14.9247
2	1	−3.8667	14.9514	13	−4.0667	16.5380	15.7247
3	1	−3.8667	14.9514	15	−2.0667	4.2712	7.9913
4	2	−2.8667	8.2180	13	−4.0667	16.5380	11.6580
5	2	−2.8667	8.2180	14	−3.0667	9.4046	8.7913
6	3	−1.8667	3.4846	14	−3.0667	9.4046	5.7246
7	3	−1.8667	3.4846	17	−0.0667	0.0044	0.1245
8	4	−0.8667	0.7512	19	1.9333	3.7376	−1.6756
9	4	−0.8667	0.7512	21	3.9333	15.4708	−3.4090
10	6	1.1333	1.2844	19	1.9333	3.7376	2.1910
11	8	3.1333	9.8176	16	−1.0667	1.1378	−3.3423
12	9	4.1333	17.0842	18	0.9333	0.8710	3.8576
13	9	4.1333	17.0842	20	2.9333	8.6042	12.1242
14	10	5.1333	26.3508	21	3.9333	15.4708	20.1908
15	11	6.1333	37.6174	22	4.9333	24.3374	30.2574
		$\sum_{i=1}^{N}(X_{1i} - \bar{X}_1)^2 = 187.7333$			$\sum_{i=1}^{N}(X_{2i} - \bar{X}_2)^2 = 138.9326$		$\sum_{i=1}^{N}(X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2) = 125.1332$
	$\bar{X}_1 = 4.8667$			$\bar{X}_2 = 17.0667$			

do not multiply the two scores for each subject. Rather, we first square the deviations from each mean (columns 4 and 7) and then sum the squared deviations for each variable. The sum of the squared deviations of each score from the mean number of arrests is equal to 187.7333; the sum of the squared deviations of each score from the mean age is equal to 138.9326. Next we take the product of those deviations, and finally we take the square root of that product.

Working It Out

$$\sqrt{\left(\sum_{i=1}^{N}(X_{1i} - \bar{X}_1)^2\right)\left(\sum_{i=1}^{N}(X_{2i} - \bar{X}_2)^2\right)} = \sqrt{(187.7333)(138.9326)}$$

$$= \sqrt{26{,}082.2755}$$

$$= 161.5001$$

This leaves us with a value of 161.5001 for the denominator of our equation.

We are now ready to calculate Pearson's r for our example. We simply take the covariation of 125.1332 and divide it by 161.5001, to get 0.7748:

Working It Out

$$\text{Pearson's } r = \frac{\sum_{i=1}^{N}(X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{\sqrt{\left(\sum_{i=1}^{N}(X_{1i} - \bar{X}_1)^2\right)\left(\sum_{i=1}^{N}(X_{2i} - \bar{X}_2)^2\right)}}$$

$$= \frac{125.1332}{\sqrt{(187.7333)(138.9326)}}$$

$$= \frac{125.1332}{161.5001}$$

$$= 0.7748$$

Our correlation is about 0.77, meaning that the correlation between age and number of arrests is a positive one. As number of arrests increases, so does the average age of the offenders in our sample. But what is the strength of this relationship? Is it large or small? As discussed in Chapter 12 when we examined the correlation coefficient eta, whether something is large or small is in good measure a value judgment. The answer depends in part on how the result compares to other research in the same area of criminal justice. For example, if other studies produced correlations that were generally much smaller, we might conclude that the relationship in our sample was a very strong one. Jacob Cohen suggests that a correlation of 0.10 may be defined as a small relationship; a correlation of 0.30, a moderate relationship; and a correlation of 0.50, a large relationship.[1] On this yardstick, the relationship observed in our example is a very strong one.

A Substantive Example: Crime and Unemployment in California

An area of study that has received extensive attention from criminologists is the relationship between crime rates and other social or economic indicators, such as unemployment. An example of such data is provided in Table 14.8, which presents the burglary rate and the unemployment

[1]See Jacob Cohen, *Statistical Power Analysis for the Behavioral Sciences* (Hillsdale, NJ: Lawrence Erlbaum, 1988), pp. 79–80. In Chapter 21, we discuss in greater detail how statisticians develop standardized estimates of "effect size."

Table 14.8

Unemployment Rate and Burglary Rate for 58 California Counties in 1999

COUNTY	UNEMPLOYMENT RATE (X_1)	BURGLARY RATE (PER 100,000) (X_2)
Alameda	3.5	837.89
Alpine	9.1	2,037.49
Amador	4.6	818.55
Butte	6.8	865.04
Calaveras	6.9	989.76
Colusa	15.9	520.06
Contra Costa	3.0	664.73
Del Norte	8.0	1,200.91
El Dorado	3.9	509.87
Fresno	13.4	924.10
Glenn	11.2	845.29
Humboldt	6.4	1,027.79
Imperial	23.4	1,526.40
Inyo	5.7	511.12
Kern	11.4	960.18
Kings	13.1	649.22
Lake	7.7	1,333.21
Lassen	7.0	361.24
Los Angeles	5.9	610.28
Madera	11.5	929.32
Marin	1.9	526.98
Mariposa	7.4	775.92
Mendocino	6.7	843.92
Merced	13.3	1,214.69
Modoc	8.5	325.08
Mono	6.7	957.95
Monterey	9.6	570.14
Napa	3.3	477.54
Nevada	4.1	455.37
Orange	2.6	464.52
Placer	3.2	646.12
Plumas	9.0	1,030.58
Riverside	5.4	1,049.18
Sacramento	4.2	925.61
San Benito	8.0	845.75
San Bernadino	4.8	883.02
San Diego	3.1	539.82
San Francisco	3.0	744.81
San Joaquin	8.8	896.85
San Luis Obispo	3.2	540.79
San Mateo	2.0	355.82
Santa Barbara	3.9	444.07
Santa Clara	3.0	347.57
Santa Cruz	6.3	647.73
Shasta	7.0	823.95
Sierra	9.2	699.71
Siskiyou	10.3	575.09
Solano	4.6	769.30
Sonoma	2.7	555.44
Stanislaus	10.5	1,057.99
Sutter	13.0	859.11
Tehama	6.7	816.55
Trinity	11.5	676.23
Tulare	16.5	1,047.32
Tuolumne	6.5	908.79
Ventura	4.8	491.86
Yolo	4.3	591.28
Yuba	11.6	1,366.76

Table 14.9 Calculations for the Correlation of Unemployment Rate (X_1) and Burglary Rate (X_2) for 58 California Counties

UNEMPLOYMENT RATE		BURGLARY RATE		
X_{1i}	$(X_{1i} - \bar{X}_1)^2$	X_{2i}	$(X_{2i} - \bar{X}_2)^2$	$(X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)$
3.5	15.2639	837.89	2,208.9342	−183.6216
9.1	2.8666	2,037.49	1,554,009.8148	2,110.6173
4.6	7.8787	818.55	765.0369	−77.6369
6.8	0.3683	865.04	5,498.1187	−45.0012
6.9	0.2569	989.76	39,548.9985	−100.8068
15.9	72.1327	520.06	73,349.2681	−2,300.1922
3.0	19.4208	664.73	15,916.5222	555.9776
8.0	0.3518	1,200.91	168,115.8264	243.1824
3.9	12.2983	509.87	78,972.6338	985.5115
13.4	35.9172	924.10	17,744.7176	798.3367
11.2	14.3876	845.29	2,959.2838	206.3420
6.4	1.0138	1,027.79	56,121.2783	−238.5339
23.4	255.7792	1,526.40	540,973.9304	11,763.0738
5.7	2.9135	511.12	78,271.6446	477.5406
11.4	15.9448	960.18	28,658.8671	675.9891
13.1	32.4114	649.22	20,070.5872	−806.5455
7.7	0.0859	1,333.21	294,110.2232	158.9538
7.0	0.1656	361.24	184,599.7240	174.8249
5.9	2.2707	610.28	32,620.2250	272.1623
11.5	16.7535	929.32	19,162.6711	566.6050
1.9	30.3259	526.98	69,648.8576	1,453.3298
7.4	0.0000	775.92	224.1219	0.1033
6.7	0.4997	843.92	2,812.1067	−37.4864
13.3	34.7286	1,214.69	179,605.8467	2,497.4917
8.5	1.1949	325.08	216,979.6082	−509.1777
6.7	0.4997	957.95	27,908.8097	−118.0942
9.6	4.8097	570.14	48,730.8716	−484.1284
3.3	16.8666	477.54	98,188.6612	1,286.9000
4.1	10.9356	455.37	112,574.1401	1,109.5334
2.6	23.1063	464.52	106,517.8338	1,568.8313
3.2	17.6980	646.12	20,958.5556	609.0359
9.0	2.5380	1,030.58	57,450.9605	381.8490
5.4	4.0276	1,049.18	66,713.3625	−518.3608
4.2	10.2842	925.61	18,149.2898	−432.0313
8.0	0.3518	845.75	3,009.5428	32.5371
4.8	6.7959	883.02	8,487.8079	−240.1719
3.1	18.5494	539.82	63,036.4964	1,081.3364
3.0	19.4208	744.81	2,123.4309	203.0730
8.8	1.9407	896.85	11,227.3733	147.6119
3.2	17.6980	540.79	62,550.3601	1,052.1486
2.0	29.2346	355.82	189,286.5140	2,352.3838
3.9	12.2983	444.07	120,284.5979	1,216.2655
3.0	19.4208	347.57	196,533.2430	1,953.6700
6.3	1.2252	647.73	20,494.9860	158.4646
7.0	0.1656	823.95	1,092.9173	−13.4518
9.2	3.2152	699.71	8,313.9201	−163.4961
10.3	8.3700	575.09	46,569.9421	−624.3330
4.6	7.8787	769.30	466.1583	60.6029
2.7	22.1549	555.44	55,437.0321	1,108.2429
10.5	9.5673	1,057.99	71,342.0361	826.1648
13.0	31.2828	859.11	4,653.8729	381.5574

(continued on next page)

Table 14.9 Calculations for the Correlation of Unemployment Rate (X_1) and Burglary Rate (X_2) for 58 California Counties (*Continued*)

UNEMPLOYMENT RATE		BURGLARY RATE		
X_{1i}	$(X_{1i} - \bar{X}_1)^2$	X_{2i}	$(X_{2i} - \bar{X}_2)^2$	$(X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)$
6.7	0.4997	816.55	658.3997	−18.1386
11.5	16.7535	676.23	13,147.0761	−469.3177
16.5	82.6845	1,047.32	65,755.9859	2,331.7373
6.5	0.8225	908.79	13,900.2449	−106.9229
4.8	6.7959	491.86	89,419.3595	779.5431
4.3	9.6528	591.28	39,844.4316	620.1705
11.6	17.5821	1,366.76	331,625.4507	2,414.6776
	$\sum_{i=1}^{N} = 1{,}010.3570$		$\sum_{i=1}^{N} = 5{,}659{,}402.5114$	$\sum_{i=1}^{N} = 37{,}128.9297$
	$\bar{X}_1 = 7.4069$		$\bar{X}_2 = 790.8907$	

rate for all 58 counties in California in 1999. The burglary rate represents the number of burglaries per 100,000 population, and the unemployment rate represents the percentage of persons actively looking for work who have not been able to find employment.

Table 14.9 presents the detailed calculations for the unemployment and burglary data from California. The sum of the covariations is 37,128.9297, the sum of the squared deviations for the unemployment rate is 1,010.3570, and the sum of the squared deviations for the burglary rate is 5,659,402.5114. After inserting these values into Equation 14.2, we find that $r = 0.4910$. The positive correlation between the unemployment rate and the burglary rate means that counties with higher rates of unemployment also tended to have higher rates of burglary, while counties with lower rates of unemployment tended to have lower rates of burglary.

Working It Out

$$r = \frac{\sum_{i=1}^{N}(X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{\sqrt{\left(\sum_{i=1}^{N}(X_{1i} - \bar{X}_1)^2\right)\left(\sum_{i=1}^{N}(X_{2i} - \bar{X}_2)^2\right)}}$$

$$= \frac{37{,}128{,}9297}{\sqrt{(1{,}010.3570)(5{,}659{,}402.5114)}}$$

$$= 0.4910$$

Nonlinear Relationships and Pearson's *r*

Pearson's *r* allows us to assess the correlation between two interval-level measures, taking into account the full amount of information that these measures provide. However, it assesses the strength of only a **linear relationship.** If the correlation between two variables is not linear, then Pearson's *r* will give very misleading results.

A simple way to see this is to look at **scatterplots,** or **scatter diagrams,** representing different types of relationships. A scatterplot positions subjects according to their scores on both variables being examined. Figure 14.1 represents the subjects in our example concerning age and number of arrests. The first case (age = 14, arrests = 0) is represented by the dot with a 1 next to it. The overall relationship in this example is basically linear and positive. That is, the dots move together in a positive direction (as age increases, so too do arrests). A scatterplot of the data in Table 14.5, where the highest number of arrests is found among younger rather than older subjects, is presented in Figure 14.2. In this case, the scatterplot shows a negative relationship (as age increases, number of arrests decreases).

But what would happen if there were a **curvilinear relationship** between age and number of arrests? That is, what if the number of arrests for both younger and older subjects was high, and the number for those of average age was low? This relationship is illustrated in Figure 14.3. In

Figure 14.1 *Scatterplot Showing a Positive Relationship Between Age and Number of Arrests for 15 Subjects*

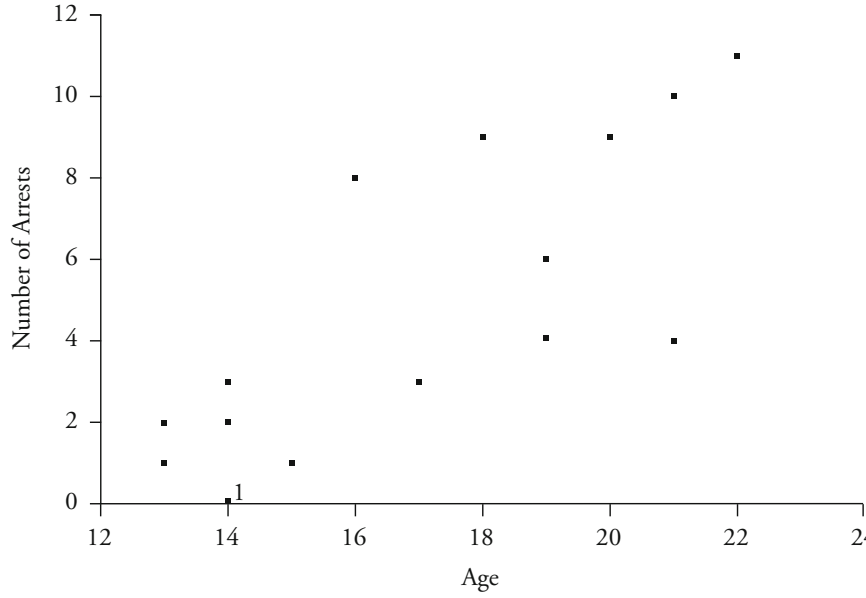


Figure 14.2

Scatterplot Showing a Negative Relationship Between Age and Number of Arrests for 15 Subjects

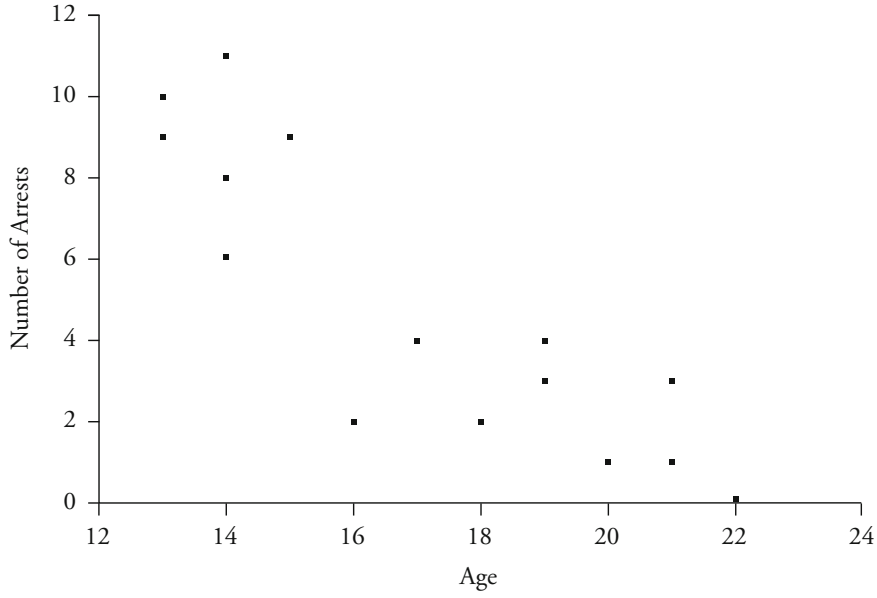


Figure 14.3

Scatterplot Showing a Curvilinear Relationship Between Age and Number of Arrests for 15 Subjects

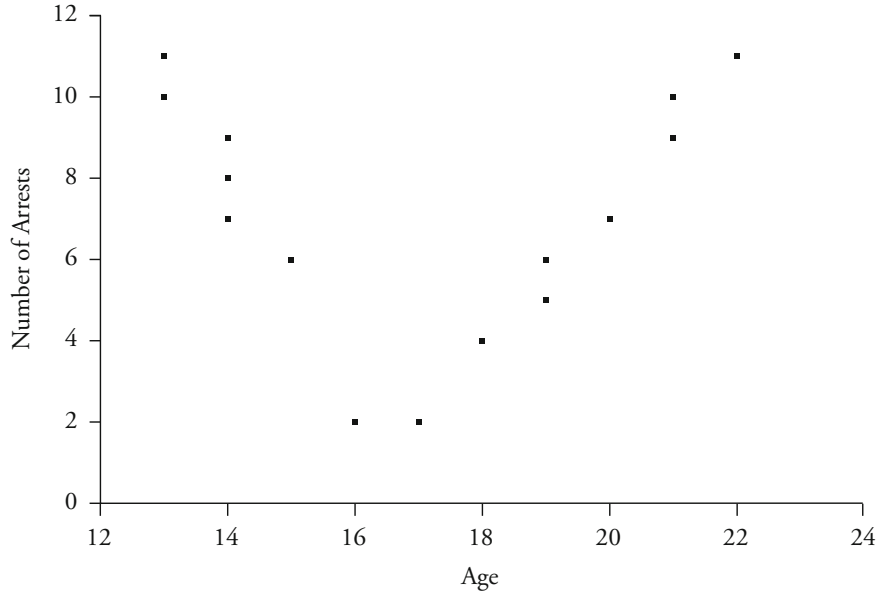


Table 14.10 Curvilinear Relatonship: Calculations for the Correlation of Number of Arrests (X_1) and Age (X_2) for 15 Young Offenders

	NUMBER OF ARRESTS			AGE			
SUBJECT	X_1	$X_{1i} - \bar{X}_1$	$(X_{1i} - \bar{X}_1)^2$	X_2	$X_{2i} - \bar{X}_2$	$(X_{2i} - \bar{X}_2)^2$	$(X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)$
1	9	1.8667	3.4846	14	−3.0667	9.4046	−5.7246
2	11	3.8667	14.9514	13	−4.0667	16.5380	−15.7247
3	6	−1.1333	1.2844	15	−2.0667	4.2712	2.3422
4	10	2.8667	8.2180	13	−4.0667	16.5380	−11.6580
5	8	0.8667	0.7512	14	−3.0667	9.4046	−2.6579
6	7	−0.1333	0.0178	14	−3.0667	9.4046	0.4088
7	2	−5.1333	26.3508	17	−0.0667	0.0044	0.3424
8	5	−2.1333	4.5510	19	1.9333	3.7376	−4.1243
9	9	1.8667	3.4846	21	3.9333	15.4708	7.3423
10	6	−1.1333	1.2844	19	1.9333	3.7376	−2.1910
11	2	−5.1333	26.3508	16	−1.0667	1.1378	5.4757
12	4	−3.1333	9.8176	18	0.9333	0.8710	−2.9243
13	7	−0.1333	0.0178	20	2.9333	8.6042	−0.3910
14	10	2.8667	8.2180	21	3.9333	15.4708	11.2756
15	11	3.8667	14.9514	22	4.9333	24.3374	19.0756
	$\sum_{i=1}^{N}(X_{1i} - \bar{X}_1)^2 = 123.7338$			$\sum_{i=1}^{N}(X_{2i} - \bar{X}_2)^2 = 138.9326$			$\sum_{i=1}^{N}(X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2) = 0.8668$
	$\bar{X}_1 = 7.1333$			$\bar{X}_2 = 17.0667$			

this scatterplot, the relationship is clear: The number of arrests declines until age 17 and then increases. However, Pearson's r for these data is close to 0. If there is a relationship, why does this happen? Table 14.10 shows why. Subjects who are either much above or much below the mean for age have large numbers of arrests. The covariance for these subjects is accordingly very high. However, for those below the mean in age, the covariance is negative, and for those above the mean, the covariance is positive. If we add these scores together, they cancel each other out. As a result, Pearson's r for this example is close to 0.

Working It Out

$$\text{Pearson's } r = \frac{\sum_{i=1}^{N}(X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{\sqrt{\left(\sum_{i=1}^{N}(X_{1i} - \bar{X}_1)^2\right)\left(\sum_{i=1}^{N}(X_{2i} - \bar{X}_2)^2\right)}}$$

$$= \frac{0.8668}{\sqrt{(123.7338)(138.9326)}}$$

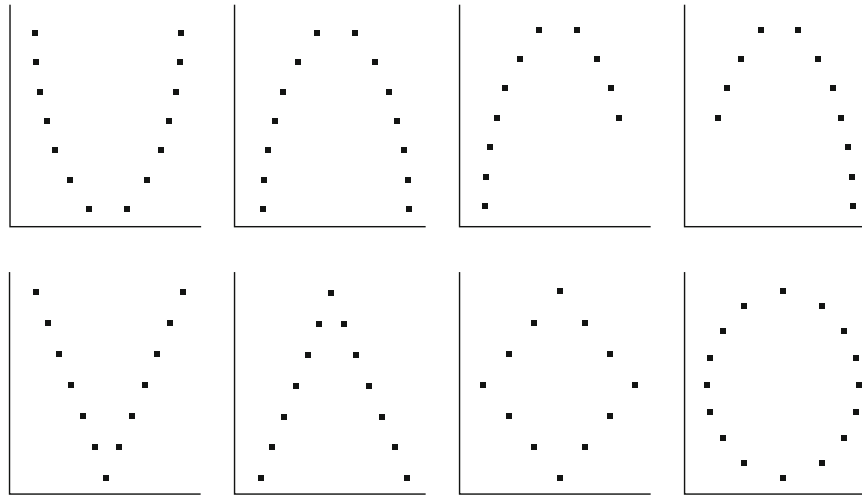
$$= 0.0066$$

Pearson's *r* will provide a good estimate of correlation when the relationship between two variables is approximately linear. However, a strong nonlinear relationship will lead to a misleading correlation coefficient. Figure 14.4 provides examples of a number of nonlinear relationships. These examples illustrate why it is important to look at the scatterplot of the relationship between two interval-level measures to establish that it is linear before calculating Pearson's correlation coefficient. Linear relationships are much more common in criminal justice than nonlinear ones. But you would not want to conclude, based on *r*, that there was a very small relationship between two variables when in fact there was a very strong nonlinear correlation between them.

What can you do if the relationship is nonlinear? Sometimes the solution is simply to break up the distribution of scores. For example, Figure 14.3 shows a nonlinear relationship that results in an *r* of 0.007. If we break this distribution at the point where it changes direction, we can calculate two separate Pearson's correlation coefficients, each for a linear relationship. The first would provide an estimate of the relationship for younger offenders (which is negative), and the second an estimate of the relationship for older offenders (which is positive).

For some nonlinear relationships, you may want to consider using alternative statistics. For example, it may be worthwhile to break up your

Figure 14.4 *Examples of Nonlinear Relationships*

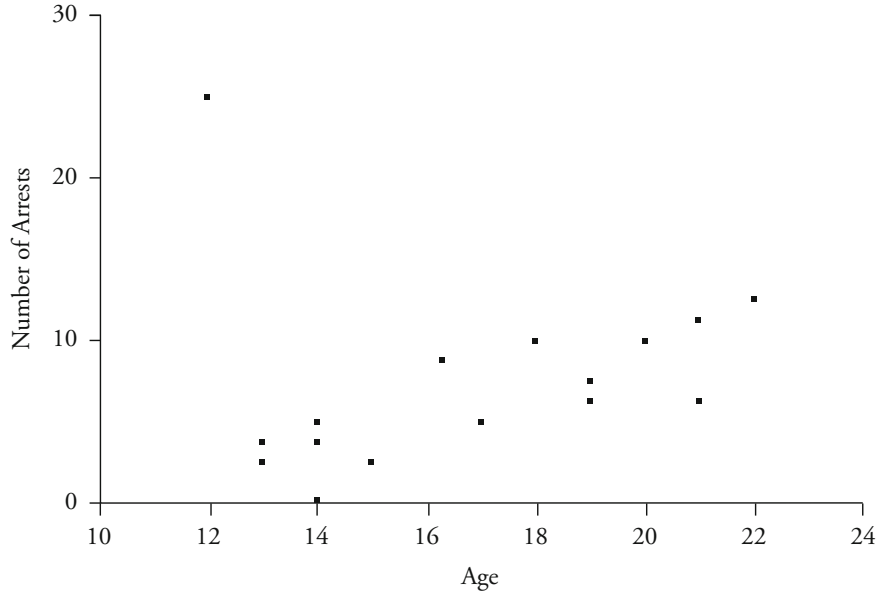


sample into a number of groups, or categories, and then look at the means for each. In some cases, it may be possible to change the form of the variables and, in doing so, increase the linearity of the relationship being examined. Although such transformations are beyond the scope of this text, you should be aware that they provide one solution to problems of nonlinearity.[2]

Beware of Outliers

For Pearson's *r*, as for other statistics based on deviations from the mean, outliers can have a strong impact on results. For example, suppose we add to our study of age and number of arrests (from Table 14.1) one subject who was very young (12) but nonetheless had an extremely large number of arrests over the last year (25), as shown in the scatterplot in Figure 14.5. If we take the covariation for this one relationship (see sub-

Figure 14.5 *Scatterplot Showing the Relationship Between Age and Number of Arrests for 16 Subjects, Including an Outlier*



[2]For a discussion of this issue, see. J. Fox, *Linear Statistical Models and Related Methods* (New York: Wiley, 1994).

Table 14.11 Calculations for the Correlation of Number of Arrests (X_1) and Age (X_2) for 16 Young Offenders

	NUMBER OF ARRESTS			AGE			
SUBJECT	X_1	$X_{1i} - \bar{X}_1$	$(X_{1i} - \bar{X}_1)^2$	X_2	$X_{2i} - \bar{X}_2$	$(X_{2i} - \bar{X}_2)^2$	$(X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)$
1	0	−6.1250	37.5156	14	−2.7500	7.5625	16.84375
2	1	−5.1250	26.2656	13	−3.7500	14.0625	19.21875
3	1	−5.1250	26.2656	15	−1.7500	3.0625	8.96875
4	2	−4.1250	17.0156	13	−3.7500	14.0625	15.46875
5	2	−4.1250	17.0156	14	−2.7500	7.5625	11.34375
6	3	−3.1250	9.7656	14	−2.7500	7.5625	8.59375
7	3	−3.1250	9.7656	17	0.2500	0.0625	−0.78125
8	4	−2.1250	4.5156	19	2.2500	5.0625	−4.78125
9	4	−2.1250	4.5156	21	4.2500	18.0625	−9.03125
10	6	−0.1250	0.0156	19	2.2500	5.0625	−0.28125
11	8	1.8750	3.5156	16	−0.7500	0.5625	−1.40625
12	9	2.8750	8.2656	18	1.2500	1.5625	3.59375
13	9	2.8750	8.2656	20	3.2500	10.5625	9.34375
14	10	3.8750	15.0156	21	4.2500	18.0625	16.46875
15	11	4.8750	23.7656	22	5.2500	27.5625	25.59375
16	25	18.8750	356.2656	12	−4.7500	22.5625	−89.65625
	$\bar{X}_1 = 6.1250$		$\sum_{i=1}^{N}(X_{1i} - \bar{X}_1)^2 = 567.7496$	$\bar{X}_2 = 16.7500$		$\sum_{i=1}^{N}(X_{2i} - \bar{X}_2)^2 = 163.000$	$\sum_{i=1}^{N}(X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2) = 29.5000$

ject 16 in Table 14.11), we see that it is very large relative to that of other subjects in our analysis. Because it is negative, it cancels out the positive covariation produced by the other subjects in the sample. Indeed, with this subject included, the correlation decreases from 0.77 to 0.10.

Working It Out

$$\text{Pearson's } r = \frac{\sum_{i=1}^{N}(X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{\sqrt{\left(\sum_{i=1}^{N}(X_{1i} - \bar{X}_1)^2\right)\left(\sum_{i=1}^{N}(X_{2i} - \bar{X}_2)^2\right)}}$$

$$= \frac{29.5000}{\sqrt{(567.7496)(163)}}$$

$$= 0.0970$$

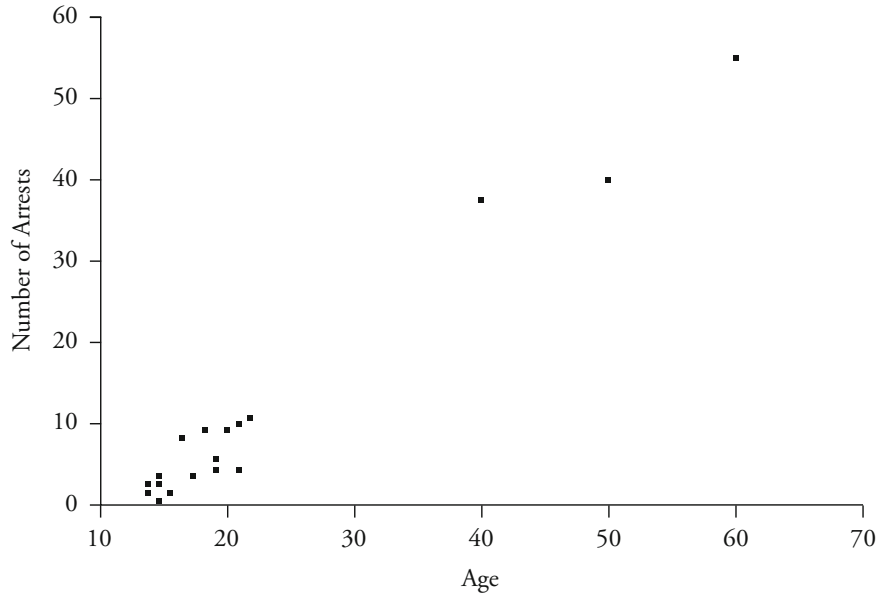
What should you do when faced with outliers? When you have just a few deviant cases in your sample, the best decision may be to exclude them from your analysis. If you take this approach, it is important to clearly state that certain cases have been excluded and to explain why. Before excluding outliers, however, you should compare the correlations with and without them. When samples are large, deviant cases may have a relatively small impact; thus, including them may not lead to misleading results.

When there are a relatively large number of outliers that follow the general pattern of relationships in your data, it may be better to choose an alternative correlation coefficient rather than exclude such cases. For example, suppose we add to our study of age and arrests three different subjects for whom the relationships are similar to those noted previously, but the number of arrests and the average age are much higher (see subjects 16, 17, and 18 in Table 14.12). These data are depicted in the scat-

Table 14.12 Calculations for the Correlation of Number of Arrests (X_1) and Age (X_2) for 18 Young Offenders

	NUMBER OF ARRESTS			AGE			
SUBJECT	X_1	$X_{1i} - \bar{X}_1$	$(X_{1i} - \bar{X}_1)^2$	X_2	$X_{2i} - \bar{X}_2$	$(X_{2i} - \bar{X}_2)^2$	$(X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)$
1	0	−11.2778	127.1888	14	−8.5556	73.1983	96.4883
2	1	−10.2778	105.6332	13	−9.5556	91.3095	98.2105
3	1	−10.2778	105.6332	15	−7.5556	57.0871	77.6549
4	2	−9.2778	86.0776	13	−9.5556	91.3095	88.6549
5	2	−9.2778	86.0776	14	−8.5556	73.1983	79.3771
6	3	−8.2778	68.5220	14	−8.5556	73.1983	70.8215
7	3	−8.2778	68.5220	17	−5.5556	30,8647	45.9881
8	4	−7.2778	52.9664	19	−3.5556	12.6423	25.8769
9	4	−7.2778	52.9664	21	−1.5556	2.4199	11.3213
10	6	−5.2778	27.8552	19	−3.5556	12.6423	18.7675
11	8	−3.2778	10.7440	16	−6.556	42.9759	21.4879
12	9	−2.2778	5.1884	18	−4.5556	20.7535	10.3767
13	9	−2.2778	5.1884	20	−2.5556	6.5311	5.8211
14	10	−1.2778	1.6328	21	−1.5556	2.4199	1.9877
15	11	−0.2778	0.0772	22	−0.5556	0.3087	0.1543
16	36	24.7222	611.1872	40	17.4444	304.3071	431.2639
17	40	28.7222	824.9648	50	27.4444	753.1951	788.2635
18	54	42.7222	1,825.1864	60	37.4444	1,402.0831	1,5999.7071
	$\bar{X}_1 = 11.2778$		$\sum_{i=1}^{N}(X_{1i} - \bar{X}_1)^2$	$\bar{X}_2 = 22.5556$		$\sum_{i=1}^{N}(X_{2i} - \bar{X}_2)^2$	$\sum_{i=1}^{N}(X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)$
			$= 4,065.6116$			$= 3,050.4446$	$= 3,472.2214$

Figure 14.6 *Scatterplot Showing the Relationship Between Age and Number of Arrests for 18 Subjects, Including Three Outliers Who Follow the General Pattern*



terplot in Figure 14.6. In such situations, Pearson's r is likely to give a misleading view of the relationship between the two variables. For our example, the correlation changes from 0.77 to 0.99.

Working It Out

$$\text{Pearson's } r = \frac{\sum_{i=1}^{N}(X_{1i} - \overline{X}_1)(X_{2i} - \overline{X}_2)}{\sqrt{\left(\sum_{i=1}^{N}(X_{1i} - \overline{X}_1)^2\right)\left(\sum_{i=1}^{N}(X_{2i} - \overline{X}_2)^2\right)}}$$

$$= \frac{3{,}472.2214}{\sqrt{(4{,}065.6116)(3{,}050.4446)}}$$

$$= 0.9859$$

In such situations, you may want to use a rank-order correlation coefficient called **Spearman's *r***. Pearson's r is generally more appropriate for interval-level data. However, where a number of outliers are found in

the distribution, **Spearman's correlation** coefficient can provide a useful alternative.

Spearman's Correlation Coefficient

Spearman's r compares the rank order of subjects on each measure rather than the relative position of each subject to the mean. Like Pearson's r, its range of possible values is between -1 and $+1$. It is calculated using Equation 14.3.

$$r_s = 1 - \frac{6\sum_{i=1}^{N} D_i^2}{N(N^2 - 1)} \qquad \textbf{Equation 14.3}$$

Let's calculate r_s for our original example with 15 cases (from Table 14.1), and for the example with the additional three outliers (from Table 14.12). To carry out the calculation, we must first rank order the cases, as

Table 14.13 Calculation of Difference in Rank (D) for Spearman's r for 15 Young Offenders

SUBJECT	NUMBER OF ARRESTS	RANK ARRESTS Rk_1	AGE	RANK AGE Rk_2	D $(Rk_1 - Rk_2)$	D^2
1	0	1	14	4	−3	9
2	1	2.5	13	1.5	1	1
3	1	2.5	15	6	−3.5	12.25
4	2	4.5	13	1.5	3	9
5	2	4.5	14	4	0.5	0.25
6	3	6.5	14	4	2.5	6.25
7	3	6.5	17	8	−1.5	2.25
8	4	8.5	19	10.5	−2	4
9	4	8.5	21	13.5	−5	25
10	6	10	19	10.5	−0.5	0.25
11	8	11	16	7	4	16
12	9	12.5	18	9	3.5	12.25
13	9	12.5	20	12	0.5	0.25
14	10	14	21	13.5	0.5	0.25
15	11	15	22	15	0	0
	$\bar{X}_1 = 4.8667$		$\bar{X}_2 = 17.0667$			$\sum_{i=1}^{N} D_i^2 = 98$

Table 14.14 Calculation of Difference in Rank (D) for Spearman's r for 18 Young Offenders

SUBJECT	NUMBER OF ARRESTS	RANK ARRESTS Rk_1	AGE	RANK AGE Rk_2	D $(Rk_1 - Rk_2)$	D^2
1	0	1	14	4	−3	9
2	1	2.5	13	1.5	1	1
3	1	2.5	15	6	−3.5	12.25
4	2	4.5	13	1.5	3	9
5	2	4.5	14	4	0.5	0.25
6	3	6.5	14	4	2.5	6.25
7	3	6.5	17	8	−1.5	2.25
8	4	8.5	19	10.5	−2	4
9	4	8.5	21	13.5	−5	25
10	6	10	19	10.5	−0.5	0.25
11	8	11	16	7	4	16
12	9	12.5	18	9	3.5	12.25
13	9	12.5	20	12	0.5	0.25
14	10	14	21	13.5	0.5	0.25
15	11	15	22	15	0	0
16	36	16	40	16	0	0
17	40	17	50	17	0	0
18	54	18	60	18	0	0
	$\bar{X}_1 = 11.2778$		$\bar{X}_2 = 22.5556$			$\sum_{i=1}^{N} D_i^2 = 98$

shown in Tables 14.13 and 14.14. We then take the squared difference in ranks for each subject on the two measures and sum it across all the cases in our example. This value is multiplied by 6, and then divided by $N(N^2 - 1)$. The final figure is then subtracted from 1.

Working It Out No Outliers

$$r_s = 1 - \frac{6\sum_{i=1}^{N} D_i^2}{N(N^2 - 1)}$$

$$= 1 - \frac{(6)(98)}{(15)(224)}$$

$$= 1 - \frac{588}{3360}$$

$$= 1 - 0.1750$$

$$= 0.8250$$

Working It Out With Outliers

$$r_s = 1 - \frac{6\sum_{i=1}^{N} D_i^2}{N(N^2 - 1)}$$

$$= 1 - \frac{(6)(98)}{(18)(323)}$$

$$= 1 - \frac{588}{5814}$$

$$= 1 - 0.1011$$

$$= 0.8989$$

The correlation coefficients for our two distributions (without outliers and with outliers) are similar. In the case without the outliers, $r_s = 0.83$; in the case with them, $r_s = 0.90$. The outliers here do not have as much of an impact because Spearman's correlation coefficient does not take into account the actual values of the scores, but only their ranks in the distribution. Note that the correlation coefficient obtained here for the 15 cases, $r_s = 0.83$, is a bit larger than, although similar to, $r = 0.77$. Which is the better estimate of the correlation between these two variables? In the case without the outliers, Pearson's *r* would be preferred because it takes into account more information (order as well as value). In the case with the outliers, however, Spearman's *r* would be preferred because it is not affected by the extreme values of the three outliers, but only by their relative positions in the distributions.

Testing the Statistical Significance of Pearson's *r*

As in Chapter 13, our emphasis in this chapter has been not on statistical inference but rather on statistical description. Our concern has been to describe the strength or nature of the relationship between two interval-level variables. Nonetheless, it is important here, as before, to define whether the differences observed in our samples can be inferred to the populations from which they were drawn.

Statistical Significance of *r*: The Case of Age and Number of Arrests

We can use the *t*-distribution introduced in Chapter 10 to test for the significance of Pearson's *r*. We begin by conducting a test of statistical significance for our example of the correlation between age and number of arrests.

Assumptions:

Level of Measurement: Interval scale.

Population Distribution: Normal distribution of Y around each value of X (must be assumed because N is not large).

Homoscedasticity.

Linearity.

Sampling Method: Independent random sampling.

Sampling Frame: Youth in one U.S. city.

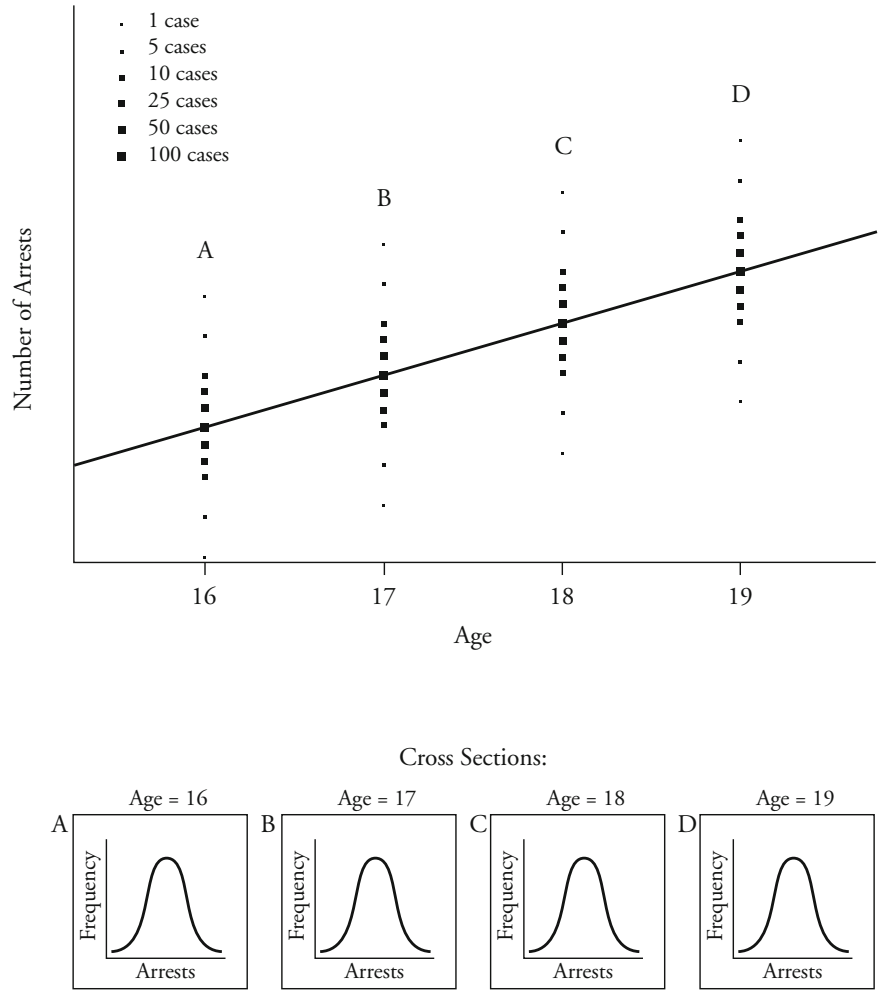
Hypotheses:

H_0: There is no linear relationship between age and number of arrests in the population of young offenders ($r_p = 0$).

H_1: There is a linear relationship between age and number of arrests in the population of young offenders ($r_p \neq 0$).

The *t*-test for Pearson's *r* assumes that the variables examined are measured on an interval scale. In practice, researchers sometimes use ordinal-scale measures for calculating these coefficients, particularly when an interval-level measure is related to an ordinal-scale variable. There is no simple answer to the question of which statistic is most appropriate in such cases, and Pearson's *r* is often considered a good solution. Nonetheless, you should keep in mind that Pearson's *r*, like other statistics that require an interval level of measurement, assumes that the categories are not only ranked but also equal to one another. When there are clear differences between the categories, the meaning of *r* becomes ambiguous. For example, suppose you were interested in the relationship between amount stolen in robberies and age, where amount stolen in robberies was measured on an ordinal scale, with the first category as $1–50, the second as $51–200, and subsequent intervals also of unequal size. If the real relationship between amount stolen and age was truly linear, with every year of age related to a measured increase in amount stolen, you would likely get a misleading correlation coefficient. In this case, an interval-scale measurement would allow you to represent the linear relationship between amount stolen and age. The ordinal scale we have described might mask or misrepresent that relationship. In practice, you should also be wary of using Pearson's correlation coefficient when the number of categories is small (for example, less than 5). While *r* is sometimes used when the researcher wants to represent the relationship between an ordinal-scale measure and an interval-level variable, it should not be used to define the relationship between two ordinal-level measures or when nominal-scale measurement is involved. As noted in earlier chapters, other statistics are more appropriate for measuring such relationships.

Figure 14.7 *Scatterplot Showing Normal Distribution and Homoscedasticity*



Because our test of the statistical significance of *r* is a parametric test of significance, we must also make assumptions regarding the population distribution. For tests of statistical significance with *r*, we must assume a normal distribution. However, in this case, it is useful to think of this distribution in terms of the joint distribution of scores between X_1 and X_2. Assume that the scatterplot in Figure 14.7 represents the relationship between age and number of arrests for ages 16–19 for the population of scores. The relationship, as in our sample, is linear. Notice how the points in the scatterplot are distributed. Suppose we put an imaginary line through the scatter of points (shown as a real line in Figure 14.7). There is a clustering of points close to the line, in the middle

of the distribution. As we move away from the center of the distribution of number of arrests for each age (represented by the imaginary line), there are fewer points. The distribution of number of arrests for each value of age is basically normal in form, as illustrated in the cross section for each of the four ages examined. This imaginary population distribution meets the normality assumption of our test.

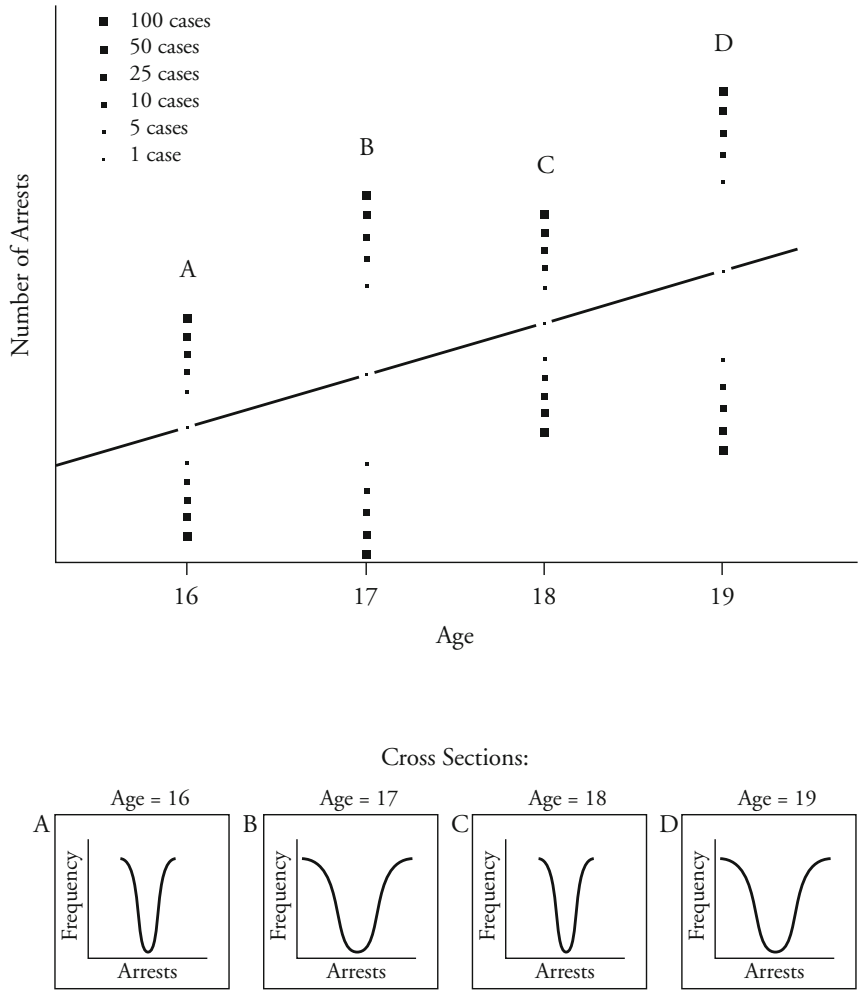
One problem in drawing conclusions about our assumptions is that they relate to the population and not to the sample. Because the population distribution is usually unknown, we generally cannot come to solid conclusions regarding our assumptions about the population. In the case of an assumption of normality, the researcher is most often aided by the central limit theorem. When the number of cases in a sample is greater than 30, the central limit theorem can be safely invoked. For our example, we cannot invoke the central limit theorem. Accordingly, our test results cannot be relied on unless the assumption of a normal distribution is true for the population to which we infer.

For our *t*-test, we must also assume that the variances of the joint distribution of scores are equal. In our example, this means that the spread of number of arrests around each value of age should be about the same. This is the assumption of homoscedasticity. To visualize this assumption, it is useful to look again at the scatterplot in Figure 14.7.

We can see that for each age examined, the variance in the distribution of scores for number of arrests is about equal. That is, the spread of the scores around our imaginary line for each value of age in this population distribution is about equal, whether we look at the cases associated with the youngest subjects (on the left side of the scatterplot), those associated with average-age subjects (in the middle of the scatterplot), or those associated with the oldest offenders (on the right side of the scatterplot). With regard to the assumption of homoscedasticity, researchers generally use the scatterplot of sample cases as an indication of the form of the population distribution. As with analysis of variance, we are generally concerned with only marked violations of the homoscedasticity assumption. Given the small number of cases in our sample distribution of scores, it is very difficult to examine the assumption of homoscedasticity. Nonetheless, if you look back at Figure 14.1, it seems reasonable to conclude that there are no major violations of this assumption.

What would a joint distribution of the relationship between arrests and age look like if both the normality and the homoscedasticity assumption were violated? In the scatterplot in Figure 14.8 (page 425), the points for each age category are not clustered in the center of the distribution of scores for number of arrests. Indeed, they form a type of bimodal distribution, with peaks above and below our imaginary line (see the cross section for each age group). **Heteroscedasticity** (or unequal variances), rather than homoscedasticity, is also represented in the scatterplot in Figure 14.8. For subjects aged 17 and 19, the scores

Figure 14.8 *Scatterplot Showing Nonnormal Distribution and Heteroscedasticity*



are scattered widely. For subjects aged 16 and 18, however, the scores are tightly clustered around the imaginary line. If this were the population distribution of the variables under study, you would want to be very cautious in applying the *t*-test to your correlation coefficient.

As with other tests of statistical significance, we assume independent random sampling. Pearson's correlation coefficient adds one new assumption to our test—that of linearity. Our null hypothesis is simply that there is no linear relationship between age and number of arrests, or that the population correlation coefficient (r_p) is equal to 0. The research hypothesis is nondirectional; that is, we simply test for a linear relationship

between age and number of arrests (i.e., $r_p \neq 0$). However, we might have proposed a directional research hypothesis—for example, that there is a positive relationship between age and number of arrests in the population, or that $r_p > 0$.

The Sampling Distribution The sampling distribution is *t*, with $N - 2$ degrees of freedom. For our example, df = 15 − 2 = 13.

Significance Level and Rejection Region With a two-tailed 0.05 significance threshold, the critical value for the *t*-test (with 13 degrees of freedom) is 2.160 (see Appendix 4). We will reject the null hypothesis if the *t*-score is greater than 2.160 or less than −2.160. In these cases, the observed significance level of the test will be less than the criterion significance level we have chosen.

The Test Statistic It is important to note that there is more than one way to test statistical significance for *r*. Equation 14.4 provides a straightforward estimate of *t*, based on our calculation of *r*.

$$t = r\sqrt{\frac{N-2}{1-r^2}} \qquad \textbf{Equation 14.4}$$

Inserting our sample estimates, we calculate that the *t*-statistic for *r* is 4.4188:

Working It Out

$$t = r\sqrt{\frac{N-2}{1-r^2}}$$
$$= 0.7748\sqrt{\frac{15-2}{1-(0.7748)^2}}$$
$$= 0.7748\sqrt{32.5256}$$
$$= 4.4188$$

The Decision Since 4.419 is greater than our critical value of *t* (2.160), we reject the null hypothesis and conclude that there is a statistically significant relationship between age and number of arrests. However, because we cannot strongly support the assumption of normality in this test or relax that assumption because *N* is large, we cannot place strong reliance on our test result.

Statistical Significance of *r*: Unemployment and Crime in California

In our analysis of unemployment and burglary rates in California counties, we found $r = 0.4910$. We can test the statistical significance of this result by following the same approach we used in the previous example. We start by outlining our assumptions and hypotheses.

Assumptions:

Level of Measurement: Interval scale.

Population Distribution: Normal distribution of Y (i.e., burglary rate) around each value of X (i.e., unemployment rate) (relaxed because N is large).

Homoscedasticity.

Sampling Method: Independent random sampling (the 58 counties represent all counties in California in 1999).

Sampling Frame: California counties.

Linearity.[3]

Hypotheses:

H_0: There is no linear relationship between unemployment rate and burglary rate ($r_p = 0$).
H_1: There is a linear relationship between unemployment rate and burglary rate ($r_p \neq 0$).

Since we have data for all counties in California for a given year, you might question why we would choose to conduct a statistical test of significance. Why do we need to make inferences? We already have the population of scores. One reason might be that we want to look at data for the year observed as a sample of the relationships that occur over a number of years. Similarly, we might want to use the data in California to represent other states. For either of these inferences, we would need to explain why this sample was representative of the population. Another reason we might choose to conduct a statistical test of significance is to see whether the correlation observed is likely to be a chance occurrence. We would expect differences across the counties simply as a product of the natural fluctuations that occur in statistics. A significance test in this case can tell us whether the relationship observed is likely to be the result of such chance fluctuations or whether it is likely to represent a real relationship between the measures examined.

---

[3]It is good practice to examine the sample scatterplot of scores to assess whether this assumption is likely to be violated. We find no reason to suspect a violation of the assumption when we examine this scatterplot (see Chapter 15, Figure 15.2).

The Sampling Distribution The sampling distribution is *t*, with $N - 2$ degrees of freedom. For this example, df = 58 − 2 = 56.

Significance Level and Rejection Region With a two-tailed 0.05 significance level and 56 degrees of freedom, interpolation yields estimates of +2.003 and −2.003 for the critical values of the *t*-test.[4]

The Test Statistic We again use Equation 14.4 to calculate a *t*-value to test the significance of *r*. Inserting the values for our data, we find the value of the *t*-statistic to be 4.2177:

$$t = r\sqrt{\frac{N-2}{1-r^2}}$$

$$= 0.491\sqrt{\frac{58-2}{1-(0.491)^2}}$$

$$= 0.491\sqrt{73.789165}$$

$$= 4.2177$$

The Decision Since 4.218 is greater than our critical value of 2.003, we reject the null hypothesis and conclude that there is a statistically significant relationship between the unemployment rate and the burglary rate.

Testing the Statistical Significance of Spearman's *r*

For Spearman's *r*, we use a nonparametric statistical test. With $N \leq 30$, we use an exact probability distribution constructed for the distribution of differences between ranked pairs (see Appendix 7). For larger samples, a normal approximation of this test is appropriate. It is constructed by taking the difference between the observed value of r_s and the parameter value under the null hypothesis ($r_{s(p)}$). This value is then divided by 1 divided by the square root of $N - 1$, as shown in Equation 14.5.

$$z = \frac{r_s - r_{s(p)}}{\frac{1}{\sqrt{N-1}}} \qquad \textbf{Equation 14.5}$$

[4]The table does not list a *t*-value for df = 56. We therefore interpolate from the values of df = 55 (2.004) and df = 60 (2.000).

Because we are examining less than 15 cases, we will use the exact probability table presented in Appendix 7.

Assumptions:

Level of Measurement: Ordinal scale.

Population Distribution: No assumption made.

Sampling Method: Independent random sampling.

Sampling Frame: Youth in one U.S. city.

Hypotheses:

H_0: There is no linear relationship between the rank order of scores in the population ($r_{s(p)} = 0$).

H_1: There is a linear relationship between the rank order of scores in the population ($r_{s(p)} \neq 0$).

Because we use a nonparametric test, we do not need to make assumptions regarding the population distribution. The null hypothesis is the same as for the correlation coefficient r; however, it is concerned with ranks rather than raw scores.

The Sampling Distribution Because N is small, we use the exact probability distribution constructed for Spearman's r in Appendix 7.

Significance Level and Critical Region As earlier, we use the conventional 0.05 significance threshold. Since our research hypothesis is not directional, we use a two-tailed rejection region. From Appendix 7, under a two-tailed 0.05 probability value and an N of 15, we find that an r_s greater than or equal to 0.525 or less than or equal to -0.525 is needed to reject the null hypothesis.

The Test Statistic In the case of the exact probability distribution, the test statistic is simply the value of r_s. As calculated earlier in this chapter (see page 420), r_s equals 0.825.

The Decision Because the observed r_s is larger than 0.525, we reject the null hypothesis and conclude that there is a statistically significant linear relationship between ranks of age and number of arrests in the population. The observed significance level of our test is less than the criterion significance level we set at the outset ($p < 0.05$).

Chapter Summary

Linear correlation coefficients describe the relationship between two interval-level measures, telling us how strongly the two are associated. **Pearson's *r*** is a widely used linear correlation coefficient. It examines the placement of subjects on both variables relative to the mean and estimates how strongly the scores move together or in opposite directions relative to the mean. The **covariation,** which is the numerator of the Pearson's *r* equation, is positive when both scores vary in the same direction relative to the mean and negative when they vary in opposite directions. Dividing the covariation by the denominator of the Pearson's *r* equation serves to standardize the coefficient so that it varies between −1 and +1. Pearson's *r* will produce a misleading correlation coefficient if there is a nonlinear relationship between the variables.

Outliers have a strong impact on Pearson's *r*. If there are several outliers that follow the general pattern of relationships in the data, **Spearman's *r*** may provide less misleading results. Spearman's *r* also varies between −1 and +1. It compares the rank order of subjects on each measure.

The *t* distribution may be used to test significance for Pearson's correlation coefficient. It is assumed that the variables examined are measured on an interval scale. There is also an assumption of normality and a requirement of homoscedasticity. These assumptions relate to the joint distribution of X_1 and X_2. The researcher must also assume linearity. For Spearman's *r*, a nonparametric test of statistical significance is used.

Key Terms

covariation A measure of the extent to which two variables vary together relative to their respective means. The covariation between the two variables serves as the numerator for the equation to calculate Pearson's *r*.

curvilinear relationship An association between two variables whose values may be represented as a curved line when plotted on a scatter diagram.

heteroscedasticity A situation in which the variances of scores on two or more variables are not equal. Heteroscedasticity violates one of the assumptions of the parametric test of statistical significance for the correlation coefficient.

linear relationship An association between two variables whose joint distribution may be represented in linear form when plotted on a scatter diagram.

Pearson's correlation coefficient See *Pearson's r.*

Pearson's *r* A commonly used measure of association between two variables. Pearson's *r* measures the strength and direction of linear relationships on a standardized scale from −1.0 to 1.0.

scatter diagram See *scatterplot.*

scatterplot A graph whose two axes are defined by two variables and upon which a point is plotted for each subject in a sample according to its score on the two variables.

Spearman's correlation coefficient See *Spearman's r.*

Spearman's *r* (r_s) A measure of association between two rank-ordered variables. Spearman's *r* measures the strength and direction of linear relationships on a standardized scale between −1.0 and 1.0.

Symbols and Formulas

r Pearson's correlation coefficient

r_s Spearman's correlation coefficient

D Difference in rank of a subject on two variables

To calculate the covariation of scores for two variables:

$$\text{Covariation of scores} = \sum_{i=1}^{N}(X_{1i} - \overline{X}_1)(X_{2i} - \overline{X}_2)$$

To calculate Pearson's correlation coefficient:

$$\text{Pearson's } r = \frac{\sum_{i=1}^{N}(X_{1i} - \overline{X}_1)(X_{2i} - \overline{X}_2)}{\sqrt{\left(\sum_{i=1}^{N}(X_{1i} - \overline{X}_1)^2\right)\left(\sum_{i=1}^{N}(X_{2i} - \overline{X}_2)^2\right)}}$$

To calculate Spearman's correlation coefficient:

$$r_s = 1 - \frac{6\sum_{i=1}^{N} D_i^2}{N(N^2 - 1)}$$

To test statistical significance for Pearson's r:

$$t = r\sqrt{\frac{N - 2}{1 - r^2}}$$

To test statistical significance for Spearman's r where N is large:

$$z = \frac{r_s - r_{s(p)}}{\frac{1}{\sqrt{N - 1}}}$$

Exercises

14.1 A researcher draws four random samples of ten offenders, aged between 30 and 35 years, all of whom are currently serving out a term of imprisonment and all of whom have been in prison before. For each sample, she compares the subjects on the following pairs of variables:

Sample 1	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
X_1: Number of convictions	3	5	1	7	6	2	4	9	10	8
X_2: Average sentence	2	2.5	0.5	3	3	1	2	4.5	5	3.5

Sample 2	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
X_1: Years of education	9	12	17	16	9	14	10	17	17	9
X_2: Age at first offense	14	17	14	16	10	17	16	10	12	12

Sample 3	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
X_1: Age at first offense	13	17	10	16	14	11	18	19	15	12
X_2: Number of convictions	7	3	10	4	6	9	1	1	6	8

Sample 4	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
X_1: Age at first offense	11	16	18	12	15	17	13	20	20	13
X_2: Average sentence	3	5	1.5	1	1	4	4.5	5	3	2.5

a. Calculate the mean scores of both variables for Samples 1, 2, 3, and 4.

b. Display the data for the four samples in four frequency distribution tables. For each score, add a positive or negative sign to indicate the direction in which the score differs from the mean (as done in Tables 14.4 and 14.5). Add an extra column in which you record a plus or a minus for the product of the two signs.

c. Draw four scatterplots, one for each sample distribution. State whether each scatterplot shows a positive relationship, a negative relationship, a curvilinear relationship, or no relationship between the two variables.

d. Would you advise against using Pearson's correlation coefficient as a measure of association for any of the four samples? Explain your answer.

14.2 Jeremy, a police researcher, is concerned that police officers may not be assigned to areas where they are needed. He wishes to find out whether there is a connection between the number of police officers assigned to a particular block and the number of violent incidents reported on that block during the preceding week. For ten different blocks (designated A

through J), the number of patrolling officers assigned and the number of prior violent incidents reported are as follows:

	A	B	C	D	E	F	G	H	I	J
X_1: Violent incidents	7	10	3	9	8	0	4	4	2	8
X_2: Officers assigned	6	9	3	10	8	1	4	5	2	7

a. Calculate the covariance for the data recorded above.

b. Calculate the value of Pearson's *r* for the data recorded above.

c. On an 11th block—block K—there are no police officers patrolling, yet in the previous week 11 violent incidents were reported there. What effect would it have on Pearson's *r* if Jeremy included block K in his calculations?

d. How do you explain this difference?

14.3 Seven subjects of different ages are asked to complete a questionnaire measuring attitudes about criminal behavior. Their answers are coded into an index, with scores ranging from 1 to 15. The subjects' scores are as follows:

X_1: Age	12	22	10	14	18	20	16
X_2: Score	6	3	3	9	9	6	13

a. Calculate Pearson's correlation coefficient for the two variables listed above.

b. Illustrate the sample distribution on a scatterplot.

c. Divide the scatterplot into two sections, and calculate the value of Pearson's *r* for each section.

d. Explain the difference between the *r* values you obtained in parts a and c.

14.4 Eight homeowners in the inner-city neighborhood of Moss Tide are asked how long they have been living in the neighborhood and how many times during that period their house has been burglarized. The results for the eight subjects are listed below:

X_1: Years in neighborhood	2	1.5	3.5	28	1	5	20	3
X_2: Number of burglaries	2	1	5	55	0	4	10	3

a. Calculate Pearson's *r* for the two variables recorded above.

b. Calculate Spearman's *r* for the same data.

c. Illustrate the sample distribution on a scatterplot.

d. Which of the two correlation coefficients is more appropriate, in your opinion, for this case? Refer to the scatterplot in explaining your answer.

14.5 Eleven defendants arrested for violent offenses were all required to post bail. The judge said that the amount of bail assigned was related to the total number of prior arrests. The results for the 11 defendants are as follows:

X_1: Number of prior arrests	0	3	9	13	2	7	1	4	7	20	5
X_2: Amount of bail assigned	100	500	2,500	10,000	1,000	10,000	100	7,500	5,000	100,000	4,000

a. Calculate Pearson's r for the two variables recorded above.

b. Calculate Spearman's r for the same data.

c. Illustrate the sample distribution on a scatterplot.

d. Which of the two correlation coefficients is more appropriate, in your opinion, for this case? Refer to the scatterplot in explaining your answer.

14.6 Researchers looking at age and lifetime assault victimization interviewed nine adults and found the following values:

X_1: Age	18	20	19	25	44	23	67	51	33
X_2: Number of times assaulted in lifetime	1	4	8	0	6	2	9	3	10

a. Calculate Pearson's r for the two variables recorded above.

b. Calculate Spearman's r for the same data.

c. Illustrate the sample distribution on a scatterplot.

d. Which of the two correlation coefficients is more appropriate, in your opinion, for this case? Refer to the scatterplot in explaining your answer.

14.7 In a study looking at the relationship between truancy and theft, a sample of ten youth were asked how many times in the last year they had skipped school and how many times they had stolen something worth $20 or less. Their responses were

X_1: Number of times skipped school	9	2	4	0	0	10	6	5	3	1
X_2: Number of thefts valued at $20 or less	25	10	13	0	2	24	31	20	1	7

a. Calculate Pearson's r for the two variables recorded above.

b. Use a 5% level of significance and outline each of the steps required in a test of statistical significance of r.

14.8 A study investigating the link between child poverty and property crime rates gathered information on a random sample of 13 counties. The values for the percentage of children under 18 living in poverty and property crime rates (given per 100,000) are

X_1: Percentage of children living in poverty	10	8	43	11	27	18	15	22	17	17	20	25	35
X_2: Property crime rate	1,000	2,000	7,000	4,000	3,000	4,500	2,100	1,600	2,700	1,400	3,200	4,800	6,300

a. Calculate Pearson's r for the two variables recorded above.

b. Use a 5% level of significance and outline each of the steps required in a test of statistical significance of r.

Computer Exercises

The process for obtaining correlation coefficients and scatterplots is similar in SPSS and Stata. Similar to other chapters, we have made available example syntax files that illustrate the commands below for both SPSS (Chapter_14.sps) and Stata (Chapter_14.do).

SPSS

Correlation Coefficients

Both correlation coefficients discussed in this chapter—Pearson's r and Spearman's r—can be obtained with similar commands. To obtain Pearson's r, use the command CORRELATIONS:

```
CORRELATIONS
  /VARIABLES=list_of_variable_names
  /STATISTICS DESCRIPTIVES.
```

The /STATISTICS DESCRIPTIVES line is optional and would simply generate a table of descriptive statistics, much like that produced with the DESCRIPTIVES command in Chapter 4. Although it is much more efficient to have SPSS compute all the correlations simultaneously for however many variables you have included in the command, rather than selectively picking out specific pairs that you might be most interested in, bear in mind that the output may be rather unwieldy if too many variables are listed in the /VARIABLES= line.

The output from the CORRELATIONS command will consist of a matrix (grid) of correlations for all the variables whose names appear in the /VARIABLES= line. You should also note that this matrix of correlations is sym-

metric; running from the upper left to the lower right corner of the matrix is a diagonal that is made up of 1s (the correlation of the variable with itself). The correlations that appear above the diagonal will be a mirror image of the correlations that appear below the diagonal. Thus, to locate the value of the correlation coefficient you are most interested in, you simply find the row that corresponds to one of the variables and the column that corresponds to the other variable. It does not matter which variable you select for the row or for the column; the correlation coefficient reported in the matrix will be the same.

The command for obtaining Spearman's *r* is NONPAR CORR (nonparametric correlation in SPSS):

```
NONPAR CORR
  /VARIABLES = list_of_variable_names
  /PRINT = SPEARMAN.
```

The output will again be a matrix of correlations for the variables listed on the /VARIABLES= line of the command. The /PRINT = SPEARMAN option ensures that SPSS will compute the correct correlation.

Scatterplots

The Computer Exercises section of Chapter 3 illustrated a wide range of graphics commands available in SPSS. One of the simple, but powerful, graphs that SPSS can produce is a scatterplot, which is obtained with the GRAPH command and /SCATTERPLOT(BIVAR) option:

```
GRAPH
/SCATTERPLOT(BIVAR) = x_axis_var WITH y_axis_var.
```

The graph produced by this command will be a simple scatterplot that can be edited to suit your needs. As we noted in Chapter 3, the interactive graphical options in SPSS are extensive, and we encourage you to explore some of the additional possibilities for creating novel and informative scatterplots.

Stata

Correlation Coefficients

Obtaining Pearson's *r* in Stata is also simple using the **pwcorr** command (pairwise correlation in Stata):

```
pwcorr list_of_variable_names, obs sig
```

where the options **obs** and **sig** will print in the correlation matrix the number of observations used in the calculation of each correlation and the observed significance level, making the output comparable to that in SPSS. The **pwcorr** command does not require these options; they are included here simply to provide more comprehensive and informative output.

There is one notable difference in the Stata output compared to that in SPSS. Rather than printing a full symmetric matrix of correlations (see previous discussion of SPSS output), Stata prints only the diagonal (where $r = 1.0$ and the variable is correlated with itself) and the correlations appearing below the diagonal (i.e., lower left half of the correlation matrix). This makes the output less overwhelming and easier to sort through a large number of correlations.

Spearman's *r* is obtained with the **spearman** command:

spearman list_of_variable_names, **stats(rho obs p) pw**

where the **stats(rho obs p)** option will print the correlation (**rho**), the number of observations used in calculating the correlation (**obs**), and the observed significance level of the correlation (**p**). The default is to simply print the value of Spearman's r without information about the number of observations or observed significance level. The **pw** option ensures that Stata uses all available cases in computing the correlation rather than deleting an observation if it is missing information on any one of the variables included in the list of variables.

The output will be a matrix of Spearman's *r* correlations.

Scatterplots

In Chapter 3's Computer Exercises section we illustrated the use of the **twoway** command for the creation of other graphs. It can also be used for the creation of simple scatterplots by changing the graph-type option to **scatter**:

twoway (**scatter** y_var_name x_var_name)

As with all other graphs in Stata, this scatterplot may be edited and customized to suit your purpose.

Problems

1. Open the California UCR data file (caucr_ 99.sav or caucr_99.dta). These are the data presented in Table 14.8. Compute Pearson's *r* for these data and note that it matches the value reported in the text.
   a. Compute Spearman's *r* for these data.
   b. How does the value of Pearson's *r* compare to that for Spearman's r? What might account for this?
   c. Generate a scatterplot for these data.
2. Enter the data from Exercise 14.2.
   a. Compute both Pearson's *r* and Spearman's *r* for these data.
   b. Add the extra case presented in part c of Exercise 14.2. Recompute Pearson's *r* and Spearman's *r* for these data.
   c. Compare your answers from parts a and b. How do you explain this pattern?
   d. Generate a scatterplot for the 11 cases.

3. Enter the data from Exercise 14.8.
   a. Compute Spearman's *r* for these data.
   b. How does the value of Pearson's *r* compare to that for Spearman's *r*? What might account for this?
   c. Generate a scatterplot for these data.
4. Open the NYS data file (nys_1.sav, nys_1_student.sav, or nys_1.dta). Use a 5% level of significance, and outline each of the steps required in a test of statistical significance for each of the following relationships:
   a. Age and number of thefts valued at less than $5 in the last year.
   b. Number of times drunk and number of thefts valued at $5 to $50 in the last year.
   c. Frequency of marijuana use and number of times the youth has hit other students in the last year.
   d. Number of times the youth has hit a parent and number of thefts valued at more than $50 in the last year.
   e. Number of times the youth has been beaten up by parent and number of times the youth has hit a teacher in the last year.
5. Generate a scatterplot for each pair of variables listed in Computer Exercise 4.
6. Open the Pennsylvania Sentencing data file (pcs_ 98.sav or pcs_98.dta). Use a 5% level of significance, and outline each of the steps required in a test of statistical significance for each of the following relationships:
   a. Age of offender and length of incarceration sentence.
   b. Severity of conviction offense and length of incarceration sentence.
   c. Prior criminal history score and age.
   d. Length of incarceration sentence and prior criminal history score.
7. Generate a scatterplot for each pair of variables listed in Computer Exercise 6.