Chapter 10
Correlation

Learning Objectives

10.1 Examine the strength of the correlation between two variables by using a scatter plot

10.2 Distinguish between positive and negative correlation

10.3 Classify as curvilinear a correlation where one variable increases as the other variable increases until the relationship reverses itself

10.4 Explain how correlation coefficients express both the strength and direction of straight-line correlation

10.5 Demonstrate how the Pearson's correlation coefficient is used to test the strength and direction of the relationship between variables measured at the interval level

10.6 Examine the necessity to properly study a scatter plot before jumping to conclusions about its implications

10.7 Explain how the introduction of a control variable in partial correlation allows researchers to control a two-variable relationship for the impact of a third variable

Introduction

Characteristics, educational attainment, vary from one person to another and are therefore referred to as **variables**. In earlier chapters, we have been concerned with establishing the presence or absence of a relationship between any two variables, which we will now label X and Y; for example, between age (X) and frequency of Internet use (Y), between intelligence (X) and achievement-motivation (Y), or between educational attainment (X) and income (Y). Aided by the t ratio, analysis of variance, or nonparametric tests, such as chi-square, we previously sought to discover whether a difference between two or more samples could be regarded as statistically significant—reflective of a true population difference—and not merely the product of sampling error. Now, we turn our attention to the existence and strength of relationship between two variables—that is, on their co-relationship or, simply, their correlation.

10.1: Strength of Correlation

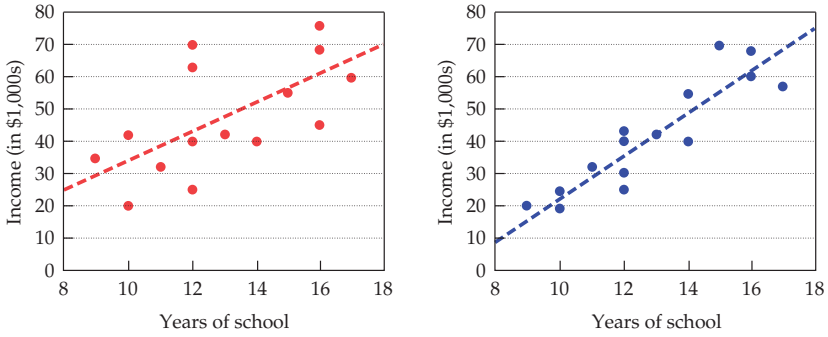
Objective: **Examine the strength of the correlation between two variables by using a scatter plot**

Finding that a relationship exists does not indicate much about the degree of association, or **correlation**, between two variables. Many relationships are statistically significant—that is, stronger than you would expect to obtain just as a result of sampling error alone, yet rather few express *perfect* correlation. To illustrate, we know that height and weight are associated, since the taller a person is, the more he or she tends to weigh. There are numerous exceptions to the rule, however. Some tall people weigh very little; some short people weigh a lot. In the same way, a relationship between age and net worth does not preclude the possibility of finding many young adults who have accumulated a greater net worth in just a few years than some older adults have over decades.

Correlations actually vary with respect to their **strength**. We can visualize differences in the strength of correlation by means of a **scatter plot** or *scatter diagram*, a graph that shows the way scores on any two variables, X and Y, are scattered throughout the range of possible score values. In the conventional arrangement, a scatter plot is set up so that the X variable is located along the horizontal base line, and the Y variable is measured on the vertical line.

Turning to Figure 10.1, we find two scatter plots, each representing the relationship between years of education (X) and income (Y). Figure 10.1(a) depicts this relationship

Figure 10.1 Strength of Relationships



(a) Males **(b)** Females

for males, and Figure 10.1(b) represents the relationship for females. Note that every point in these scatter plots depicts two scores, education and income, obtained by one respondent. In Figure 10.1(a), for example, we see that a male having 9 years of education earned just about $35,000, whereas a male with 17 years of education made about $60,000.

We can say that the strength of the correlation between *X* and *Y* increases as the points in a scatter plot more closely form an imaginary diagonal line through the center of the scatter of points in the graph. Although not necessarily included in a scatter plot, we have drawn these lines with light gray dashes to illustrate how the strength of correlation translates to closeness of points to a straight line. Actually, we will discuss this line in great detail in Chapter 11.

As should be clear, Figure 10.1(a) represents a weaker correlation than does Figure 10.1(b), although both scatter plots indicate that income tends to increase with greater education. Such data would indeed support the view that the income of women (relative to that of men) is more related to the level of education they attain. Alternatively, this suggests that income levels for men are more a result of factors other than education than is the case for women.

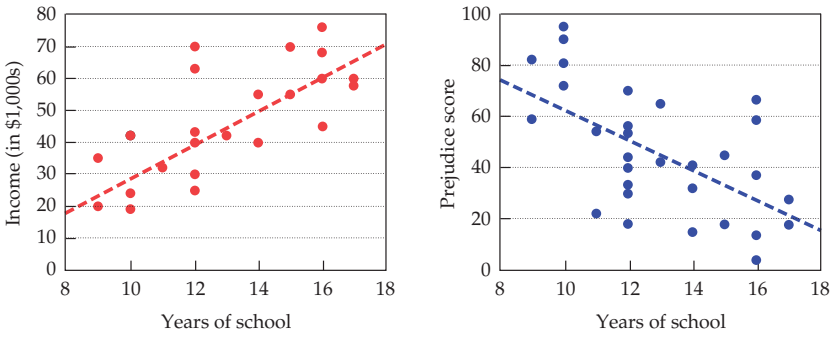
10.2: Direction of Correlation

Objective: **Distinguish between positive and negative correlation**

Correlation can often be described with respect to direction as either positive or negative. A *positive correlation* indicates that respondents getting *high* scores on the *X* variable also tend to get *high* scores on the *Y* variable. Conversely, respondents who get *low* scores on *X* also tend to get *low* scores on *Y*. Positive correlation can be illustrated by the relationship between education and income. As we have previously seen, respondents completing many years of school tend to make large annual incomes, whereas those who complete only a few years of school tend to earn very little annually. The overall relationship between education and income for males and females combined is shown in Figure 10.2(a), again with an imaginary straight line through the scatter of points drawn to show the direction of the relationship.

A *negative correlation* exists if respondents who obtain *high* scores on the *X* variable tend to obtain *low* scores on the *Y* variable. Conversely, respondents achieving *low* scores on *X* tend to achieve *high* scores on *Y*. The relationship between education and income would certainly *not* represent a negative correlation, because respondents completing many

Figure 10.2 Direction of Relationship for Straight-line Correlations



(a) Years of School and Income **(b)** Years of School and Prejudice

years of school *do not* tend to make small annual incomes. A more likely example of negative correlation is the relationship between education and the degree of prejudice against minority groups. Prejudice tends to diminish as the level of education increases. Therefore, as shown in Figure 10.2(b), individuals having little formal education tend to hold strong prejudices, whereas individuals completing many years of education tend to be low with respect to prejudice.

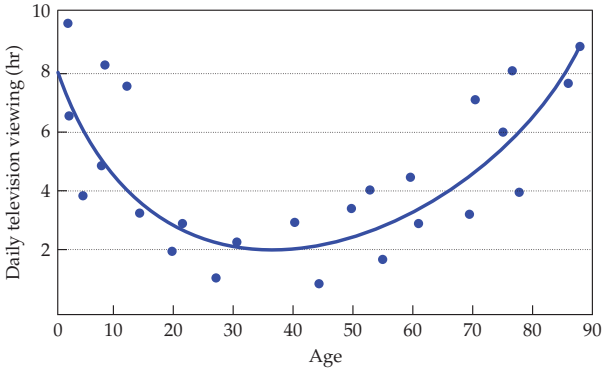
A positive or negative correlation represents a type of *straight-line* relationship. Depicted graphically, the points in a scatter plot tend to form an imaginary straight line through the center of the points in the graph. If a positive correlation exists, then the points in the scatter plot will cluster around an imaginary straight line like the one drawn in Figure 10.2(a). In contrast, if a negative correlation is present, the points in the scatter plot will surround an imaginary straight line like the one shown in Figure 10.2(b).

10.3: Curvilinear Correlation

Objective: **Classify as curvilinear a correlation where one variable increases as the other variable increases until the relationship reverses itself**

For the most part, social researchers seek to establish a straight-line correlation, whether positive or negative. It is important to note, however, that not all relationships between X and Y can be regarded as forming a straight line. There are many **curvilinear** relationships, indicating, for example, that one variable increases as the other variable increases until the relationship reverses itself, so that one variable finally decreases while the other continues to increase. That is, a relationship between X and Y that begins as positive becomes negative; or a relationship that starts as negative becomes positive. To illustrate a curvilinear correlation, consider the relationship between age and hours of television viewing. As shown in Figure 10.3, the points in the scatter plot tend to form a U-shaped curve rather than a straight line. Thus, television viewing tends to decrease with age, until the thirties, after which viewing tends to increase with age.

Figure 10.3 The Curvilinear Relationship between Age (X) and Television Viewing (Y)



10.4: The Correlation Coefficient

Objective: **Explain how correlation coefficients express both the strength and direction of straight-line correlation**

The process for finding curvilinear correlation lies beyond the scope of this text. Instead, we turn our attention to **correlation coefficients**, which numerically express both strength and direction of straight-line correlation. Such correlation coefficients range between −1.00 and +1.00 as follows:

−1.00 ← perfect negative correlation
⋮
−.60 ← strong negative correlation
⋮
−.30 ← moderate negative correlation
⋮
−.10 ← weak negative correlation
⋮
.00 ← no correlation
⋮
+.10 ← weak positive correlation
⋮
+.30 ← moderate positive correlation
⋮
+.60 ← strong positive correlation
⋮
+1.00 ← perfect positive correlation

We see, then, that negative numerical values, such as −1.00, −.60, −.30, and −.10, signify negative correlation, whereas positive numerical values, such as +1.00, +.60, +.30, and +.10, indicate positive correlation. Regarding degree of association, the closer to 1.00 in either direction, the greater the strength of the correlation. Because the strength of a correlation is independent of its direction, we can say that −.10 and +.10 are equal in strength (both are weak); and −.80 and +.80 have equal strength (both are strong).

10.5: Pearson's Correlation Coefficient

Objective: **Demonstrate how the Pearson's correlation coefficient is used to test the strength and direction of the relationship between variables measured at the interval level**

With the aid of **Pearson's correlation coefficient** (r), we can determine the strength and the direction of the relationship between X and Y variables, both of which have been measured at the interval level. For example, we might be

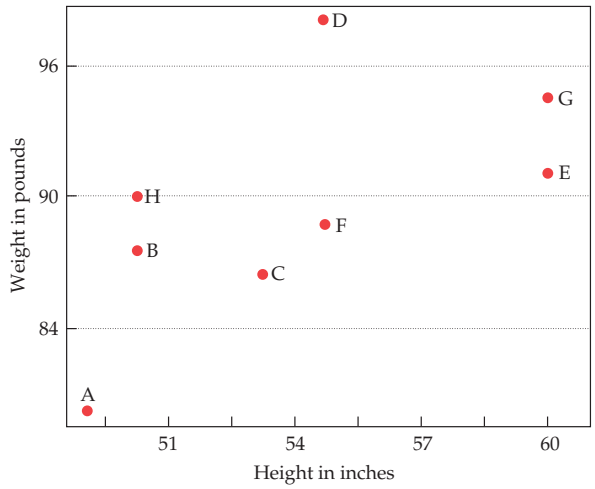
interested in examining the association between height and weight for the following sample of eight children:

Table 10.1 Heights and Weights of a Sample of Eight Children

Child	Height (in.) (X)	Weight (lb) (Y)
A	49	81
B	50	88
C	53	87
D	55	99
E	60	91
F	55	89
G	60	95
H	50	90

In the scatter plot in Figure 10.4, the positive association that one would anticipate between height (X) and weight (Y) in fact appears. But note that there are some exceptions. Child C is taller but weighs less than child H; child D is shorter but weighs more than child E. These exceptions should not surprise us because the relationship between height and weight is not perfect. Nonetheless, the general rule that "the taller one is the heavier one is" holds true: F is taller and heavier than A, as is G compared to H.

Figure 10.4 Scatter Plot of Height and Weight

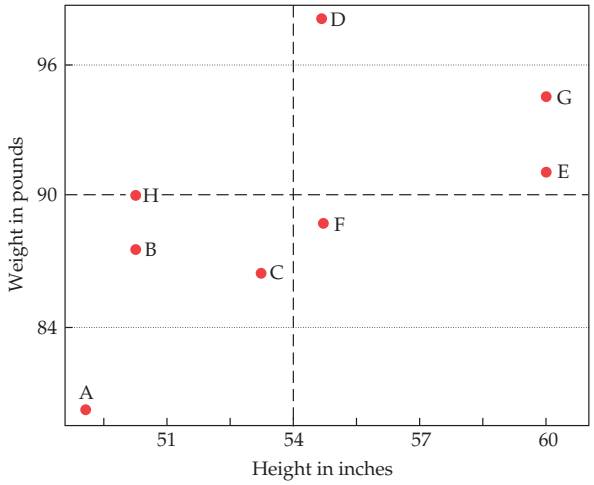


Pearson's r does more than just consider if subjects are simply taller or heavier than other subjects; it considers precisely how much heavier and how much taller. The quantity that Pearson's r focuses on is the product of the X and Y deviations from their respective means. The deviation $(X - \overline{X})$ tells how much taller or shorter than average a particular child is; the deviation $(Y - \overline{Y})$ tells how much heavier or lighter than average a particular child is.

With Pearson's r, we add the products of the deviations to see if the positive products or negative products are more abundant and sizable. Positive products indicate cases in which the variables go in the same direction (that is, both taller and heavier than average or both shorter and lighter than average); negative products indicate cases in which the variables go in opposite directions (that is, taller but lighter than average or shorter but heavier than average).

In Figure 10.5, dashed lines are added to the scatter plot of X and Y to indicate the location of the mean height ($\overline{X} = 54$ inches) and the mean weight ($\overline{Y} = 90$ pounds). Child G is apparently much taller and much heavier than average. His deviations on the two variables are $(X - \overline{X}) = 60 - 54 = 6$ (inches) and $(Y - \overline{Y}) = 95 - 90 = 5$ (pounds), which when multiplied yield $+30$. Child A is much shorter and much lighter than average. Her deviations (-5 and -9) multiply to $+45$. On the other hand, child C is only slightly shorter and lighter than average; her product of deviations $(-1 \times -3 = 3)$ is far less dramatic. This is as it would seem intuitively: The more dramatically a child demonstrates the rule "the taller, the heavier," the larger the product of the X and Y deviations. Finally, child F is a slight exception: He is a little taller than average yet lighter than average. As a result, his $+1$ deviation on X and -1 deviation on Y produce a negative product (-1).

Figure 10.5 Scatter Plot of Height and Weight with Mean Axes



We can compute the sum of the products for the data shown in Table 10.2. Columns 2 and 3 reproduce the heights and weights for the eight children in the sample. Columns 4 and 5 give the deviations from the means for the X and Y values. In column 6, these deviations are multiplied and then summed.

The sum of the final column (denoted SP, for sum of products) is positive—indicating a positive association between X and Y. But, as we have learned, correlation coefficients are constrained to range from -1 to $+1$ to aid in their interpretation. The formula for r accomplishes this by

Table 10.2 Sum of Products for the Data in Table 10.1

Child	X	Y	$(X - \bar{X})$	$(Y - \bar{Y})$	$(X - \bar{X})(Y - \bar{Y})$	
A	49	81	−5	−9	45	$N = 8$
B	50	88	−4	−2	8	$\bar{X} = 54$
C	53	87	−1	−3	3	$\bar{Y} = 90$
D	55	99	1	9	9	
E	60	91	6	1	6	
F	55	89	1	−1	−1	
G	60	95	6	5	30	
H	50	90	−4	0	0	
	$\Sigma X = 432$	$\Sigma Y = 720$			SP = 100	

dividing the SP value by the square root of the product of the sum of squares of both variables (SS_X and SS_Y). Thus, we need to add two more columns to our table in which we square and sum the deviations for X and for Y (see Table 10.3).

The formula for Pearson's correlation is as follows:

$$r = \frac{\Sigma (X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma (X - \bar{X})^2 \Sigma (Y - \bar{Y})^2}} = \frac{SP}{\sqrt{SS_X SS_Y}}$$

$$= \frac{100}{\sqrt{(132)(202)}}$$

$$= \frac{100}{\sqrt{26{,}664}}$$

$$= \frac{100}{\sqrt{163.3}}$$

$$= +.61$$

Therefore, Pearson's correlation indicates, as suggested by the scatter plot, that height and weight are fairly strongly correlated in the positive direction.

10.5.1: A Computational Formula for Pearson's *r*

Computing Pearson's r from deviations helps relate the topic of correlation to our earlier discussion. However, the previous formula for Pearson's r requires lengthy and time-consuming calculations. Fortunately, there is an alternative formula for Pearson's r that works directly with raw scores, thereby eliminating the need to obtain deviations for the X and Y variables. Similar to the computational formulas for variance and standard deviation we encountered earlier in the course, there are raw-score formulas for SP, SS_X, and SS_Y.

$$SP = \Sigma XY - N\bar{X}\bar{Y}$$

$$SS_X = \Sigma X^2 - N\bar{X}^2$$

$$SS_Y = \Sigma Y^2 - N\bar{Y}^2$$

Using these expressions in our formula for Pearson's correlation, we obtain the following computational formula for r:

$$r = \frac{\Sigma XY - N\bar{X}\bar{Y}}{\sqrt{(\Sigma X^2 - N\bar{X}^2)(\Sigma Y^2 - N\bar{Y}^2)}}$$

To illustrate the use of Pearson's r computational formula, consider the following data on the number of years of school completed by the father (X) and the number of years of school completed by the child (Y). To apply our formula, we must obtain the sums of X and Y (to calculate the means) and of X^2, Y^2, and XY:

X	Y	X^2	Y^2	XY	
12	12	144	144	144	$N = 8$
10	8	100	64	80	$\Sigma X = 84$
6	12	36	144	72	$\Sigma Y = 92$
16	11	256	121	176	$\bar{X} = \frac{\Sigma X}{N} = \frac{84}{8} = 10.5$
8	10	64	100	80	
9	8	81	64	72	$\bar{Y} = \frac{\Sigma Y}{N} = \frac{92}{8} = 11.5$
12	16	144	256	192	$\Sigma X^2 = 946$
11	15	121	225	165	$\Sigma Y^2 = 1{,}118$
84	92	946	1,118	981	$\Sigma XY = 981$

Table 10.3 Sum of Products and Sum of Squares of Deviations for the Data in Table 10.1

X	Y	$(X - \bar{X})$	$(Y - \bar{Y})$	$(X - \bar{X})(Y - \bar{Y})$	$(X - \bar{X})^2$	$(Y - \bar{Y})^2$
49	81	−5	−9	45	25	81
50	88	−4	−2	8	16	4
53	87	−1	−3	3	1	9
55	99	1	9	9	1	81
60	91	6	1	6	36	1
55	89	1	−1	−1	1	1
60	95	6	5	30	36	25
50	90	−4	0	0	16	0
$\Sigma X = 432$	$\Sigma Y = 720$			SP = 100	$SS_X = 132$	$SS_Y = 202$

The Pearson's correlation is then equal to

$$
\begin{aligned}
r &= \frac{\Sigma XY - N\bar{X}\bar{Y}}{\sqrt{(\Sigma X^2 - N\bar{X}^2)(\Sigma Y^2 - N\bar{Y}^2)}} \\
&= \frac{981 - 8(10.5)(11.5)}{\sqrt{[946 - 8(10.5)][1{,}118 - 8(11.5)]}} \\
&= \frac{981 - 966}{\sqrt{(946 - 882)(1{,}118 - 1{,}058)}} \\
&= \frac{15}{\sqrt{(64)(60)}} \\
&= \frac{15}{\sqrt{3{,}840}} \\
&= \frac{15}{61.97} \\
&= +.24
\end{aligned}
$$

10.5.2: Testing the Significance of Pearson's *r*

Pearson's *r* gives us a precise measure of the strength and direction of the correlation in the sample being studied. If we have taken a random sample from a specified population, we may seek to determine whether the obtained association between *X* and *Y* exists in the *population* and is not due merely to sampling error.

To test the significance of a measure of correlation, we usually set up the null hypothesis that no correlation exists in the population. With respect to the Pearson correlation coefficient, the null hypothesis states that the population correlation ρ (rho) is zero. That is,

$$\rho = 0$$

whereas the research hypothesis says that

$$\rho \neq 0$$

As was the case in earlier chapters, we test the null hypothesis by selecting the alpha level of .05 or .01 and computing an appropriate test of significance. To test the significance of Pearson's *r*, we can compute a *t* ratio with the degrees of freedom equal to $N - 2$ (*N* equals the number of pairs of scores). For this purpose, the *t* ratio can be computed by the formula,

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

where t = *t* ratio for testing the statistical significance of Pearson's *r*

N = number of pairs of *X* and *Y* scores

r = obtained Pearson's correlation coefficient

Returning to the previous example, we can test the significance of a correlation coefficient equal to +.24 between a father's educational level (*X*) and that of his child (*Y*):

$$
\begin{aligned}
t &= \frac{.24\sqrt{8-2}}{\sqrt{1-(.24)^2}} \\
&= \frac{(.24)(2.45)}{\sqrt{1-.0576}} \\
&= \frac{.59}{\sqrt{.9424}} \\
&= \frac{.59}{.97} \\
&= .61
\end{aligned}
$$

When we turn to Table C in Appendix C, we find that the critical value of *t* with 6 degrees of freedom and $\alpha = .05$ is 2.447. Because our calculated *t* value does not even come close to exceeding this critical value, we cannot reject the null hypothesis that $\rho = 0$. Although a correlation of +.24 is not especially weak, with a sample size of only 8, it is not nearly statistically significant. That is, given a small sample size of 8, it is very possible that the obtained *r* of +.24 is a result of sampling error. Thus, we are forced to retain the null hypothesis that the population correlation (ρ) is zero, at least until we have more data bearing on the relationship between father's and child's educational attainment.

10.5.3: A Simplified Method for Testing the Significance of *r*

Fortunately, the process of testing the significance of Pearson's *r* as previously illustrated has been simplified, so it becomes unnecessary actually to compute a *t* ratio. Instead, we turn to Table H in Appendix C, where we find a list of significant values of Pearson's *r* for the .05 and .01 levels of significance, with the number of degrees of freedom ranging from 1 to 90. Directly comparing our calculated value of *r* with the appropriate table value yields the same result as though we had actually computed a *t* ratio. If the calculated Pearson's correlation coefficient (in absolute value) does not exceed the appropriate table value, we must retain the null hypothesis that $\rho = 0$ if, on the other hand, the calculated *r* is greater than the table critical value, we reject the null hypothesis and accept the research hypothesis that a correlation exists in the population.

For illustrative purposes, let us return to our previous example in which a correlation coefficient equal to +.24 was tested by means of a *t* ratio and found not to be statistically significant. Turning to Table H in Appendix C, we now find that the value of *r* must be at least .7067 to reject the null hypothesis at the .05 level of significance with 6 degrees of freedom. Hence, this simplified method leads us to the same conclusion as the longer procedure of computing a *t* ratio.

Step-by-Step Illustration: Obtaining Pearson's Correlation Coefficient

To illustrate the step-by-step procedure for obtaining a Pearson's correlation coefficient (r), let us examine the relationship between years of school completed (X) and prejudice (Y) as found in the following sample of 10 immigrants:

Respondent	Years of School (X)	Prejudice (Y)[a]
A	10	1
B	3	7
C	12	2
D	11	3
E	6	5
F	8	4
G	14	1
H	9	2
I	10	3
J	2	10

[a] Higher scores on the measure of prejudice (from 1 to 10) indicate greater prejudice.

To obtain Pearson's r, we must proceed through the following steps:

Step 1 Find the values of $\Sigma X, \Sigma Y, \Sigma X^2, \Sigma Y^2$, and ΣXY, as well as $\overline{X}$ and $\overline{Y}$.

X	Y	X^2	Y^2	XY	
10	1	100	1	10	$N = 10$
3	7	9	49	21	$\Sigma X = 85$
12	2	144	4	24	$\Sigma Y = 38$
11	3	121	9	33	$\overline{X} = \frac{\Sigma X}{N} = \frac{85}{10} = 8.5$
6	5	36	25	30	
8	4	64	16	32	$\overline{Y} = \frac{\Sigma Y}{N} = \frac{38}{10} = 3.8$
14	1	196	1	14	$\Sigma X^2 = 855$
9	2	81	4	18	$\Sigma Y^2 = 218$
10	3	100	9	30	$\Sigma XY = 232$
2	10	4	100	20	
85	38	855	218	232	

Step 2 Plug the values from Step 1 into Pearson's correlation formula.

$$r = \frac{\Sigma XY - N\overline{X}\,\overline{Y}}{\sqrt{(\Sigma X^2 - N\overline{X}^2)(\Sigma Y^2 - N\overline{Y}^2)}}$$

$$= \frac{232 - (10)(8.5)(3.8)}{\sqrt{[855 - (10)(8.5)^2][218 - (10)(3.8)^2]}}$$

$$= \frac{232 - 323}{\sqrt{(855 - 722.5)(218 - 144.4)}}$$

$$= \frac{-91}{\sqrt{(1325)(73.6)}}$$

$$= \frac{-91}{\sqrt{9{,}752}}$$

$$= \frac{-91}{98.75}$$

$$= -.92$$

Our result indicates a rather strong negative correlation between education and prejudice.

Step 3 Find the degrees of freedom.

$$\text{df} = N - 2$$
$$= 10 - 2$$
$$= 8$$

Step 4 Compare the obtained Pearson's r with the appropriate value of Pearson's r in Table H.

$$\text{Obtained } r = -.92$$
$$\text{Table } r = .6319$$
$$\text{df} = 8$$
$$\alpha = .05$$

As indicated, to reject the null hypothesis that $\rho = 0$ at the .05 level of significance with 8 degrees of freedom, our calculated value of Pearson's r must exceed .6319 in absolute value. Because our obtained r equals −.92, we reject the null hypothesis and accept the research hypothesis. That is, our result suggests that a negative correlation between education and prejudice is present in the immigrant population from which our sample was taken.

10.5.4: Requirements for the Use of Pearson's *r* Correlation Coefficient

To employ Pearson's correlation coefficient correctly as a measure of association between X and Y variables, the following requirements must be taken into account:

1. *A straight-line relationship.* Pearson's r is only useful for detecting a straight-line correlation between X and Y.
2. *Interval data.* Both X and Y variables must be measured at the interval level so that scores may be assigned to the respondents.
3. *Random sampling.* Sample members must have been drawn at random from a specified population to apply a test of significance.
4. *Normally distributed characteristics.* Testing the significance of Pearson's r requires both X and Y variables to be normally distributed in the population. In small samples, failure to meet the requirement of normally distributed characteristics may seriously impair the validity of the test. However, this requirement is of minor importance when the sample size equals or exceeds 30 cases.

10.6: The Importance of Scatter Plots

Objective: **Examine the necessity to properly study a scatter plot before jumping to conclusions about its implications**

It seems instinctive to look for shortcuts and time-saving devices in our lives. For social researchers, the development of high-speed computers and simple statistical software has become what the advent of the automatic washer and liquid detergent was for the housekeeper. Unfortunately, these statistical programs have been used too often without sufficient concern for their appropriateness. This is particularly true in correlational analysis.

The correlation coefficient is a very powerful statistical measure. Moreover, for a data set containing several variables, with a very short computer run, one can obtain in just seconds a *correlation matrix*, such as that in Table 10.4.

Table 10.4 A Correlation Matrix

	Respondent's Age *X1*	Respondent's Education *X2*	Family Income *X3*	Spouse's Education *X4*
X1	1.00	−.48	.35	−.30
X2	−.48	1.00	.67	.78
X3	.35	.67	1.00	.61
X4	−.30	.78	.61	1.00

A correlation matrix displays in compact form the interrelationships of several variables simultaneously. Along the diagonal from the upper-left corner to the bottom-right corner is a series of 1.00s. These represent the correlation of each variable with itself, and so they are necessarily perfect and therefore equal to 1. The off-diagonal entries are the intercorrelations. The entry in the second row, fourth column (.78) gives the correlation of *X*2 and *X*4 (respondent's and spouse's education). The matrix is symmetrical—that is, the triangular portion above the diagonal is identical to that below the diagonal. Thus, the entry for the fourth row, second column is .78 as well.

The value of computer programs that produce results like this is that the researcher can quickly glance at the intercorrelations of a large number of variables—say, 10—and quickly pick out the strong and interesting correlations. One immediate problem, as we discussed earlier in reference to analysis of variance, is that such a fishing expedition for a large number of correlations will tend to pick up correlations that are significant by chance. An even greater pitfall, however, is that correlations may gloss over some major violations of the assumptions of Pearson's *r*. That is, a correlation matrix only provides linear correlation coefficients; it does not tell if the relationships are linear in the first place or whether there are peculiarities in the data that are worth noting. To prevent falling victim to data peculiarities, one really should inspect scatter plots before jumping to conclusions about what is related to what.

It is a far more tedious task to look at scatter plots in conjunction with the correlation matrix, because they must be examined one pair at a time. For example, to inspect scatter plots for all pairs of 10 variables would require 45 plots and a great deal of time and effort. As a result, far too many students and researchers skip over this step, often with misleading or disastrous results. Sometimes, as we shall see, what seems like a strong association on the basis of the correlation coefficient may be proved illusory after seeing the scatter plot. Conversely, truly important associations may be misrepresented by the single summary value of Pearson's *r*.

Consider, for example, the following data on homicide and suicide rates (per 100,000 population) for the six New England states in Table 10.5.

Table 10.5 Homicide and Suicide Rates for the Six New England States

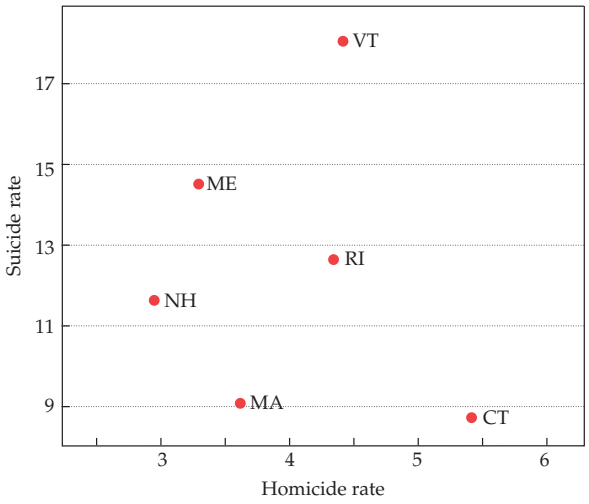
State	Homicide Rate	Suicide Rate
Maine	3.2	14.3
New Hampshire	2.9	11.3
Vermont	4.3	17.8
Massachusetts	3.6	8.9
Rhode Island	4.2	12.3
Connecticut	5.4	8.6

The correlation coefficient is −.17, suggesting a weak to moderate negative relationship. This would seem to support the contention of some sociologists that these two forms of violence (other-directed and self-directed) are trade-offs; when one rate is high, the other rate is low.

INSPECTING SCATTER PLOTS FOR OUTLIERS Before we get too excited about this result, however, let's inspect the scatter plot in Figure 10.6, based on the data in Table 10.5. Although the scatter plot appears to show a slight negative association, the lower-right point deserves further consideration. This corresponds to Connecticut. There is some justification for suspecting that Connecticut is in fact systematically different from the rest of the New England states. Suppose, for the sake of argument, we exclude Connecticut and recalculate the correlation. By using only the five other states, $r = .44$. Indeed, Connecticut has both the lowest suicide rate and the highest homicide rate in New England, which seems to have distorted the initial correlation coefficient.

There are statistical procedures for determining if this or any other data point should be excluded; they are,

Figure 10.6 Scatter Plot of Homicide and Suicide Rates in New England



however, beyond the scope of this text. Nevertheless, the importance of inspecting for these outliers is a lesson well worth learning. It can be distressing to promote a particular correlation as substantively meaningful, only to find later that the exclusion of one or two observations radically alters the results and interpretation.

10.7: Partial Correlation

Objective: **Explain how the introduction of a control variable in partial correlation allows researchers to control a two-variable relationship for the impact of a third variable**

In this chapter we have considered a powerful method for studying the association or relationship between two interval-level variables. It is important to consider if a correlation between two measures holds up when controlling for additional variables. That is, does our interpretation of the relationship between two variables change in any way when looking at the broader context of other related factors?

To see this most easily, we will focus again on scatter plots. A scatter plot visually displays all the information contained in a correlation coefficient—both its direction (by the trend underlying the points) and its strength (by the closeness of the points to a straight line). We can construct separate scatter plots for different subgroups of a sample to see if the correlation observed for the full sample holds when controlling for the subgroup or control variable. For example, there has been some research by social psychologists in recent years on the relationship between physical characteristics (such as attractiveness) and professional attainment (for example, salary or goal fulfillment). Suppose that within the context of studying the relationship between personal attributes and salary, a social psychologist stumbles on a strong positive association between height and salary, as shown in Figure 10.7. This would make sense to the social psychologist; he or she reasons that taller people tend to be more assertive and are afforded greater respect from others, which pays off in being successful in requests for raises.

But this social psychologist could be misled—in total or in part—if he or she fails to bring into the analysis other relevant factors that might alternatively account for the height–salary correlation. Gender of employee is one such possible variable. Men tend to be taller than women, and, for a variety of reasons, tend to be paid more. Perhaps this could explain all or part of the strong correlation between height and salary. Figure 10.7 also provides scatter plots of height and salary separately for males and females in the sample. It is important to note, first, that if we superimposed these two scatter plots they would produce the original plot.

Apparently, when we control for gender, the height–salary correlation weakens substantially—in fact, disappears. If any correlation remains in either of the two gender-specific subplots, it is nowhere near as strong as that which we saw at first in the uncontrolled scatter plot. Thus, had the social psychologist failed to consider the influence of gender, he or she would have been greatly misled.

Figure 10.8 illustrates additional possible outcomes when a control variable is introduced. Each scatter plot represents a positive correlation between *X* and *Y*. Observations in group 1 are symbolized by empty circles and those in group 2 by dark circles. This allows us to see the

Figure 10.7 Scatter Plot of Salary and Height Controlling for Gender

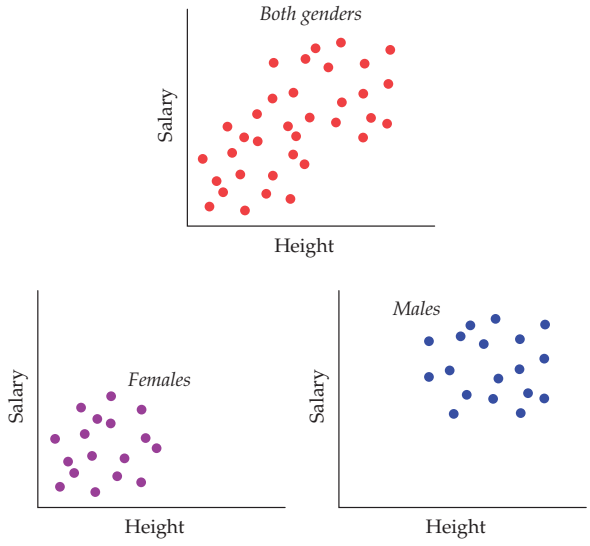
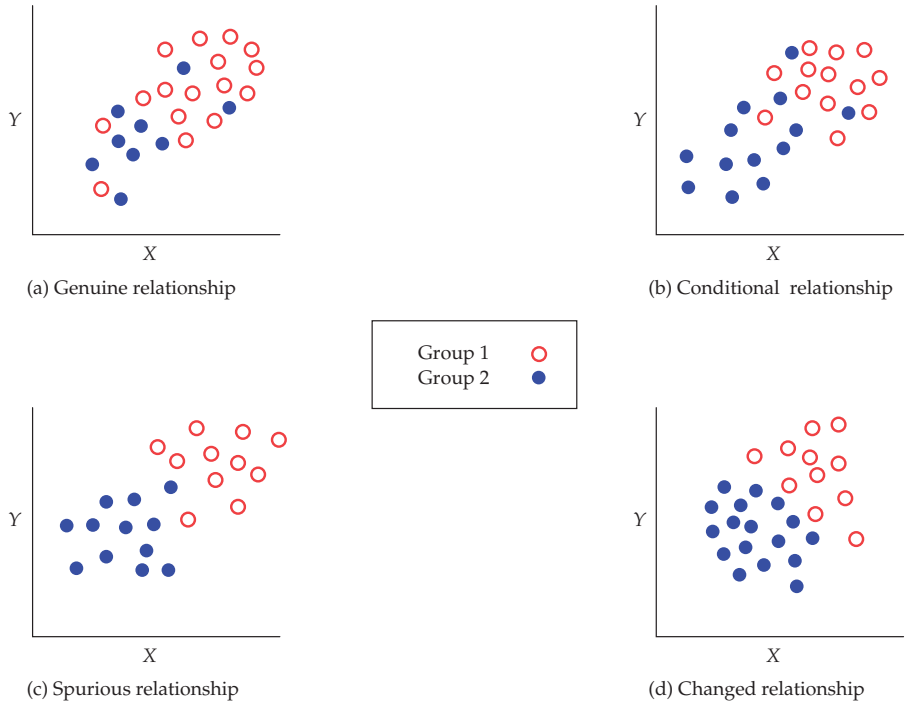


Figure 10.8 Controlling for a Third Variable



X–Y relationship within the two subgroups separately. Note that these are prototypes—in practice, one may not observe such clear-cut situations.

In scatter plot (a), we see that the X–Y association observed overall holds for each subgroup as well. Group 1 tends to exceed group 2 on both X and Y, and within these two groups X and Y are still related positively and strongly. That is, controlling for the grouping variable does not alter the X–Y relationship. For example, in Figure 10.8 (a) the positive relationship between education and income holds both for whites and non-whites. If one observes this kind of outcome when testing for a range of control variables (for example, race, sex, and age), one develops confidence in interpreting the association (for example, between education and income) as causal. In other words, one can conclude that increased education generally leads to increased income.

Scatter plot (b) shows a conditional relationship. Again, there is a strong relationship between X and Y for one group, but no relationship for the other. If the grouping variable is ignored, the correlation between X and Y misrepresents the more accurate picture within the subgroups.

Scatter plot (c) illustrates a spurious or misleading correlation. Within both subgroups, X and Y are unrelated. Overall, group 1 tends to be higher on both variables. As a result, when ignoring the subgroup distinction, it appears as if X and Y are related. Our association noted previously between height and salary is an example of a spurious correlation. Spurious correlations frequently occur in practice, and one should always be wary that two variables are related only because of their having a common cause.

Finally, scatter plot (d) shows a relationship that changes direction when a third variable is controlled. That is, the original positive association between X and Y becomes negative within the two subgroups. That group 1 was so much greater than group 2 on both X and Y overshadowed the negative relationship within each subgroup. This type of situation occurs rarely in practice, but one still should be aware that an apparent finding could be just the opposite of what it should be.

10.7.1: Control Variables with Three or More Categories

All the comparisons we have considered thus far involve dichotomous (two-category) control variables. The same approach applies to control variables having three or more levels or categories. For example, one could investigate the influence of religion on the relationship between two variables by computing Pearson's r separately for Protestants, Catholics, and Jews.

How would one handle an interval-level control variable like age? There is a temptation to categorize age into a number of subgroups (for example, under 18, 18–34, 35–49, 50 and over) and then to plot the X–Y association separately for each age category. However, this would be both inefficient and a waste of information (for example, the distinction between 18-year-olds and 34-year-olds is lost because these

two ages are within the same category). Perhaps, then, we could use narrower age groups, but we still are being less precise than we could be. Fortunately, a simple method exists for adjusting a correlation between two variables for the influence of a third variable when all three are interval level. That is, we do not have to categorize any variables artificially.

The **partial correlation coefficient** is the correlation between two variables, after removing (or partialing out) the common effects of a third variable. Like simple correlations, a partial correlation can range from −1 to +1 and is interpreted exactly the same way as a simple correlation. The formula for the partial correlation of X and Y controlling for Z is

$$r_{XY \cdot Z} = \frac{r_{XY} - r_{XZ}r_{YZ}}{\sqrt{1 - r_{XZ}^2}\sqrt{1 - r_{YZ}^2}}$$

In the notation $r_{XY \cdot Z}$ the variables before the period are those being correlated, and the variable after the period is the control variable. The partial correlation is computed exclusively on the basis of three quantities: the correlations between X and Y, X and Z, and Y and Z.

For example, consider the following correlation matrix for height (X), weight (Y), and age (Z). Not only are height and weight positively correlated, but both increase with age. One might wonder, then, how much of the correlation between height and weight ($r = .90$) is due to the common influence of age and how much remains after the influence of age is controlled:

	Height	Weight	Age
Height (X)	1.00	.90	.80
Weight (Y)		1.00	.85
Age (Z)			1.00

$$\begin{aligned} r_{XY.Z} &= \frac{r_{XY} - r_{XZ}r_{YZ}}{\sqrt{1 - r^2_{XZ}}\sqrt{1 - r^2_{YZ}}} \\ &= \frac{.90 - (.80)(.85)}{\sqrt{1 - (.80)^2}\sqrt{1 - (.85)^2}} \\ &= \frac{.90 - .68}{\sqrt{1 - .64}\sqrt{1 - .7225}} \\ &= \frac{.22}{\sqrt{.36}\sqrt{.2775}} \\ &= \frac{.22}{(.60)(.5268)} \\ &= \frac{.22}{.3161} \\ &= +.70 \end{aligned}$$

Thus, the strong initial correlation between height and weight ($r = .90$) weakens somewhat when the effects of age are removed ($r_{xyz} = .70$).

SIMPLE VERSUS PARTIAL CORRELATION We saw in Figure 10.8 that there are many possible patterns when controlling for a third variable. Similarly, partial correlations can be smaller, equal to, or greater than the two-variable simple correlation. Consider, for example, the following correlation matrix for education (X), salary (Y), and age (Z).

	Education	Salary	Age
Education (X)	1.00	.40	−.30
Salary (Y)		1.00	.50
Age (Z)			1.00

The simple correlation between education and salary is .40 but the partial correlation between education and salary controlling for age is even higher:

$$\begin{aligned} r_{XY.Z} &= \frac{r_{XY} - r_{XZ}r_{YZ}}{\sqrt{1 - r^2_{XZ}}\sqrt{1 - r^2_{YZ}}} \\ &= \frac{.40 - (-.30)(.50)}{\sqrt{1 - (-.30)^2}\sqrt{1 - (.50)^2}} \\ &= \frac{.55}{\sqrt{.91}\sqrt{.75}} \\ &= \frac{.55}{(.9539)(.8660)} \\ &= \frac{.55}{.8261} \\ &= +.67 \end{aligned}$$

Thus, ignoring age suppresses the observed association between education and salary. Younger employees, because of their low seniority, have lower salaries, despite their higher level of educational attainment. As a result, the influence of education is dwarfed in the simple correlation because highly educated employees, who you would think should be paid more than they are, do not have the salary expected because they tend to be younger and newer employees. By controlling for age, we isolate the effect of education on salary, absent the influence of age.

THE PARTIAL CORRELATION COEFFICIENT AND SPURIOUS RELATIONSHIPS The partial correlation coefficient is a very useful statistic for finding spurious relationships, as is demonstrated in this classic case of a "vanishing" correlation.[1] The correlation between the rate of forcible rape (per 100,000) in 1982 and the circulation of *Playboy* (per 100,000) in 1979 for 49 U.S. states (Alaska is an outlier on rape and is excluded) is $r = +.40$. Because of this substantial correlation, many observers have asked, if *Playboy* has this kind of effect on sex crimes, imagine what harm may be caused by truly hard-core pornography?

This concern stems from the unjustified assumption that the correlation implies cause. Before making such a leap, however, we need to consider whether the two variables have a third variable as a common cause, thereby producing a spurious result.

[1]We thank Rodney Stark and Cognitive Development, Inc., for this fine illustration and for these data.

As it turns out, both the rape and the *Playboy* subscription rate are related to the rate of homes without an adult female (per 1,000 households): For the rape rate (Y) and the rate of homes without an adult female (Z), $r_{YZ} = +.48$; for the rate of subscription to *Playboy* (X) and the rate of homes without an adult female (Z), $r_{YZ} = +.85$. Apparently, both types of sexual outlet (one illegal and one legal) sometimes stem from the absence of adult females in the home.

To determine the correlation of *Playboy* (X) with rape (Y), controlling for homes without adult females (Z), we calculate the partial correlation:

$$\begin{aligned} r_{XY \cdot Z} &= \frac{r_{XY} - r_{XZ}r_{YZ}}{\sqrt{1 - r_{XZ}^2}\sqrt{1 - r_{YZ}^2}} \\ &= \frac{.40 - (.85)(.48)}{\sqrt{1 - (.85)^2}\sqrt{1 - (.48)^2}} \\ &= \frac{.40 - .41}{\sqrt{1 - .7225}\sqrt{1 - .2304}} \\ &= \frac{-.01}{\sqrt{.2775}\sqrt{.7696}} \\ &= \frac{-.01}{(.53)(.88)} \\ &= \frac{-.01}{.47} \\ &= -.02 \end{aligned}$$

As a result, after controlling for one common variable, the original correlation disappears.

10.7.2: Testing the Significance of a Partial Correlation

Earlier in the chapter, we followed up the calculation of Pearson's correlation (the non-partial r) by testing the null hypothesis that the population correlation (ρ) was zero. Similarly, in the case of partial correlation, we seek a significance test concerning its magnitude in the population. As it happens, the testing procedure, using either the simplified method in Table H or a t ratio with Table C, is the same for a partial correlation, but with a slight modification. Because there are three variables involved—the two being correlated and the third held constant, we have $N - 3$ degrees of freedom. Thus, we can compare our calculated value of $r_{XY \cdot Z}$ against the critical value contained in Table H corresponding to $N - 3$ df and the desired level of significance. Alternatively, we can use the t formula after replacing $N - 2$ with $N - 3$:

$$t = \frac{r_{XY \cdot Z}\sqrt{N - 3}}{\sqrt{1 - r_{XY \cdot Z}^2}}$$

To illustrate, let's consider testing the significance of the partial correlation between education and salary with age held constant. We determined that $r_{XY \cdot Z} = .67$, a strong association between education and salary that grew even stronger after removing the effects of age. If this were based on a sample size $N = 25$, we would obtain

$$\begin{aligned} t &= \frac{.67\sqrt{25 - 3}}{\sqrt{1 - .67^2}} \\ &= \frac{.67(22)}{\sqrt{.55}} \\ &= \frac{14.74}{.74} \\ &= 19.9 \end{aligned}$$

Checking Table C in Appendix C with 22 df and a .05 level of significance, we find that the two-tail critical value of t is 2.074. Thus, we can easily reject the null hypothesis of no partial correlation in the population between education and salary with age held constant.

Summary: Correlation

In this chapter, we went beyond the task of establishing the presence or absence of a relationship between two variables. In correlation, the social researcher is interested in the degree of association between two variables. With the aid of the correlation coefficient known as Pearson's r, it is possible to obtain a precise measure of both the strength—from 0.0 to 1.0—and direction—positive versus negative—of a relationship between two variables that have been measured at the interval level. Moreover, if a researcher has taken a random sample of scores, he or she may also determine whether the obtained relationship between X and Y exists in the population and is not due merely to sampling error, by consulting a table of critical values or computing a t ratio. In addition, the partial correlation coefficient allows the researcher to control a two-variable relationship for the impact of a third variable.

Homework 10.1: Practice Correlation

1. The following six students were questioned regarding (*X*) their attitudes toward the legalization of prostitution and (*Y*) their attitudes toward the legalization of marijuana. Compute a Pearson's correlation coefficient for these data and determine whether the correlation is significant.

Student	X	Y
A	1	2
B	6	5
C	4	3
D	3	3
E	2	1
F	7	4

2. A high school guidance counselor is interested in the relationship between proximity to school and participation in extracurricular activities. He collects the data on distance from home to school (in miles) and number of clubs joined for a sample of 10 juniors. Using the following data, compute a Pearson's correlation coefficient and indicate whether the correlation is significant.

	Distance to School (miles)	Number of Clubs Joined
Lee	4	3
Ronda	2	1
Jess	7	5
Evelyn	1	2
Mohammed	4	1
Steve	6	1
George	9	9
Juan	7	6
Chi	7	5
David	10	8

3. An urban sociologist interested in neighborliness collected data for a sample of 10 adults on (*X*) how many years they have lived in their neighborhood and (*Y*) how many of their neighbors they regard as friends. Compute a Pearson's correlation coefficient for these data and determine whether the correlation is significant.

X	Y	X	Y
1	1	2	1
5	4	5	2
6	2	9	6
1	3	4	7
8	5	2	0

4. An economist is interested in studying the relationship between length of unemployment and job-seeking activity among white-collar workers. He interviews a sample of 12 unemployed accountants as to the number of weeks they have been unemployed (*X*) and seeking a job during the past year (*Y*). Compute a Pearson's correlation coefficient for these data and determine whether the correlation is significant.

Accountant	X	Y
A	2	8
B	7	3
C	5	4
D	12	2
E	1	5
F	10	2
G	8	1
H	6	5
I	5	4
J	2	6
K	3	7
L	4	1

5. A psychiatrist is concerned about his daughter, who has suffered from extremely low self-esteem and high anxiety since entering high school last year. Wondering if many of his daughter's peers have the same troubles, he collects a random sample of high school girls and anonymously asks them how much they agree or disagree (on a scale from 1 to 7, with 1 being "strongly disagree" and 7 being "strongly agree") with the following statements: (*X*) "I felt better about myself before I started high school" and (*Y*) "I have felt very anxious since I started high school." Compute a Pearson's correlation coefficient for the following data and indicate whether the correlation is significant.

X	Y
7	5
6	4
4	3
5	6
3	2
6	7
5	5

6. A new high school special education teacher wonders if there really is a correlation between reading disabilities and attention disorders. She collects data from six of her students on their reading abilities (*X*) and their attention abilities (*Y*), with a higher score indicating greater ability for both variables. Compute a Pearson's correlation

coefficient for the following data and indicate whether the correlation is significant.

X	Y
2	3
1	2
5	3
4	2
2	4
1	3

7. A researcher wonders if there is a correlation between (*X*) people's opinions of bilingual education and (*Y*) their opinions about whether foreign-born citizens should be allowed to run for president. She collects the following data, with both variables being measured on a scale from 1 to 9 (1 being strongly opposed and 9 being strongly in favor).

X	Y
2	5
5	2
8	6
6	9
1	3
2	1
8	5
3	1

Calculate a Pearson's correlation coefficient and determine whether the correlation is significant.

8. Obesity in children is a major concern because it puts them at risk for several serious medical problems. Some researchers believe that a major issue related to this is that children these days spend too much time watching television and not enough time being active. Based on a sample of boys of roughly the same age and height, data were collected regarding hours of television watched per day and weight. Compute a Pearson's correlation coefficient and indicate whether the correlation is significant.

TV Watching (hrs)	Weight (lbs)
1.5	79
5.0	105
3.5	96
2.5	83
4.0	99
1.0	78
0.5	68

9. Is there a relationship between (*X*) rate of poverty (measured as percent of population below poverty level) and (*Y*) rates of teen pregnancy (measured per 1,000 females aged 15 to 17)? A researcher selected random states and collected the following data. Compute a Pearson's correlation coefficient and determine whether the correlation is significant.

State	X	Y
A	10.4	41.7
B	8.9	38.6
C	13.3	43.2
D	6.9	35.7
E	16.0	46.9
F	5.2	33.5
G	14.5	43.3
H	15.3	44.8

10. A researcher set out to determine whether suicide and homicide rates in metropolitan areas around the country are correlated and, if so, whether they vary inversely (negative correlation) or together (positive correlation). Using available data for a recent year, he compared the following sample of 10 metropolitan areas with respect to their rates (number per 100,000), of suicide and homicide:

Metropolitan Area	Suicide Rate	Homicide Rate	Metropolitan Area	Suicide Rate	Homicide Rate
A	20.2	22.5	F	21.4	19.5
B	22.6	28.0	G	9.8	13.2
C	23.7	15.4	H	13.7	16.0
D	10.9	12.3	I	15.5	17.7
E	14.0	12.6	J	18.2	20.8

What is the strength and direction of correlation between suicide and homicide rates among the 10 metropolitan areas sampled? Test the null hypothesis that rates of suicide and homicide are not correlated in the population.

11. An educational researcher interested in the consistency of school absenteeism over time studied a sample of eight high school students for whom complete school records were available. The researcher counted the number of days each student had missed while in the sixth grade and then in the tenth grade. He obtained the following results:

Student	Days Missed (6th)	Days Missed (10th)
A	4	10
B	2	4
C	21	11
D	1	3
E	3	1
F	5	5
G	4	9
H	8	5

What is the strength and direction of the relationship between the number of days these students were absent from elementary school (sixth grade) and how many days they missed when they reached high school (tenth grade)? Can the correlation be generalized to a larger population of students?

12. Do reading and television viewing compete for leisure time? To find out, a communication specialist interviewed a sample of 10 children regarding the number of books they had read during the last year and the number of hours they had spent watching television on a daily basis. Her results are as follows:

Number of Books	Hours of TV Viewing
0	3
7	1
2	2
1	2
5	0
4	1
3	3
3	2
0	7
1	4

What is the strength and direction of the correlation between number of books read and hours of television viewing daily? Is the correlation significant?

13. An education specialist is interested in how the number of words learned within the first two years of a child's life affects the trajectory of academic performance later in high school.

Words Known at Age 2	High School GPA
20	3.52
14	3.14
18	3.96
8	2.82
7	2.34
17	3.67

a. To test his idea, calculate Pearson's correlation coefficient for the relationship between "words known at age 2" and "high school grade point average (GPA)" for a sample of six teenagers. What is the strength and direction of this relationship?

b. Using the correlation coefficient results, calculate the statistical significance of the relationship. Can the null hypothesis be rejected? In other words, does this relationship hold true for the population?

14. A sociologist researching the link between race and inequality believes that younger people generally have more tolerant views. To test this hypothesis, the sociologist gathers a random sample of 10 Caucasians ranging in age from teenager to elderly and asks them for their beliefs about African-Americans on a 10-point scale (1 representing acceptance of negative beliefs, 10 representing acceptance of positive beliefs).

Age	Beliefs about African-Americans
16	9
18	10
20	7
29	8
34	6
46	3
51	5
65	4
69	5
82	3

a. Check to see if his idea about the relationship between age and racial stereotyping is correct by calculating and interpreting the strength and direction of Pearson's correlation coefficient.

b. Using the correlation coefficient results, determine the statistical significance of the relationship. Can the null hypothesis be rejected? In other words, does this relationship hold true for the population?

15. In addition to job-seeking activity, the age of a white-collar worker may be related to his or her length of unemployment. Suppose then that age (Z) is added to the two variables in Problem 15.

Accountant	Weeks Unemployed (X)	Weeks Seeking (Y)	Age (Z)
A	2	8	30
B	7	3	42
C	5	4	36
D	12	2	47
E	1	5	29
F	10	2	56
G	8	1	52
H	6	5	40
I	5	4	27
J	2	6	31
K	3	7	36
L	4	1	33

Find the partial correlation of weeks unemployed and weeks seeking a job, holding the age of the worker constant, and then determine if this partial correlation is significant.

16. In preparing for an examination, some students in a class studied more than others. Besides studying

time, intelligence itself may be related to test performance.

	Hours Studied (*X*)	Exam Grade (*Y*)	IQ (*Z*)
Barbara	4	5	100
Bob	1	2	95
Deidra	3	1	95
Owen	5	5	108
Charles	8	9	110
Emma	2	7	117
Sanford	7	6	110
Luis	6	8	115

Find the partial correlation of studying time and exam grade, holding IQ constant.

17. The following is a correlation matrix among family size (*X*), weekly grocery bill (*Y*), and income (*Z*) for a random sample of 50 families.

	X	*Y*	*Z*
X	1.00	.60	.20
Y	.60	1.00	.30
Z	.20	.30	1.00

a. Which of the correlations are significant at the .05 level?

b. Calculate the partial correlation between family size and grocery bill, holding income constant, and then determine if this partial correlation is significant.

Discuss the difference between the simple correlation r_{XY} and the partial correlation $r_{XY.Z}$.

18. It has become common, though potentially inaccurate, to believe that playing classical music to infants can increase their brain development, spatial awareness, and creativity. A child development specialist wants to test whether the exposure to Mozart's calming compositions does in fact increase childhood neural development. He gathers a sample of 10 children chosen at random and tracks them from infancy through college. When each of his subjects turns 16, he administers an IQ test and then compares the results with the average hours a year the participant was played Mozart's melodies as an infant.

IQ Score	Hours of Mozart/Year
128	68
104	51
96	18
107	26
106	36
99	47
91	10
111	72
131	94
117	48

a. Calculate Pearson's correlation coefficient and interpret the strength and direction of the results.

b. Using the correlation coefficient results, calculate the statistical significance of the relationship. Can the null hypothesis be rejected? In other words, does this relationship hold true for the population?

19. With regard to the earlier problems, many sociologists and developmental psychologists have noted that it might not be listening to Mozart per se that increases a child's neural development. Instead, they argue, increased intelligence may be attributable to having the type of parents who would intentionally play music to their infant because it reputedly stimulates child development. This leads to a different hypothesis: There is a positive correlation between children's intelligence and parents' intervention in general, and not the specific nature of the intervention.

IQ Score	Hours of Mozart/Year	Parental Intervention Hours
128	68	14
104	51	6
96	18	4
107	26	7
106	36	7
99	47	5
91	10	2
111	72	9
131	94	14
117	48	12

a. To test this alternative explanation, calculate and interpret the strength and direction of a partial correlation between hours of Mozart (*X*) and IQ score at age 16 (*Y*) holding constant the amount of time (in hours per week) parents spent on interventions with their child (*Z*). Make sure to construct a correlation matrix, it will aid you in solving for $r_{XY.Z}$.

b. Using the correlation coefficient results calculate the statistical significance of the one correlation between *X* and *Y* and then between *X* and *Y* holding constant *Z*. How does controlling for *Z* change the significance of the correlation between *X* and *Y*?

Homework 10.2: General Social Survey Practice

1. Analyze the General Social Survey to generate a single correlation matrix which will allow you to test the following null hypotheses: Remember to weight the cases by WTSSALL.

   Null hypothesis 1: *There is no relationship between days of poor mental health during the past 30 days (MNTLHLTH) and days of poor physical health during the past 30 days (PHYSHLTH).*

   Null hypothesis 2: *There is no relationship between days of poor mental health during the past 30 days (MNTLHLTH) and age (AGE).*

   Null hypothesis 3: *There is no relationship between days of poor physical health during the past 30 days (PHYSHLTH) and age (AGE).*

   Create a correlation matrix of the three variables. Report the strength and direction of the Pearson's *r* correlation coefficients. *Hint*: ANALYZE, CORRELATE, BIVARIATE, and choose the variables.

2. Generate a single correlation matrix which will allow you to test the null hypothesis of no relationship for all of the following pairs of variables from the General Social Survey:

   Job satisfaction (SATJOB1) and income (REALRINC)
   Job satisfaction (SATJOB1) and years of education (EDUC)
   Income (REALRINC) and education (EDUC)
   Days of poor mental health during the past 30 days (MNTLHLTH) and job satisfaction (SATJOB1)

   a. Create the correlation matrix.
   b. Report the strength and direction of the Pearson's *r* correlation coefficients for each pair of variables.
   c. What other pairs of variables could be tested using this same correlation matrix?

3. Use the General Social Survey to generate a single correlation matrix to test the null hypothesis of no relationship for all of the following pairs of variables:

   Age your first child was born (AGEKDBRN) and year of birth (COHORT)
   Age your first child was born (AGEKDBRN) and the number of children you have (CHILDS)
   Ideal number of children (CHLDIDEL) and year of birth (COHORT)
   Another variable of your choice to correlate with number of children (CHILDS)

   a. Create the correlation matrix.
   b. Report the strength and direction of the Pearson's *r* correlation coefficients for each pair of variables.

4. Using the General Social Survey, calculate Pearson's *r* to test the following null hypotheses:

   Null hypothesis 1: *Respondent's highest year of school completed (EDUC) is not related to their fathers' highest year of school completed (PAEDUC).*

   Null hypothesis 2: *Respondent's highest year of school completed (EDUC) is not related to personal income (REALRINC).*

   a. Create the correlation matrix.
   b. Report the strength and direction of the Pearson's *r* correlation coefficients for each pair of variables.

5. Choose two variables from the General Social Survey so that you may generate and interpret a Pearson's *r* correlation to test the null hypothesis of no correlation between the variables.

Chapter 10 Quiz: Correlation