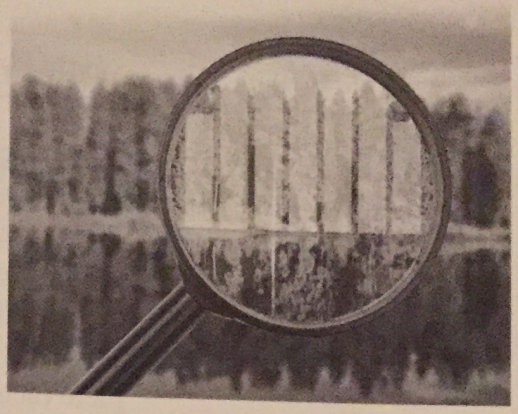
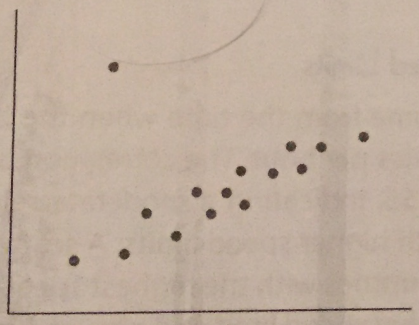
CHAPTER 11

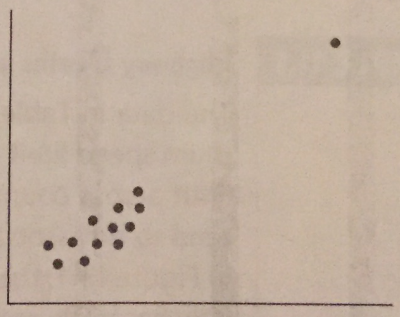


Relationships Can Be Deceiving

Thought Questions

1. Use the following two pictures to speculate on what influence outliers have on correlation. For each picture, do you think the correlation is higher or lower than it would be without the outlier? (*Hint:* Remember that correlation measures how closely points fall to a straight line.)





2. A strong correlation has been found in a certain city in the northeastern United States between sales of hot chocolate and sales of facial tissues measured weekly for a year. Would you interpret that to mean that hot chocolate causes people to need facial tissues? Explain.

3. Researchers have shown that there is a positive correlation between the average fat intake and the breast cancer rate across countries. In other words, countries with higher fat intake tend to have higher breast cancer rates. Does this correlation prove that dietary fat is a contributing cause of breast cancer? Explain.

4. If you were to draw a scatterplot of *number of women in the workforce* versus *number of Christmas trees sold* in the United States for each year between 1930 and the present, you would find a very strong correlation. Why do you think this would be true? Does one cause the other?

11.1 Illegitimate Correlations

In Chapter 10, we learned that the correlation between two measurement variables provides information about how closely related they are. A strong correlation implies that the two variables are closely associated or related. With a positive correlation, they increase together, and with a negative correlation, one variable tends to increase as the other decreases.

However, as with any numerical summary, correlation does not provide a complete picture. A number of anomalies can cause misleading correlations. Ideally, all reported correlations would be accompanied by a scatterplot. Without a scatterplot, however, you need to ascertain whether any of the problems discussed in this section may be distorting the correlation between two variables.

Watch out for these problems with correlations:

- Outliers can substantially inflate or deflate correlations.
- Groups combined inappropriately may mask relationships.

The Impact Outliers Have on Correlations

In a manner similar to the effect we saw on means, outliers can have a large impact on correlations. This is especially true for small samples. An outlier that is consistent with the trend of the rest of the data will inflate the correlation. An outlier that is not consistent with the rest of the data can substantially decrease the correlation.

EXAMPLE 11.1 Highway Deaths and Speed Limits

The data in Table 11.1 come from the time when the United States still had a maximum speed limit of 55 miles per hour. The correlation between death rate and speed limit across countries is .55, indicating a moderate relationship. Higher death rates tend to be associated with higher speed limits. A scatterplot of the data is presented in Figure 11.1; the two countries with the highest speed limits are labeled. Notice that Italy has both a much higher speed limit and a much higher death rate than any other country. That fact alone is responsible for the magnitude of the correlation. In fact, if Italy is removed, the correlation drops to .098, a negligible association. Of course, we could now claim that Britain is responsible for the almost zero magnitude of the correlation, and we would be right. If we remove Britain from the plot, the correlation is no longer negligible; it jumps to .70. You can see how much influence outliers have, sometimes inflating correlations and sometimes deflating them. (Of course, the actual relationship between speed limit and death rate is complicated by many other factors, a point we discuss later in this chapter.) ■

One of the ways in which outliers can occur in a set of data is through erroneous recording of the data. Statisticians speculate that at least 5% of all

TABLE 11.1 Highway Death Rates and Speed Limits

Country	Death Rate (Per 100 Million Vehicle Miles)	Speed Limit (in Miles Per Hour)
Norway	3.0	55
United States	3.3	55
Finland	3.4	55
Britain	3.5	70
Denmark	4.1	55
Canada	4.3	60
Japan	4.7	55
Australia	4.9	60
Netherlands	5.1	60
Italy	6.1	75

Source: Rivkin, 1986.

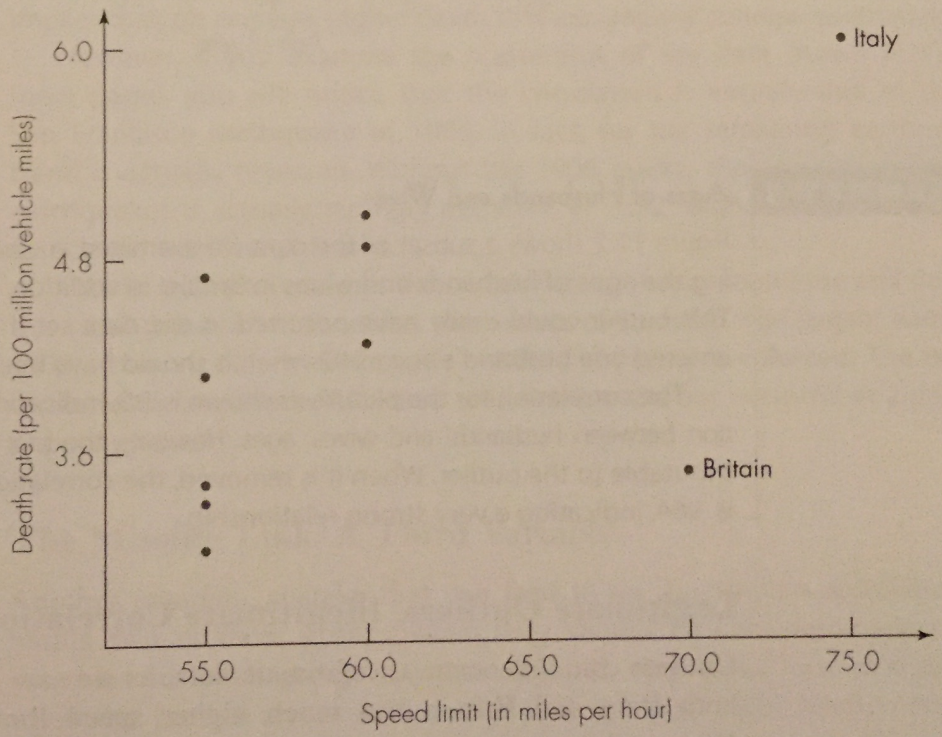


Figure 11.1
An example of how an outlier can inflate correlation
Source: Rivkin, 1986.

data points are corrupted, either when they are initially recorded or when they are entered into the computer. Good researchers check their data using scatterplots, stemplots, and other methods to ensure that such errors are detected and corrected. However, they do sometimes escape notice, and they can play havoc with numerical measures like correlation.

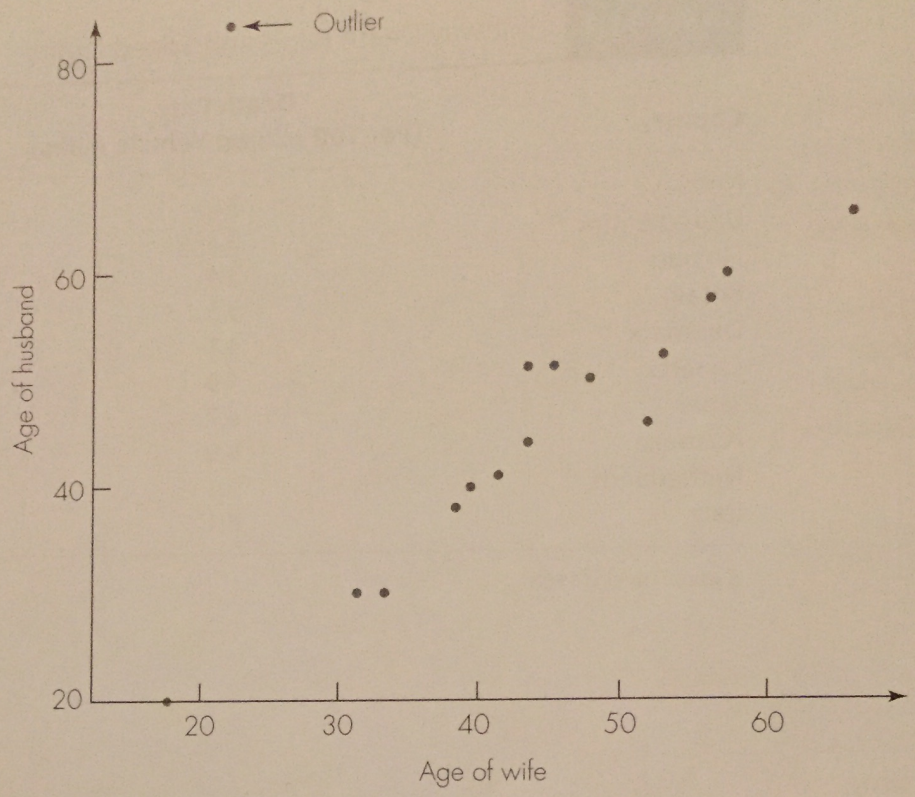


Figure 11.2
An example of how an outlier can deflate correlation
Source: Adapted from Figure 10.1.

EXAMPLE 11.2 Ages of Husbands and Wives

Figure 11.2 shows a subset of the data we examined in Chapter 10, Figure 10.1, relating the ages of husbands and wives in Britain. In addition, an outlier has been added. This outlier could easily have occurred in the data set if someone had erroneously entered one husband's age as 82 when it should have been 28.

The correlation for the picture as shown is .39, indicating a somewhat low correlation between husbands' and wives' ages. However, the low correlation is completely attributable to the outlier. When it is removed, the correlation for the remaining points is .964, indicating a very strong relationship. ■

Legitimate Outliers, Illegitimate Correlation

Outliers can also occur as legitimate data, as we saw in the example for which both Italy and Britain had much higher speed limits than other countries. However, the theory of correlation was developed with the idea that both measurements were from bell-shaped distributions, so outliers would be unlikely to occur. As we have seen, correlations are quite sensitive to outliers. Be very careful when you are presented with correlations for data in which outliers are likely to occur or when correlations are presented for a small sample, as shown in Example 11.3. Not all researchers or reporters are aware of the havoc outliers can play with correlation, and they may innocently lead you astray by not giving you the full details.

TABLE 11.2 Major Earthquakes in the Continental United States, 1880–2012

Date	Location	Deaths	Magnitude
August 31, 1886	Charleston, SC	60	6.6
April 18–19,1906	San Francisco, CA	503	8.3
March 10, 1933	Long Beach, CA	115	6.2
February 9, 1971	San Fernando Valley, CA	65	6.6
October 17, 1989	San Francisco area (CA)	62	7.1
June 28, 1992	Yucca Valley, CA	1	7.5
January 17, 1994	Northridge, CA	61	6.8

Source: http://earthquake.usgs.gov/earthquakes/states/historical.php, accessed June 13, 2013

EXAMPLE 11.3 **Earthquakes in the Continental United States**

Table 11.2 lists major earthquakes that occurred in the continental United States between 1880 and 2012. The correlation between deaths and magnitude for these seven earthquakes is .689, showing a relatively strong association. This relationship implies that, on average, higher death tolls accompany stronger earthquakes.

However, if you examine the scatterplot of the data shown in Figure 11.3 (next page), you will notice that the correlation is entirely due to the famous San Francisco earthquake of 1906. In fact, for the remaining earthquakes, the trend is actually reversed. Without the 1906 quake, the correlation for these six earthquakes is actually strongly negative, at −.92. Higher-magnitude quakes are associated with fewer deaths.

Clearly, trying to interpret the correlation between magnitude and death toll for this small group of earthquakes is a misuse of statistics. The largest earthquake, in 1906, occurred before earthquake building codes were enforced. The next largest quake, with magnitude 7.5, killed only one person but occurred in a very sparsely populated area. ■

The Missing Link: A Third Variable

Another common mistake that can lead to an illegitimate correlation is combining two or more groups when they should be considered separately. The variables for each group may actually fall very close to a straight line, but when the groups are examined together, the individual relationships may be masked. As a result, it will appear that there is very little correlation between the two variables.

This problem is a variation of "Simpson's Paradox" for count data, a phenomenon we will study in the next chapter. However, statisticians do not seem to be as alert to this problem when it occurs with measurement data. When you read that two variables have a very low correlation, ask yourself whether data may have been combined into one correlation when groups should, instead, have been considered separately.

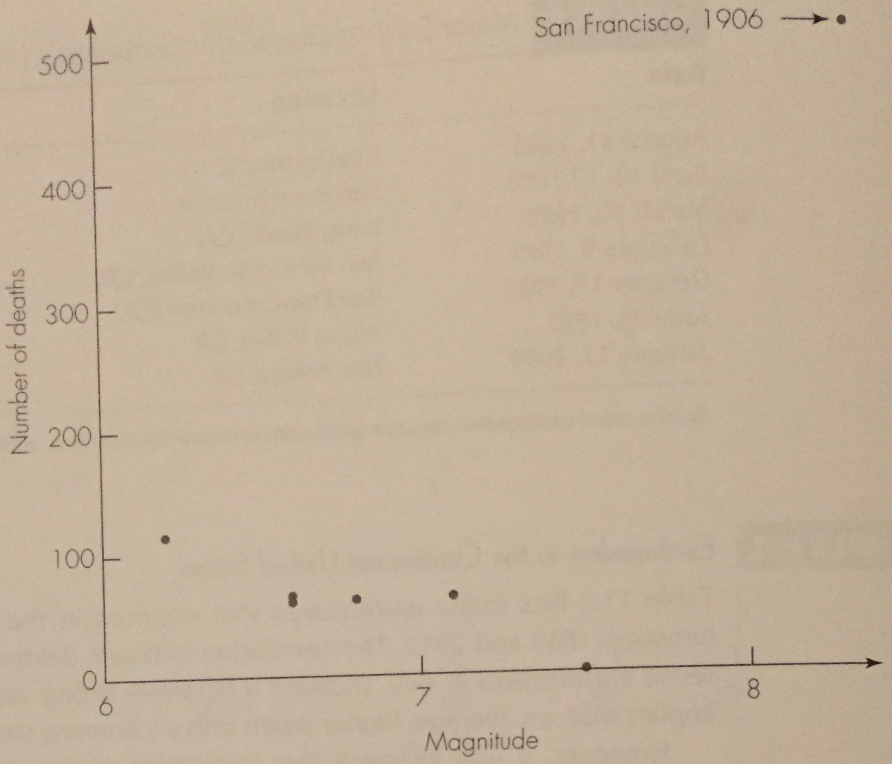


Figure 11.3
A data set for which correlation should not be used.
Source: Data from Table 11.2.

EXAMPLE 11.4 **The Fewer the Pages, the More Valuable the Book?**

If you peruse the bookshelves of a typical college professor, you will find a variety of books ranging from textbooks to esoteric technical publications to paperback novels. To determine whether the price of a book can be determined by the number of pages it contains, a college professor recorded the number of pages and price for 15 books on one shelf. The numbers are shown in Table 11.3. Is there a relationship between number of pages and the price of the book? The correlation for these figures is −.312. The negative correlation indicates that the more pages a book has, the less it costs, which is certainly a counterintuitive result.

Figure 11.4 illustrates what has gone wrong. It displays the data in a scatterplot, but it also identifies the books by type. The letter H indicates a hardcover book; the letter S indicates a softcover book. The collection of books on the professor's shelf consisted of softcover novels, which tend to be long but inexpensive, and hardcover technical books, which tend to be shorter but very expensive. If the correlations are calculated within each type, we find the result we would expect. The correlation between number of pages and price is .64 for the softcover books alone, and .35 for the hardcover books alone. Combining the two types of books into one collection not only masked the positive association between length and price, but produced an illogical negative association. ■

TABLE 11.3 Pages versus Price for the Books on a Professor's Shelf

Pages	Price	Pages	Price	Pages	Price
104	32.95	342	49.95	436	5.95
188	24.95	378	4.95	458	60.00
220	49.95	385	5.99	466	49.95
264	79.95	417	4.95	469	5.99
336	4.50	417	39.75	585	5.95

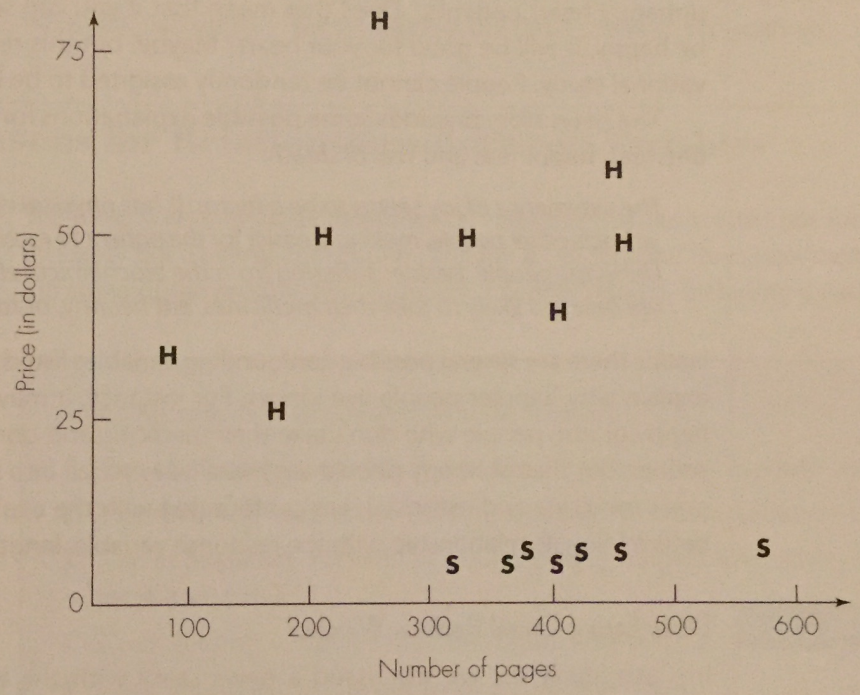


Figure 11.4 Combining groups produces misleading correlations (H = hardcover; S = softcover).
Source: Data from Table 11.3.

11.2 Legitimate Correlation Does Not Imply Causation

Even if two variables are legitimately related or correlated, do not fall into the trap of believing there is a causal connection between them. Although "correlation does not imply causation" is a very well known saying among researchers, relationships and correlations derived from observational studies are often *reported* as if the connection were causal.

It is easy to construct silly, obvious examples of correlations that do not result from causal connections. For example, a list of weekly tissue sales and weekly hot chocolate sales for a city with extreme seasons would probably exhibit a correlation because both

tend to go up in the winter and down in the summer. A list of shoe sizes and vocabulary words mastered by school children would certainly exhibit a correlation because older children tend to have larger feet and to know more words than younger children.

The problem is that sometimes the connections do seem to make sense, and it is tempting to treat the observed association as if there were a causal link. Remember that data from an observational study, in the absence of any other evidence, simply cannot be used to establish causation.

EXAMPLE 11.5 **Happiness and Heart Disease**

News Story 4 in the Appendix and on the companion website notes that "heart patients who are happy are much more likely to be alive 10 years down the road than unhappy heart patients." Does that mean that if you can somehow force yourself to be happy, it will be good for your heart? Maybe, but this research is clearly an observational study. People cannot be randomly assigned to be happy or not.

The news story provides some possible explanations for the observed relationship between happiness and risk of death:

> *The experience of joy seems to be a factor. It has physical consequences and also attracts other people, making it easier for the patient to receive emotional support. Unhappy people, besides suffering from the biochemical effects of their sour moods, are also less likely to take their medicines, eat healthy, or to exercise.* (p. 9)

Notice there are several possible confounding variables listed in this quote that may help explain why happier people live longer. For instance, it may be the case that whether happy or not, people who don't take their medicine and don't exercise are likely to die sooner, but that unhappy people are more likely to fall into that category. Thus, taking one's medicine and exercising are confounded with the explanatory variable, mood, in determining its relationship with the response variable, length of life. ■

EXAMPLE 11.6 **Does Eating Cereal Reduce Weight?**

In Case Study 6.2 we discussed a news story with the heading "Breakfast cereal tied to lower BMI for kids" (http://www.reuters.com/article/2013/04/09/us-health-breakfast-idUSBRE93815320130409), which reported on a dietary study conducted with low income children in Austin, Texas. The study was originally reported in the *Journal of the Academy of Nutrition and Dietetics* (Frantzen et al, 2013). The researchers asked children what they ate for three consecutive days when they were in each of grades 4, 5, and 6. In one part of the study, the explanatory variable was number of days of eating breakfast cereal, and the response variable was the child's percentile for body mass index (BMI). In other words, the response was a measure of how overweight the child was at that time. The study reported that on average, children who ate breakfast cereal had lower BMI than children who did not, and the more cereal they ate the lower their BMI percentile was.

We may be tempted to believe that eating breakfast cereal causes children to weigh less. That's possible, but it is more likely that the observed relationship can be explained by something else. For instance, a possible confounding variable is general dietary habits in the home. Children who are fed a healthy breakfast are more likely to

be fed healthy meals at other times of the day and thus are likely to weigh less than those who don't eat as well. Another possibility is that both variables have a common cause, such as high metabolism. People with high metabolism are less likely to be overweight than people who have lower metabolism and are also less able to skip breakfast. (Note that the study did not distinguish between eating no breakfast and eating something other than cereal. The explanatory variable was simply the number of days the child ate cereal.) ■

In the next section we will examine seven possible reasons for a relationship between two variables. The first reason is that a change in the explanatory variable really is causing a change in the response variable. But there are six other explanations that could account for an observed relationship. Remember the well-worn phrase, "correlation does not imply causation." Always think about other possible explanations.

1.3 Some Reasons for Relationships Between Variables

We have seen numerous examples of variables that are related but for which there is probably not a causal connection. To help us understand this phenomenon, let's examine some of the reasons two variables could be related, including a causal connection.

Some reasons two variables could be related:

1. The explanatory variable is the direct cause of the response variable.
2. The response variable is causing a change in the explanatory variable.
3. The explanatory variable is a contributing but not sole cause of the response variable.
4. Confounding variables may be responsible for the observed relationship.
5. Both variables may result from a common cause.
6. Both variables are changing over time.
7. The association may be nothing more than coincidence.

Reason 1: The explanatory variable is the direct cause of the response variable. Sometimes, a change in the explanatory variable is the direct cause of a change in the response variable. For example, if we were to measure amount of food consumed in the past hour and level of hunger, we would find a relationship. We would probably agree that the differences in the amount of food consumed were responsible for the difference in levels of hunger.

Unfortunately, even if one variable is the direct cause of another, we may not see a strong association. For example, even though intercourse is the direct cause of pregnancy, the relationship between having intercourse and getting pregnant is not strong; most occurrences of intercourse do not result in pregnancy.

Reason 2: The response variable is causing a change in the explanatory variable. Sometimes the causal connection is the opposite of what might be expected. For example, what do you think you would find if you studied hotels and defined the response variable as the hotel's occupancy rate and the explanatory variable as advertising sales (in dollars) per room? You would probably expect that higher advertising expenditures would cause higher occupancy rates. Instead, it turns out that the relationship is negative because, when occupancy rates are low, hotels spend more money on advertising to try to raise them. Thus, although we might expect higher advertising dollars to cause higher occupancy rates, if they are measured at the same point in time, we instead find that low occupancy rates cause higher advertising revenues.

Reason 3: The explanatory variable is a contributing but not sole cause of the response variable. The complex kinds of phenomena most often studied by researchers are likely to have multiple causes. Even if there were a causal connection between diet and a type of cancer, for instance, it would be unlikely that the cancer was caused solely by eating that certain type of diet. It is particularly easy to be misled into thinking you have found a sole cause for a particular outcome, when what you have found is actually a *necessary contributor* to the outcome. For example, scientists generally agree that in order to have AIDS, you must be infected with HIV. In other words, HIV is *necessary* to develop AIDS. But it does not follow that HIV is the *sole* cause of AIDS, and there has been some controversy over whether that is actually the case.

Another possibility, discussed in earlier chapters, is that one variable is a contributory cause of another, but only for a subgroup of the population. If the researchers do not examine separate subgroups, that fact can be masked, as the next example demonstrates.

EXAMPLE 11.7 **Delivery Complications, Rejection, and Violent Crime**

A study summarized in *Science* (Mann, March 1994) and conducted by scientists at the University of Southern California reported a relationship between violent crime and complications during birth. The researchers found that delivery complications at birth were associated with much higher incidence of violent crime later in life. The data came from an observational study of males born in Copenhagen, Denmark, between 1959 and 1961.

However, the connection held only for those men whose mothers rejected them. Rejection meant that the mother had not wanted the pregnancy, had tried to have the fetus aborted, and had sent the baby to an institution for at least a third of his first year of life. Men who were accepted by their mothers did not exhibit this relationship. Men who were rejected by their mothers but for whom there were no complications at birth did not exhibit the relationship either. In other words, it was the interaction of delivery complications and maternal rejection that was associated with higher levels of violent crime.

This example was based on an observational study, so there may not be a causal link at all. However, even if there is a causal connection between delivery complications and subsequent violent crime, the data suggest that it holds only for a particular

subset of the population. If the researchers had not measured the additional variable of maternal rejection, the data would have erroneously been interpreted as suggesting that the connection held for all men. ■

Reason 4: Confounding variables may be responsible for the observed relationship. We defined confounding variables in Chapter 5, but it is worth reviewing the concept here because it is relevant for explaining relationships. Remember that a confounding variable is one that has two properties. First, a confounding variable is related to the explanatory variable in the sense that individuals who differ for the explanatory variable are also likely to differ for the confounding variable. Second, a confounding variable affects the response variable. Thus, both the explanatory and one or more confounding variables may help cause the change in the response variable, but there is no way to establish how much is due to the explanatory variable and how much is due to the confounding variables. Example 5 in this chapter illustrates the point with several possibilities for confounding variables. For instance, people with differing levels of happiness (the explanatory variable) may have differing levels of emotional support, and emotional support affects one's will to live. Thus, emotional support is a confounding variable for the relationship between happiness and length of life.

Reason 5: Both variables may result from a common cause. We have seen numerous examples in which a change in one variable was thought to be associated with a change in the other, but for which we speculated that a third variable was responsible. For example, a study by Glaser et al. (1992) found that meditators had levels of an enzyme normally associated with people of a younger age. We could speculate that something in the personality of the meditators caused them to want to meditate and also caused them to have lower enzyme levels than others of the same age.

As another example, recall the scatterplot and correlation between verbal SAT scores and college GPAs, exhibited in Chapters 9 and 10. We would certainly not conclude that higher SAT scores caused higher grades in college, except perhaps for a slight benefit of boosted self-esteem. However, we could probably agree that the causes responsible for one variable being high (or low) are the same as those responsible for the other being high (or low). Those causes would include such factors as intelligence, motivation, and ability to perform well on tests.

EXAMPLE 11.8 **Do Smarter Parents Have Heavier Babies?**

News Story 18 in the Appendix describes a study that found for babies in the normal birth weight range, there was a relationship between birth weight and intelligence in childhood and early adulthood. The study was based on a cohort of about 3900 babies born in Britain in 1946. But there is a genetic component to intelligence, so smarter parents are likely to have smarter offspring. The researchers did include mother's education and father's social class in the analysis, to rule them out as possible confounding variables. However, there are many other variables that may contribute to birth weight, such as mother's diet and alcohol consumption, for which smarter parents may have provided more favorable conditions. Thus, it's possible that heavier birth weight and higher intelligence in the child both result from a common cause, such as parents' intelligence. ■

Reason 6: Both variables are changing over time. Some of the most nonsensical associations result from correlating two variables that have both changed over time. If each one is steadily increasing or decreasing across time, you will indeed see a strong correlation, but it may not have any causal link. For example, you would certainly see a correlation between winning times in two different Olympic events because winning times have all decreased over the years.

Sociological variables are the ones most likely to be manipulated in this way, as demonstrated by the next example, relating decreasing marriage rates and increasing life expectancy. Watch out for reports of a strong association between two such variables, especially when you know that both variables are likely to have consistently changed over time.

EXAMPLE 11.9 **Marriage Rates and Life Expectancy**

Table 11.4 shows the marriage rate (marriages per 1000 people) and the life expectancy in the United States for the years 2000 to 2011, and Figure 11.5 (next page) shows a scatterplot of these numbers with the least squares regression line superimposed. There is a very strong negative correlation between these two sets of numbers, at –0.97. Does that mean that avoiding marriage causes people to live longer? No! The explanation is that marriage rates have been declining across time for various social, political, and religious reasons, and life expectancy has been increasing across time due to advanced health care, safety programs, and lifestyle decisions. In fact, both variables are more highly correlated with year than they are with each other, with a correlation of 0.98 for year with life expectancy and –0.984 for year with marriage rate. Any two variables that both change across time will display a correlation with each other. ■

TABLE 11.4 Marriage Rates and Life Expectancy in the United States, 2000 to 2011

Year	Marriage Rate (per 1000)	Life Expectancy
2000	8.2	76.8
2001	8.2	76.9
2002	8.0	76.9
2003	7.7	77.1
2004	7.8	77.5
2005	7.6	77.4
2006	7.5	77.7
2007	7.3	77.9
2008	7.1	78.0
2009	6.8	78.5
2010	6.8	78.3
2011	6.8	78.7

Source: Life expectancy, http://www.census.gov/compendia/statab/2012/tables/12s0104.pdf Marriage rates, http://www.cdc.gov/nchs/nvss/marriage_divorce_tables.htm

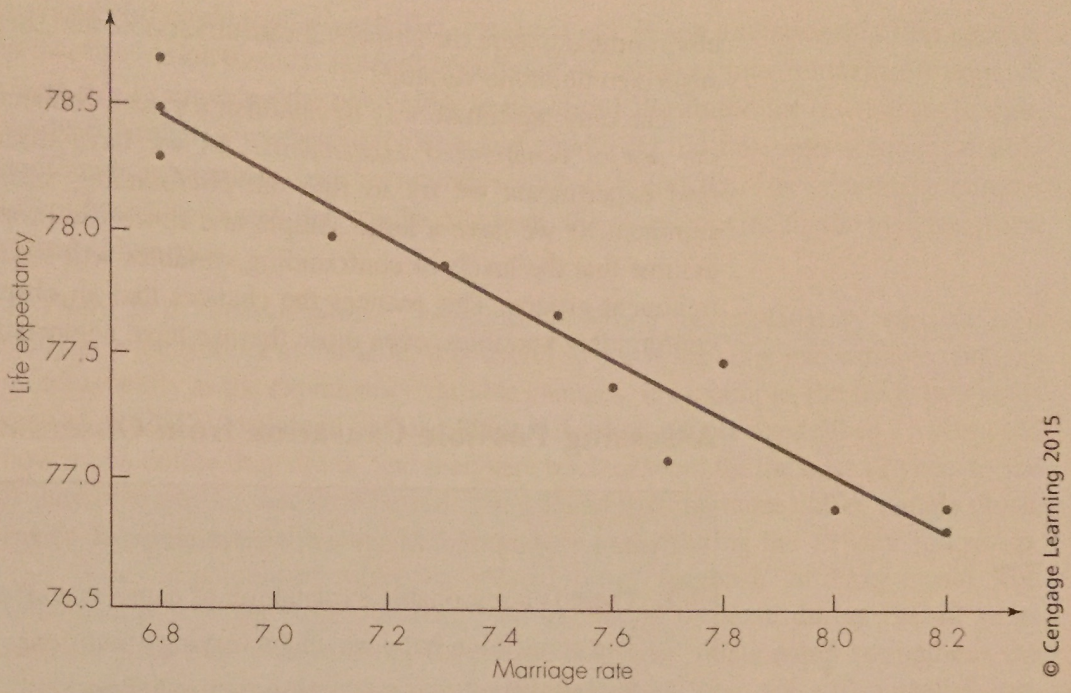


Figure 11.5
Marriage rate versus life expectancy in the United States, 2000 to 2011

Reason 7: The association may be nothing more than coincidence. Sometimes an association between two variables is nothing more than coincidence, even though the odds of it happening appear to be very small. For example, suppose a new office building opened, and within a year, there was an unusually high rate of brain cancer among workers in the building. Suppose someone calculated that the odds of having that many cases in one building were only 1 in 10,000. We might immediately suspect that something wrong in the environment was causing people to develop brain cancer.

The problem with this reasoning is that it focuses on the odds of seeing such a rare event occurring in that particular building in that particular city. It fails to take into account the fact that there are thousands of new office buildings. If the odds really were only 1 in 10,000, we should expect to see this phenomenon just by chance in about 1 of every 10,000 buildings. And that would just be for this particular type of cancer. What about clusters of other types of cancer or other diseases? It would be unusual if we did not occasionally see clusters of diseases as chance occurrences.

We will study this phenomenon in more detail in Part 3. For now, be aware that a connection of this sort should be expected to occur relatively often, even though each individual case has low probability.

11.4 Confirming Causation

Given the number of possible explanations for the relationship between two variables, how do we ever establish that there actually is a causal connection? It isn't easy. Ideally, in establishing a causal connection, we would change nothing in the

environment except the suspected causal variable and then measure the result on the suspected outcome variable.

The only legitimate way to establish a causal connection statistically is *through the use of randomized experiments.* As we have discussed earlier, in randomized experiments we try to rule out confounding variables through random assignment. If we have a large sample and if we use proper randomization, we can assume that the levels of confounding variables will be about equal in the different treatment groups. This reduces the chances that an observed association is due to confounding variables, even those that we have neglected to measure.

Assessing Possible Causation from Observational Studies

Evidence of a possible causal connection exists when

1. There is a reasonable explanation of cause and effect.
2. The connection happens under varying conditions.
3. Potential confounding variables are ruled out.
4. There is a "dose-response" relationship.

If a randomized experiment cannot be done, then nonstatistical considerations must be used to determine whether a causal link is reasonable. Following are some features that lend evidence to a causal connection:

1. *There is a reasonable explanation of cause and effect.* A potential causal connection will be more believable if an explanation exists for how the cause and effect occur. For instance, in Example 11.4 in this chapter, we established that for hardcover books, the number of pages is correlated with the price. We would probably not contend that higher prices result in more pages, but we could reasonably argue that more pages result in higher prices. We can imagine that publishers set the price of a book based on the cost of producing it and that the more pages there are, the higher the cost of production. Thus, we have a reasonable explanation for how an increase in the length of a book could cause an increase in the price.

2. *The connection happens under varying conditions.* If many observational studies conducted under different conditions all find the same link between two variables, the evidence for a causal connection is strengthened. This is especially true if the studies are not likely to have the same confounding variables. The evidence is also strengthened if the same type of relationship holds when the explanatory variable falls into different ranges.

For example, numerous observational studies have related cigarette smoking and lung cancer. Further, the studies have shown that the higher the number of cigarettes smoked, the greater the chances of developing lung cancer; similarly, a connection has been established between lung cancer and the age at which smoking began. These facts make it more plausible that smoking actually causes lung cancer.

3. *Potential confounding variables are ruled out.* When a relationship first appears in an observational study, potential confounding variables may immediately come to mind. For example, as the news story in Figure 6.1 illustrates, the researchers in Case Study 6.4 relating mother's smoking and child's IQ did take into account possible confounding variables such as mother's education and IQ. The greater the number of confounding factors that can be ruled out, the more convincing the evidence for a causal connection.

4. *There is a "dose-response" relationship.* When the explanatory variable in a study is a measurement variable, it is useful to see if the response variable changes systematically as the explanatory variable changes. For example, the study by Freedman et al (2012) explained in Case Study 6.5 asked people aged 50 to 71 years old how much coffee they drank, and then kept track of them for the next 12 years to see if they died during that time period. They found that the more coffee people drank (up to 5 cups a day), the less likely they were to die during the 12 year follow-up. The "dose" of amount of coffee was related to the "response" of dying or not. The relationship between coffee drinking and likelihood of death during the 12 years held for both men and women. This "dose-response" relationship strengthens the likelihood that there is a causal connection, although there are still potential confounding variables that could explain the relationship. For instance, people seem to have a genetic predisposition to metabolize caffeine at differing rates, and perhaps that is related to both amount of coffee they drink and age of death.

A Final Note

As you should realize by now, it is very difficult to establish a causal connection between two variables by using anything except randomized experiments. Because it is virtually impossible to conduct a flawless experiment, potential problems crop up even with a well-designed experiment. This means that you should look with skepticism on claims of causal connections. Having read this chapter, you should have the tools necessary for making intelligent decisions and for discovering when an erroneous claim is being made.

Thinking About Key Concepts

- An *outlier* in a scatterplot that fits the pattern of the rest of the data will increase the correlation. An outlier that does not fit the pattern of the rest of the data can substantially decrease the correlation, depending on where it falls.
- *Combining two or more groups* in a scatterplot can distort the relationship between the explanatory and response variable. It is better to use a separate symbol in the plot to identify group membership.
- There are many reasons that two variables could be related other than a direct cause-and-effect connection. The most common reason is that there are *confounding variables* that are related to the explanatory variable and affect the response variable.