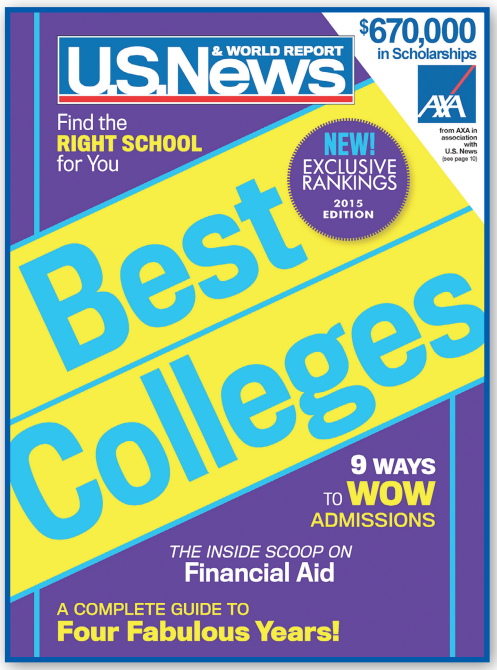
Describing Relationships: Scatterplots and Correlation

14



CASE STUDY The news media have a weakness for lists. Best places to live, best colleges, healthiest foods, worst-dressed women … a list of best or worst is sure to find a place in the news. When the state-by-state SAT scores come out each year, it's therefore no surprise that we find news articles ranking the states from best (North Dakota in 2014) to worst (District of Columbia in 2014) according to the average Mathematics SAT score achieved by their high school seniors. Unfortunately, such reports leave readers believing that schools in the District of Columbia must be much worse than those in North Dakota. Where does your home state rank? And do you believe the ranking reflects the quality of education you received?

The College Board, which sponsors the SAT exams, doesn't like this practice at all. "Comparing or ranking states on the basis of SAT scores alone is invalid and strongly discouraged by the College Board," says the heading on their table of state average SAT scores. To see why, let's look at the data.

Figure 14.1 shows the distribution of average scores on the SAT Mathematics exam for the 50 states and the District of Columbia. North Dakota leads at 620, and the District of Columbia trails at 438 on the SAT scale of 200 to 800. The distribution has an unusual shape: it has one clear peak and perhaps a second, small one. This may be a clue that the data mix two distinct groups. But, we need to explore the data further to be sure that this is the case.

In this chapter, we will learn that to understand one variable, such as SAT scores, we must look at how it is related to other variables. By the end of this chapter, you will be able to use what you have learned to understand why Figure 14.1 has such an unusual shape and to appreciate why the College Board discourages ranking states on SAT scores alone.

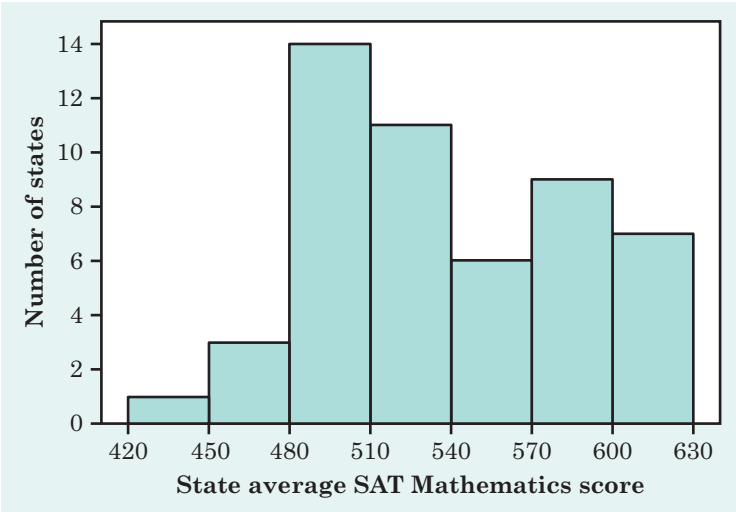


Figure 14.1 Histogram of the average scores of students in the 50 states and the District of Columbia on the SAT Mathematics exam.

A medical study finds that short women are more likely to have heart attacks than women of average height, while tall women have the fewest heart attacks. An insurance group reports that heavier cars are involved in fewer fatal accidents per 10,000 vehicles registered than are lighter cars. These and many other statistical studies look at the relationship between two variables. To understand such a relationship, we must often examine other variables as well. To conclude that shorter women have higher risk from heart attacks, for example, the researchers had to eliminate the effect of other variables such as weight and exercise habits. Our topic in this and the following chapters is relationships between variables. One of our main themes is that the relationship between two variables can be strongly influenced by other variables that are lurking in the background.

Most statistical studies examine data on more than one variable. Fortunately, statistical analysis of several-variable data builds on the tools we used to examine individual variables. The principles that guide our work also remain the same:

- First plot the data, then add numerical summaries.
- Look for overall patterns and deviations from those patterns.
- When the overall pattern is quite regular, there is sometimes a way to describe it very briefly.

Scatterplots

The most common way to display the relation between two quantitative variables is a *scatterplot*.

EXAMPLE 1 The Big Bang

NASA/JPL-Caltech/University of Arizona/STScI

How did the universe begin? One popular theory is known as the "Big Bang." The universe began with a big bang and matter expanded outward, like a balloon inflating. If the Big Bang theory is correct, galaxies farthest away from the origin of the bang must be moving faster than those closest to the origin. This also means that galaxies close to the earth must be moving at a similar speed to that of earth, and galaxies far from earth must be moving at very different speeds from earth. Hence, relative to earth, the farther away a galaxy is, the faster it appears to be moving away from earth. Are data consistent with this theory? The answer is Yes.

In 1929, Edwin Hubble investigated the relationship between the distance from the earth and the recession velocity (the speed at which an object is moving away from an observer) of galaxies. Using data he had collected, Hubble estimated the distance, in megaparsecs, from the earth to 24 galaxies. One parsec equals 3.26 light-years (the distance light travels in one year), and a megaparsec is one million parsecs. The recession velocities, in kilometers per second, of the galaxies were also measured. Figure 14.2 is a scatterplot that shows how recession velocity is related to distance from the earth. We think that "distance from the earth" will help explain "recession velocity." That is, "distance from the earth" is the *explanatory variable*, and "recession velocity" is the *response variable*. We want to see how recession velocity changes when distance from the earth changes, so we put distance from the earth (the explanatory variable) on the horizontal axis. We can then see that, as distance from the earth goes up, recession velocity goes up. Each point on the plot represents one galaxy. For example, the point with a different plotting symbol corresponds to a galaxy that is 1.7 megaparsecs from the earth and that has a recession velocity of 960 kilometers per second.

Hubble's discovery turned out to be one of the most important discoveries in all of astronomy. The data helped establish Hubble's law, which is recession velocity $= H_0 \times$ Distance, where H_0 is the value known as the Hubble constant. Hubble's law says that the apparent recession velocities of galaxies are directly proportional to their distances. This relationship is the key evidence for the idea of the expanding universe, as suggested by the Big Bang.

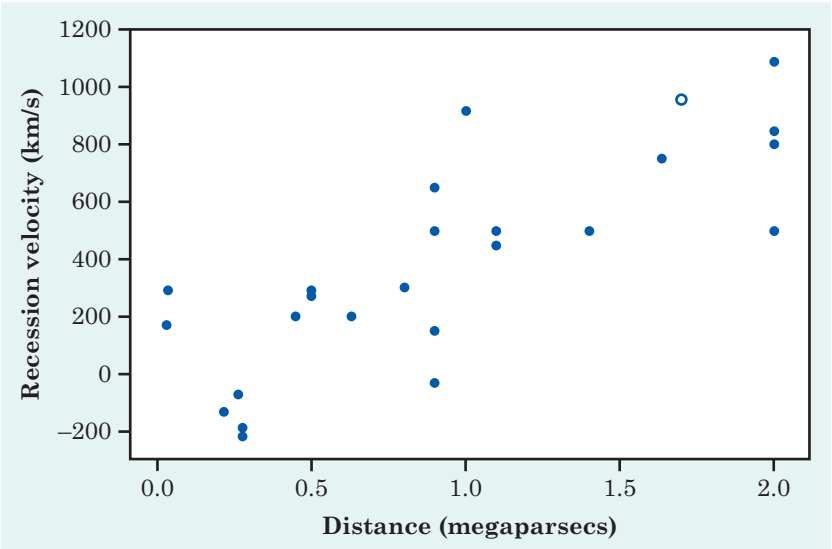


Figure 14.2 Scatterplot of recession velocity against distance from the earth, Example 1.

Scatterplot

A **scatterplot** shows the relationship between two quantitative variables measured on the same individuals. The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis. Each individual in the data appears as the point in the plot fixed by the values of both variables for that individual.

Always plot the explanatory variable, if there is one, on the horizontal axis (the x axis) of a scatterplot. As a reminder, we usually call the explanatory variable x and the response variable y. If there is no explanatory-response distinction, either variable can go on the horizontal axis.

EXAMPLE 2 Health and wealth

Figure 14.3 is a scatterplot of data from the World Bank for 2010. The individuals are all the world's nations for which data are available. The explanatory variable is a measure of how rich a country is: the gross domestic product (GDP) per capita. GDP is the total value of the goods and services produced in a country, converted into dollars. The response variable is life expectancy at birth.

We expect people in richer countries to live longer. The overall pattern of the scatterplot does show this, but the relationship has an interesting shape. Life expectancy tends to rise very quickly as GDP increases, then

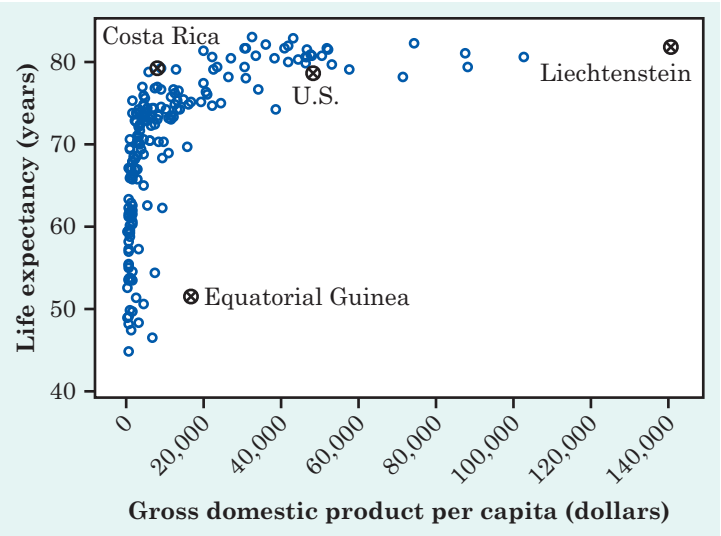


Figure 14.3 Scatterplot of the life expectancy of people in many nations against each nation's GDP per person, Example 2.

levels off. People in very rich countries such as the United States typically live no longer than people in poorer but not extremely poor nations. Some of these countries, such as Costa Rica, do almost as well as the United States.

Two nations are outliers. In one, Equatorial Guinea, life expectancies are similar to those of its neighbors but its GDP is higher. Equatorial Guinea produces oil. It may be that income from mineral exports goes mainly to a few people and so pulls up GDP per capita without much effect on either the income or the life expectancy of ordinary citizens. That is, GDP per person is a mean, and we know that mean income can be much higher than median income.

The other outlier is Liechtenstein, a tiny nation bordering Switzerland and Austria. Liechtenstein has a strong financial sector and is considered a tax haven.

NOW IT'S YOUR TURN

14.1 Brain size and intelligence. For centuries, people have associated intelligence with brain size. A recent study used magnetic resonance imaging to measure the brain size of several individuals. The IQ and brain size (in units of 10,000 pixels) of six individuals are as follows:

Brain size:	100	90	95	92	88	106
IQ:	140	90	100	135	80	103

Is there an explanatory variable? If so, what is it and what is the response variable? Make a scatterplot of these data.

Interpreting scatterplots

To interpret a scatterplot, apply the usual strategies of data analysis.

Examining a scatterplot

In any graph of data, look for the **overall pattern** and for striking **deviations** from that pattern.

You can describe the overall pattern of a scatterplot by the **direction**, **form**, and **strength** of the relationship.

An important kind of deviation is an **outlier**, an individual value that falls outside the overall pattern of the relationship.

Statistics in Your World

After you plot your data, think! Abraham Wald (1902–1950), like many statisticians, worked on war problems during World War II. Wald invented some statistical methods that were military secrets until the war ended. Here is one of his simpler ideas. Asked where extra armor should be added to airplanes, Wald studied the location of enemy bullet holes in planes returning from combat. He plotted the locations on an outline of the plane. As data accumulated, most of the outline filled up. Put the armor in the few spots with no bullet holes, said Wald. That's where bullets hit the planes that didn't make it back.

Both Figures 14.2 and 14.3 have a clear *direction*: recession velocity goes up as distance from the earth increases, and life expectancy generally goes up as GDP increases. We say that Figures 14.2 and 14.3 show a *positive association*. Figure 14.4 is a scatterplot of the gas mileages (in miles per gallon) and the engine size (or engine displacement, in liters) of 1252 model-year 2015 cars. The response variable is gas mileage and the explanatory variable is engine size. We see that gas mileage decreases as engine size goes up. We say that Figure 14.4 shows a *negative association*.

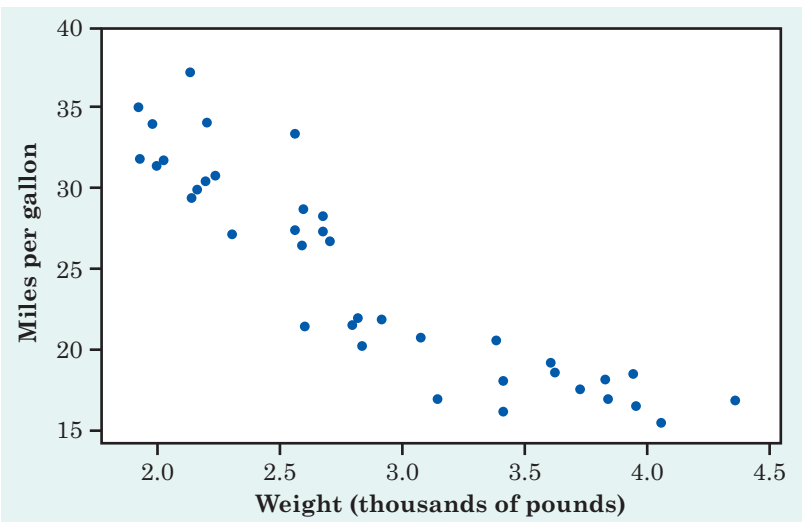


Figure 14.4 Scatterplot of miles per gallon against weight for 38 cars.

Positive association, negative association

Two variables are **positively associated** when above-average values of one tend to accompany above-average values of the other and below-average values also tend to occur together. The scatterplot slopes upward as we move from left to right.

Two variables are **negatively associated** when above-average values of one tend to accompany below-average values of the other, and vice versa. The scatterplot slopes downward from left to right.

Each of our scatterplots has a distinctive *form*. Figure 14.2 shows a roughly straight-line trend, and Figure 14.3 shows a *curved relationship*. Figure 14.4 shows a slightly curved relationship. The *strength* of a relationship in a scatterplot is determined by how closely the points follow a clear form. The relationships in Figures 14.2 and 14.3 are not strong. Galaxies with similar distances from the earth show quite a bit of scatter in their recession velocities, and nations with similar GDPs can have quite different life expectancies. The relationship in Figure 14.4 is moderately strong. Here is an example of a stronger relationship with a simple form.

EXAMPLE 3 Classifying fossils

James L. Amos/Science Source

Archaeopteryx is an extinct beast having feathers like a bird but teeth and a long bony tail like a reptile. Only six fossil specimens are known. Because these fossils differ greatly in size, some scientists think they are different species rather than individuals from the same species. We will examine data on the lengths in centimeters of the femur (a leg bone) and the humerus (a bone in the upper arm) for the five fossils that preserve both bones. Here are the data:

Femur:	38	56	59	64	74
Humerus:	41	63	70	72	84

Because there is no explanatory-response distinction, we can put either measurement on the x axis of a scatterplot. The plot appears in Figure 14.5.

The plot shows a *strong, positive, straight-line association*. The straight-line form is important because it is common and simple. The association is strong because the points lie close to a line. It is positive because

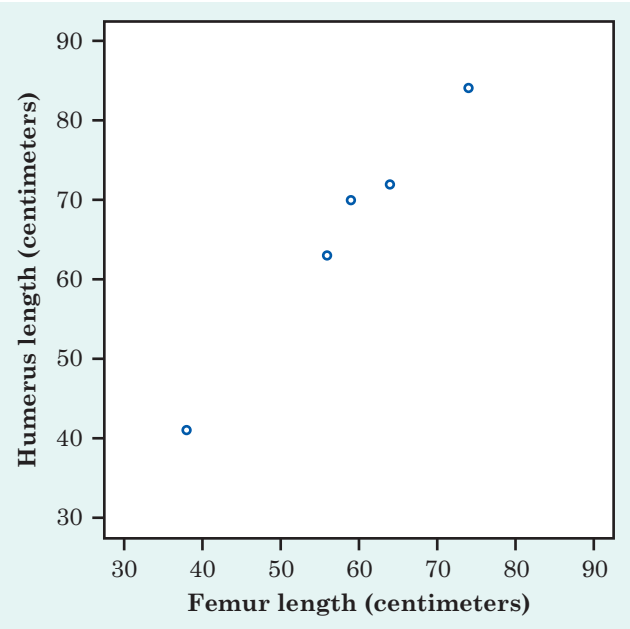


Figure 14.5 Scatterplot of the lengths of two bones in five fossil specimens of the extinct beast archaeopteryx, Example 3.

as the length of one bone increases, so does the length of the other bone. These data suggest that all five fossils belong to the same species and differ in size because some are younger than others. We expect that a different species would have a different relationship between the lengths of the two bones, so that it would appear as an outlier.

NOW IT'S YOUR TURN

14.2 Brain size and intelligence. For centuries, people have associated intelligence with brain size. A recent study used magnetic resonance imaging to measure the brain size of several individuals. The IQ and brain size (in units of 10,000 pixels) of six individuals are as follows:

Brain size:	100	90	95	92	88	106
IQ:	140	90	100	135	80	103

Make a scatterplot of these data if you have not already done so. What is the form, direction, and strength of the association? Are there any outliers?

Correlation

A scatterplot displays the direction, form, and strength of the relationship between two variables. Straight-line relations are particularly important because a straight line is a simple pattern that is quite common. A straight-line relation is strong if the points lie close to a straight line and weak if they are widely scattered about a line. Our eyes are not good judges of how strong a relationship is. The two scatterplots in Figure 14.6 depict the same data, but the right-hand plot is drawn smaller in a large field. The right-hand plot seems to show a stronger straight-line relationship. Our eyes can be fooled by changing the plotting scales or the amount of blank space around the cloud of points in a scatterplot. We need to follow our strategy for data analysis by using a numerical measure to supplement the graph. *Correlation* is the measure we use.

Correlation

The **correlation** describes the direction and strength of a straight-line relationship between two quantitative variables. Correlation is usually written as r.

Calculating a correlation takes a bit of work. You can usually think of r as the result of pushing a calculator button or giving a command in software and concentrate on understanding its properties and use. Knowing how we obtain r from data, however, does help us understand how correlation works, so here we go.

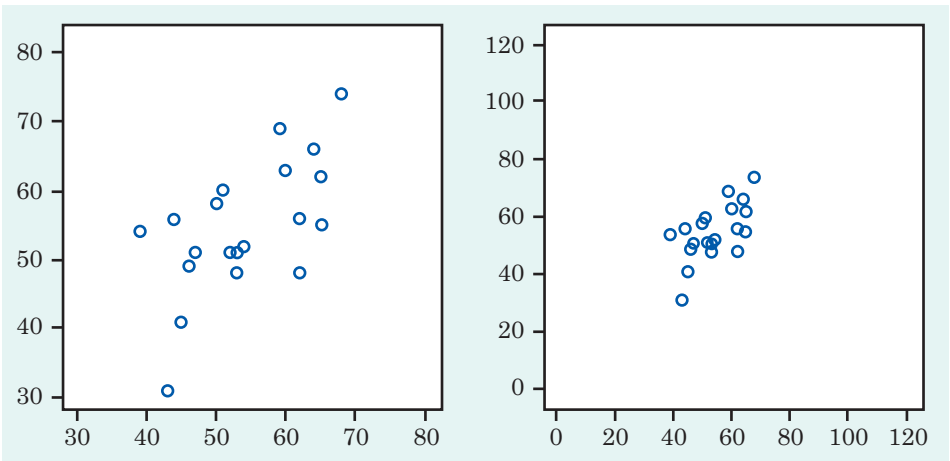


Figure 14.6 Two scatterplots of the same data. The right-hand plot suggests a stronger relationship between the variables because of the surrounding space.

EXAMPLE 4 Calculating correlation

We have data on two variables, x and y, for n individuals. For the fossil data in Example 3, x is femur length, y is humerus length, and we have data for $n = 5$ fossils.

Step 1. Find the mean and standard deviation for both x and y. For the fossil data, a calculator tells us that

Femur:	$\bar{x} = 58.2$ cm	$s_x = 13.20$ cm
Humerus:	$\bar{y} = 66.0$ cm	$s_y = 15.89$ cm

We use s_x and s_y to remind ourselves that there are two standard deviations, one for the values of x and the other for the values of y.

Step 2. Using the means and standard deviations from Step 1, find the standard scores for each x-value and for each y-value:

Value of x	Standard score $(x - \bar{x})/s_x$	Value of y	Standard score $(y - \bar{y})/s_y$
38	$(38 - 58.2)/13.20 = -1.530$	41	$(41 - 66.0)/15.89 = -1.573$
56	$(56 - 58.2)/13.20 = -0.167$	63	$(63 - 66.0)/15.89 = -0.189$
59	$(59 - 58.2)/13.20 = 0.061$	70	$(70 - 66.0)/15.89 = 0.252$
64	$(64 - 58.2)/13.20 = 0.439$	72	$(72 - 66.0)/15.89 = 0.378$
74	$(74 - 58.2)/13.20 = 1.197$	84	$(84 - 66.0)/15.89 = 1.133$

Step 3. The correlation is the average of the products of these standard scores. As with the standard deviation, we "average" by dividing by $n - 1$, one fewer than the number of individuals:

$$r = \frac{1}{4}\big[(-1.530)(-1.573) + (-0.167)(-0.189) + (0.061)(0.252) + (0.439)(0.378) + (1.197)(1.133)\big]$$

$$= \frac{1}{4}(2.4067 + 0.0316 + 0.0154 + 0.1659 + 1.3562)$$

$$= \frac{3.9758}{4} = 0.994$$

The algebraic shorthand for the set of calculations in Example 4 is

$$r = \frac{1}{n-1}\sum\left(\frac{x - \bar{x}}{s_x}\right)\left(\frac{y - \bar{y}}{s_y}\right)$$

The symbol $\sum$, called "sigma," means "add them all up."

Understanding correlation

More important than calculating r (a task for technology) is understanding how correlation measures association. Here are the facts:

- **Positive r indicates positive association between the variables, and negative r indicates negative association**. The scatterplot in Figure 14.5 shows strong positive association between femur length and humerus length. In three fossils, both bones are longer than their average values, so their standard scores are positive for both x and y. In the other two fossils, the bones are shorter than their averages, so both standard scores are negative. The products are all positive, giving a positive r.

- **The correlation r always falls between -1 and 1**. Values of r near 0 indicate a very weak straight-line relationship. The strength of the relationship increases as r moves away from 0 toward either -1 or 1. Values of r close to -1 or 1 indicate that the points lie close to a straight line. The extreme values $r = -1$ and $r = 1$ occur only when the points in a scatterplot lie exactly along a straight line.

  The result $r = 0.994$ in Example 4 reflects the strong positive straight-line pattern in Figure 14.5. The scatterplots in Figure 14.7 illustrate how r measures both the direction and the strength of a straight-line relationship. Study them carefully. Note that the sign of r matches the direction of the slope in each plot, and that r approaches -1 or 1 as the pattern of the plot comes closer to a straight line.

- Because r uses the standard scores for the observations, **the correlation does not change when we change the units of measurement** of x, y, or both. Measuring length in inches rather than centimeters in Example 4 would not change the correlation $r = 0.994$.

  Our descriptive measures for one variable all share the same units as the original observations. If we measure length in centimeters, the median, quartiles, mean, and standard deviation are all in centimeters. The correlation between two variables, however, has no unit of measurement; it is just a number between -1 and 1.

- **Correlation ignores the distinction between explanatory and response variables**. If we reverse our choice of which variable to call x and which to call y, the correlation does not change.

- **Correlation measures the strength of only straight-line association between two variables**. Correlation does not describe curved relationships between variables, no matter how strong they are.

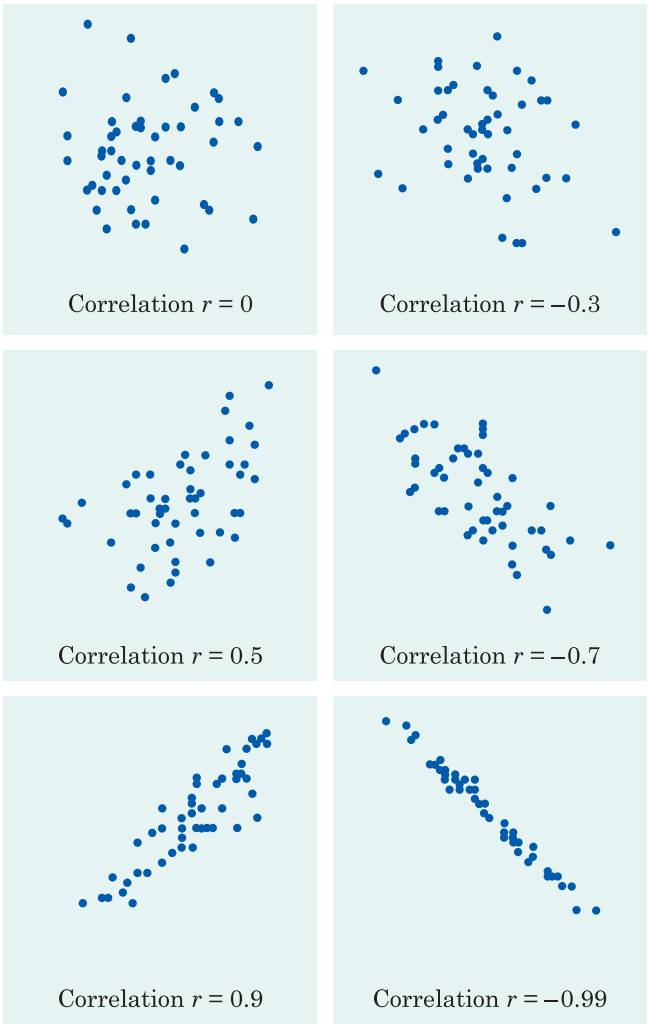


Figure 14.7 How correlation measures the strength of a straight-line relationship. Patterns closer to a straight line have correlations closer to 1 or –1.

- Like the mean and standard deviation, **the correlation is strongly affected by a few outlying observations**. Use r with caution when outliers appear in the scatterplot. Look, for example, at Figure 14.8. We changed the femur length of the first fossil from 38 to 60 centimeters. Rather than falling in line with the other fossils, the first is now an outlier. The correlation drops from $r = 0.994$ for the original data to $r = 0.640$.

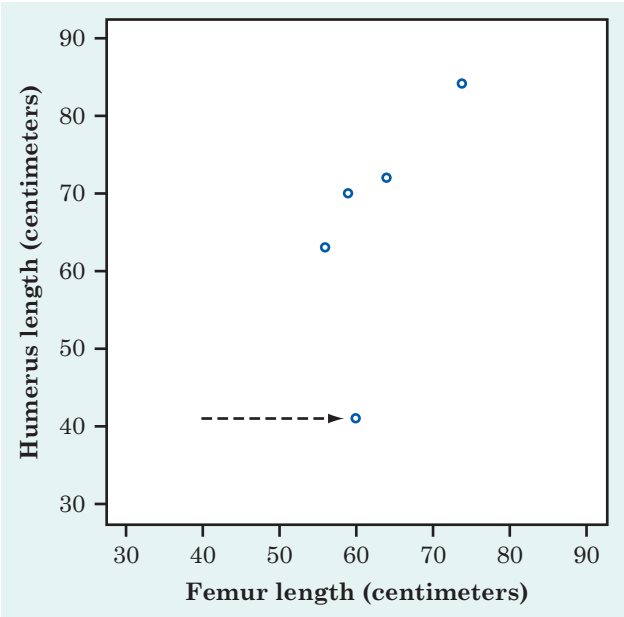


Figure 14.8 Moving one point reduces the correlation from $r = 0.994$ to $r = 0.640$.

NOW IT'S YOUR TURN

14.3 Brain size and intelligence. For centuries, people have associated intelligence with brain size. A recent study used magnetic resonance imaging to measure the brain size of several individuals. The IQ and brain size (in units of 10,000 pixels) of six individuals are as follows:

Brain size:	100	90	95	92	88	106
IQ:	140	90	100	135	80	103

Make a scatterplot of these data if you have not already done so. Compare your plot with those in Figure 14.7. What would you estimate the correlation r to be?

There are many kinds of relationships between variables and many ways to measure them. Although correlation is very common, remember its limitations. Correlation makes sense only for quantitative variables—we can speak of the relationship between the sex of voters and the political party they prefer, but not of the correlation between these variables. Even for quantitative variables such as the length of bones, correlation measures only straight-line association.

Remember also that correlation is not a complete description of two-variable data, even when there is a straight-line relationship between the variables. You should give the means and standard deviations of both x and y along with the correlation. Because the formula for correlation uses the means and standard deviations, these measures are the proper choice to accompany a correlation.

STATISTICS IN SUMMARY

Chapter Specifics

- A **scatterplot** is a graph of the relationship between two quantitative variables. If you have an explanatory and a response variable, put the explanatory variable on the x (horizontal) axis of the scatterplot.
- When you examine a scatterplot, look for the **direction**, **form**, and **strength** of the relationship and also for possible **outliers**.
- If there is a clear direction, is it positive (the scatterplot slopes upward from left to right) or negative (the plot slopes downward)?
- Is the form straight or curved? Are there clusters of observations? Is the relationship strong (a tight pattern in the plot) or weak (the points scatter widely)?
- The **correlation** r measures the direction and strength of a straight-line relationship between two quantitative variables.
- Correlation is a number between -1 and 1. The sign of r shows whether the association is positive or negative. The value of r gets closer to -1 or 1 as the points cluster more tightly about a straight line. The extreme values -1 and 1 occur only when the scatterplot shows a perfectly straight line.

Chapters 11, 12, and 13 discussed graphical and numerical summaries suitable for a single quantitative variable. In practice, most statistical studies examine relationships between two or more variables. In this chapter, we learn about scatterplots, a type of graph that displays the relationship between two quantitative variables, and correlation, a number that measures the direction and strength of a straight-line relationship between two quantitative variables.

As with other graphics and numbers that summarize data, scatterplots and correlations help us see what the data are telling us, in this case about the possible relationship between two quantitative variables. The goal is to draw conclusions about whether the relationships observed in our data are true in general. In the next chapter, we discuss this in detail.

CASE STUDY EVALUATED Figure 14.9 is a scatterplot of each state's average SAT Mathematics score and the proportion of that state's high school seniors who took the SAT exam. SAT score is the response variable, and the proportion of a state's high school seniors who took the SAT exam is the explanatory variable.

1. Describe the overall pattern in words. Is the association positive or negative? Is the relationship strong?
2. The plot shows two groups of states. In one group, no more than 20% took the SAT. In the other, at least 36% took the exam and the average scores tend to be lower. There are two common college entrance exams, the SAT and the ACT. In ACT states, only students applying to selective colleges take the SAT. Which group of states in the plot corresponds to states in which most students take the ACT exam?
3. Write a paragraph, in language that someone who knows no statistics would understand, explaining why comparing states on the basis of average SAT scores alone would be misleading as a way of comparing the quality of education in the states.

LaunchPad macmillan learning **Online Resources**

StatBoards video

- The ~~StatBoard Video~~ *Creating and Interpreting Scatterplots* describes how to create and interpret a scatterplot in the context of an example exploring the relationship between the cost of attending a baseball game and the team's performance.
- The ~~StatBoard Video~~ *Computing a Correlation* describes how to compute the correlation in the context of an example exploring the relationship between the cost of attending a baseball game and the team's performance.

StatBoards video

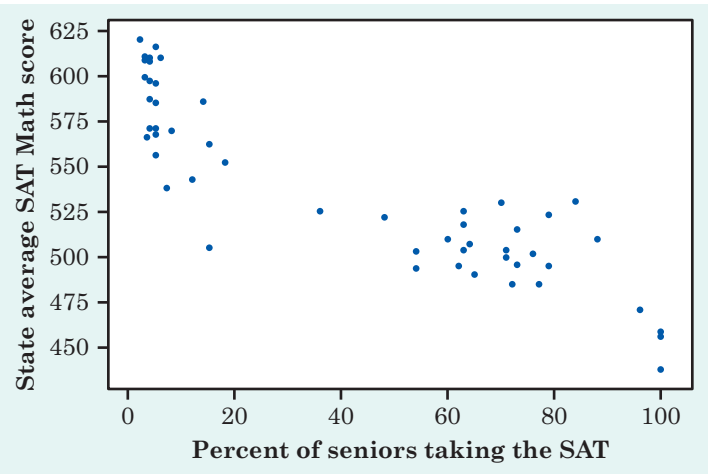


Figure 14.9 Scatterplot of average SAT Mathematics score for each state against the proportion of the state's high school seniors who took the SAT.

CHECK THE BASICS

For Exercise 14.1, see page 319; for Exercise 14.2, see page 322; for Exercise 14.3, see page 327.

14.4 Creating scatterplots. When creating a scatterplot,

(a) always put the categorical variable on the horizontal axis.

(b) always put the categorical variable on the vertical axis.

(c) if you have an explanatory variable, put it on the horizontal axis.

(d) if you have a response variable, put it on the horizontal axis.

14.5 Interpreting scatterplots. If the points in a scatterplot of two variables slope upward from left to right, we say the direction of the relationship between the variables is

(a) positive.

(b) negative.

(c) strong.

(d) weak.

14.6 Interpreting scatterplots. Which of the following patterns might one observe in a scatterplot?

(a) The points in the plot follow a curved pattern.

(b) The points in the plot group into different clusters.

(c) One or two points are clear outliers.

(d) All of the above.

14.7 Correlation. Which of the following is true of the correlation r?

(a) It cannot be greater than 1 or less than -1.

(b) It measures the strength of the *straight-line* relationship between two quantitative variables.

(c) A correlation of $+1$ or -1 can only happen if there is a perfect straight-line relationship between two quantitative variables.

(d) All of the above.

14.8 Correlation and scatterplots. If the points in a scatterplot are very tightly clustered around a straight line, the correlation must be

(a) close to 0.

(b) close to $+1$.

(c) close to -1.

(d) close to either $+1$ or -1.

CHAPTER 14 EXERCISES

14.9 What number can I be?

(a) What are all the values that a correlation r can possibly take?

(b) What are all the values that a standard deviation s can possibly take?

(c) What are all the values that a mean $\bar{x}$ can possibly take?

14.10 Measuring mice. For a biology project, you measure the tail length (millimeters) and weight (grams) of 10 mice.

(a) Explain why you expect the correlation between tail length and weight to be positive.

(b) If you measured tail length in centimeters, how would the correlation change?

14.11 Living on campus. A February 2, 2008, article in the *Columbus Dispatch* reported a study on the distances students lived

from campus and average GPA. Here is a summary of the results:

Residence	Avg. GPA
Residence hall	3.33
Walking distance	3.16
Near campus, long walk or short drive	3.12
Within the county, not near campus	2.97
Outside the county	2.94

Based on these data, is the association between the distance a student lives from campus and average GPA positive, negative, or near 0?

14.12 The endangered manatee. Manatees are large, gentle, slow-moving creatures found along the coast of Florida. Many manatees are injured or killed by boats. Figure 14.10 is a scatterplot of the number of manatee deaths by boats versus the number of boats registered in Florida (in thousands) for the years between 1977 and 2014.

(a) Describe the overall pattern of the relationship in words.

(b) About what are the number of boats registered and manatee deaths for point A?

(c) Suppose there was a point near B. Would this be an outlier? If so, say how it is unusual (for example, "a moderately high number of deaths but a low number of boats registered").

AU: Please check proofreading edit here.

14.13 Calories and salt in hot dogs. Figure 14.11 shows the calories and sodium content in 17 brands of meat hot dogs. Describe the overall pattern of these data. In what way is the point marked A unusual?

14.14 The endangered manatee. Is the correlation r for the data in Figure 14.10 near -1, clearly negative but not near -1, near 0, clearly positive but not near 1, or near 1? Explain your answer.

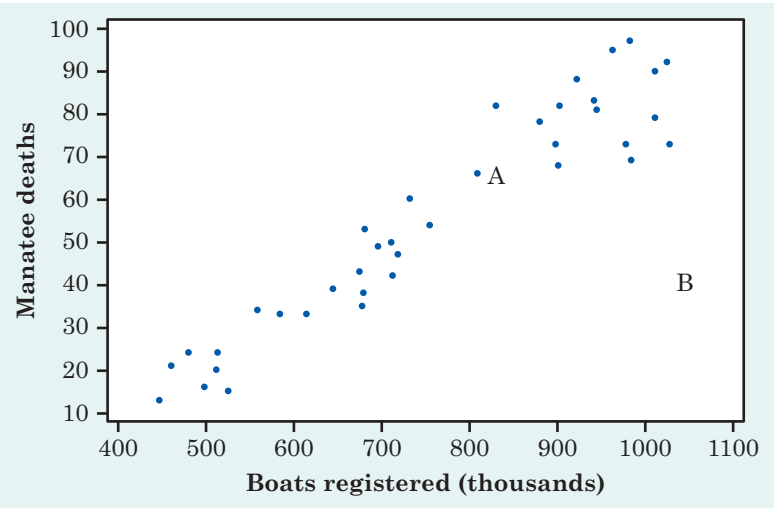


Figure 14.10 Manatee deaths by boats and boats registered in Florida, Exercises 14.12 and 14.14.

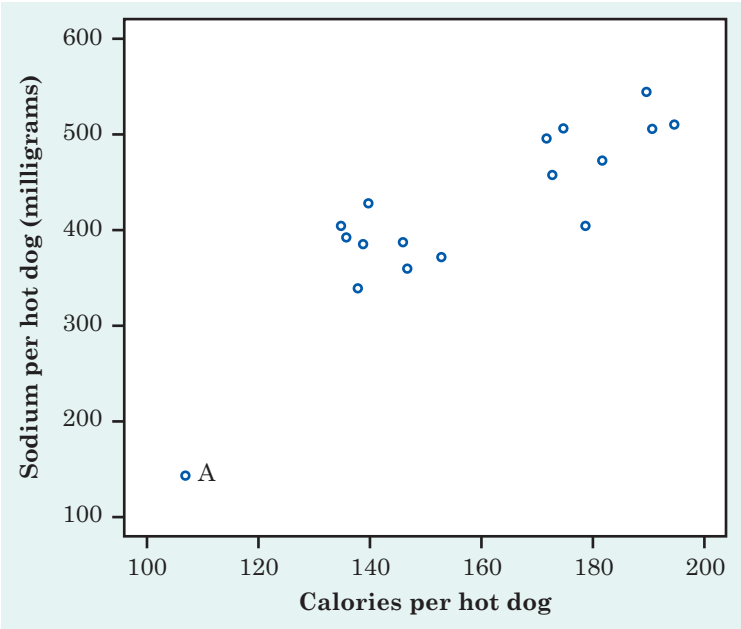


Figure 14.11 Calories and sodium content for 17 brands of meat hot dogs, Exercises 14.13 and 14.15.

14.15 Calories and salt in hot dogs. Is the correlation r for the data in Figure 14.11 near -1, clearly negative but not near -1, near 0, clearly positive but not near 1, or near 1? Explain your answer.

14.16 Comparing correlations. Which of Figures 14.2, 14.10, and 14.11 has a correlation closer to 0? Explain your answer.

14.17 Outliers and correlation. In Figure 14.11, the point marked A is an outlier. Will removing the outlier *increase* or *decrease* r? Why?

14.18 Global warming. Have average global temperatures been increasing in recent years? Here are annual average global temperatures for the last 21 years in degrees Celsius.

Year	1994	1995	1996
Temperature	14.23	14.35	14.22
Year	1997	1998	1999
Temperature	14.42	14.54	14.36
Year	2000	2001	2002
Temperature	14.33	14.45	14.51
Year	2003	2004	2005
Temperature	14.52	14.48	14.55
Year	2006	2007	2008
Temperature	14.50	14.49	14.41
Year	2009	2010	2011
Temperature	14.50	14.56	14.43
Year	2012	2013	2014
Temperature	14.48	14.52	14.59

(a) Make a scatterplot. (Which is the explanatory variable?)

(b) Is the association between these variables positive or negative? Explain why you expect the relationship to have this direction.

(c) Describe the form and strength of the relationship.

14.19 Death by intent. Homicide and suicide are both intentional means of ending a life. However, the reason for committing a homicide is different than that for suicide, and we might expect homicide and suicide rates to be uncorrelated. On the other hand, both can involve some degree of violence, so perhaps we might expect some level of correlation in the rates. Table 14.1 gives data from 2008–2010 for 26 counties in Ohio. Rates are per 100,000 people. The data also indicate that the homicide rates for some counties should be treated with caution because of low counts (Y = Yes, treat with caution, and N = No, do not treat with caution).

(a) Make a scatterplot of the data for the counties for which the data do not need to be treated with caution. Use homicide rate as the explanatory variable.

(b) Is the association between these variables positive or negative? What is the form of the relationship? How strong is the relationship?

(c) Now add the data for the counties for which the data do need to be treated with caution to your graph, using a different color or a different plotting symbol. Does the pattern of the relationship that you observed in part (b) hold for the counties for which the data do need to be treated with caution also?

TABLE 14.1 Homicide and suicide rates per 100,000 people

County	Homicide rate	Suicide rate	Caution	County	Homicide rate	Suicide rate	Caution
Allen	4.2	9.2	Y	Lorrain	3.1	11.0	Y
Ashtabula	1.8	15.5	Y	Lucas	7.4	13.3	N
Butler	2.6	12.7	Y	Mahoning	10.9	12.4	N
Clermont	1.0	16.0	Y	Medina	0.5	10.0	Y
Clark	5.6	14.5	N	Miami	2.6	9.2	Y
Columbiana	3.5	16.6	N	Montgomery	9.5	15.2	N
Cuyahoga	9.2	9.5	N	Portage	1.6	9.6	Y
Delaware	0.8	7.6	Y	Stark	4.7	13.5	N
Franklin	8.7	11.4	N	Summit	4.9	11.5	N
Greene	2.7	12.8	Y	Trumbull	5.8	16.6	N
Hamilton	8.9	10.8	N	Warren	0.7	11.3	Y
Lake	1.8	11.3	Y	Wayne	1.8	8.9	Y
Licking	4.5	12.9	N	Wood	1.0	7.4	Y

14.20 Marriage. Suppose that men always married women three years younger than themselves. Draw a scatterplot of the ages of six married couples, with the husband's age as the explanatory variable. What is the correlation r for your data? Why?

14.21 Stretching a scatterplot. Changing the units of measurement can greatly alter the appearance of a scatterplot. Return to the fossil data from Example 3:

Femur:	38	56	59	64	74
Humerus:	41	63	70	72	84

These measurements are in centimeters. Suppose a deranged scientist measured the femur in meters and the humerus in millimeters. The data would then be

Femur:	0.38	0.56	0.59	0.64	0.74
Humerus:	410	630	700	720	840

(a) Draw an x axis extending from 0 to 75 and a y axis extending from 0 to 850. Plot the original data on these axes. Then plot the new data on the same axes in a different color. The two plots look very different.

(b) Nonetheless, the correlation is exactly the same for the two sets of measurements. Why do you know that this is true without doing any calculations?

14.22 Global warming. Exercise 14.18 gives data on the average global temperatures, in degrees Celsius, for the years 1994 to 2014.

(a) Use a calculator to find the correlation r. Explain from looking at the scatterplot why this value of r is reasonable.

(b) Suppose that the temperatures had been recorded in degrees Fahrenheit. For example, the 1994 temperature of 14.23°C would be 57.61°F. How would the value of r change?

14.23 Death by intent. Table 14.1 gives data on on homicide and suicide rates from 2008–2010 for 26 counties in Ohio. The homicide rates for 14 of the counties should be treated with caution because of low counts. You made a scatterplot of these data in Exercise 14.19.

(a) Do you think the correlation will be about the same for the counties for which the data do need to be treated with caution and for the counties for which the data do not need to be treated with caution, or quite different for the two groups? Why?

(b) Calculate r for the counties for which the data do need to be treated with caution alone and also for for the counties for which the data do not need to be treated with caution alone. (Use your calculator.)

14.24 Strong association but no correlation. The gas mileage of an automobile first increases and then decreases as the speed increases. Suppose that this relationship is very regular, as shown by the following data on speed (miles per hour) and mileage (miles per gallon):

Speed:	25	35	45	55	65
Mileage:	20	24	26	24	20

Make a scatterplot of mileage versus speed. Use a calculator to show that the correlation between speed and mileage is $r = 0$. Explain why the correlation is 0 even though there is a strong relationship between speed and mileage.

14.25 Death by intent. The data in Table 14.1 are given in deaths per 100,000 people. If we changed the data from deaths per 100,000 people to deaths per 1,000 people how would the rates change? How would the correlation between homicide and suicide rates change?

14.26 What are the units? How sensitive to changes in water temperature are coral reefs? To find out, measure the growth of corals in aquariums (where growth is the change in weight, in pounds, of the coral before and after the experiment) when the water temperature (in degrees Fahrenheit) is controlled at different levels. In what units are each of the following descriptive statistics measured?

(a) the mean growth of the coral

(b) the standard deviation of the growth of the coral

(c) the correlation between weight gain and temperature

(d) the median growth of the coral

14.27 Teaching and research. A college newspaper interviews a psychologist about student ratings of the teaching of faculty members. The psychologist says, "The evidence indicates that the correlation between the research productivity and teaching rating of faculty members is close to zero." The paper reports this as "Professor McDaniel said that good researchers tend to be poor teachers, and vice versa." Explain why the paper's report is wrong. Write a statement in plain language (don't use the word "correlation") to explain the psychologist's meaning.

14.28 Sloppy writing about correlation. Each of the following statements contains a blunder. Explain in each case what is wrong.

(a) "There is a high correlation between the manufacturer of a car and the gas mileage of the car."

(b) "We found a high correlation ($r = 1.09$) between the horsepower of a car and the gas mileage of the car."

(c) "The correlation between the weight of a car and the gas mileage of the car was found to be $r = 0.53$ miles per gallon."

14.29 Guess the correlation. Measurements in large samples show that the correlation

(a) between this semester's GPA and the previous semester's GPA of an upper-class student is about _____.

(b) between IQ and the scores on a test of the reading ability of seventh-grade students is about _____.

(c) between the number of hours a student spends studying per week and the average number of hours spent studying by his or her roommates is about _____.

The answers (in scrambled order) are

$$r = 0.2 \qquad r = 0.5 \qquad r = 0.8$$

Match the answers to the statements and explain your choice.

14.30 Guess the correlation. For each of the following pairs of variables, would you expect a substantial negative correlation, a substantial positive correlation, or a small correlation?

(a) the cost of a cable TV service and the number of channels provided by the service

(b) the weight of a road-racing bicycle and the cost of the bicycle

TABLE 14.2 2015 Hot dog and beer (per ounce) prices (in dollars) at some Major League Baseball stadiums

Team	Hot dog	Beer	Team	Hot dog	Beer	Team	Hot dog	Beer
Angels	4.50	0.28	Giants	5.50	0.50	Rays	5.00	0.42
Astros	4.75	0.36	Indians	3.00	0.33	Reds	1.00	0.44
Blue Jays	4.98	0.49	Marlins	6.00	0.50	Red Sox	5.25	0.65
Braves	4.75	0.45	Mets	6.25	0.48	Rockies	4.75	0.38
Brewers	3.50	0.38	Padres	4.00	0.36	Royals	5.00	0.41
Cardinals	4.25	0.42	Phillies	3.75	0.37	Tigers	4.50	0.42
Diamondbacks	2.75	0.29	Pirates	3.25	0.34	Twins	4.50	0.38
Dodgers	5.50	0.31	Rangers	5.00	0.31	White Sox	4.00	0.41

(c) the number of hours a student spends on Facebook and the student's GPA

(d) the heights and salaries of faculty members at your university

14.31 Investment diversification. A mutual funds company's newsletter says, "A well-diversified portfolio includes assets with low correlations." The newsletter includes a table of correlations between the returns on various classes of investments. For example, the correlation between municipal bonds and large-cap stocks is 0.50, and the correlation between municipal bonds and small-cap stocks is 0.21.

(a) Rachel invests heavily in municipal bonds. She wants to diversify by adding an investment whose returns do not closely follow the returns on her bonds. Should she choose large-cap stocks or small-cap stocks for this purpose? Explain your answer.

(b) If Rachel wants an investment that tends to increase when the return on her bonds drops, what kind of correlation should she look for?

14.32 Take me out to the ball game. What is the relationship between the price charged for a hot dog and the price charged, per ounce, for beer in Major League Baseball stadiums? Table 14.2 gives some data. Make a scatterplot appropriate for showing how beer price helps explain hot dog price. Describe the relationship that you see. Are there any outliers?

14.33 When it rains, it pours. Figure 14.12 plots the highest *yearly* precipitation ever recorded in each state against the highest *daily* precipitation ever recorded in that state. The points for Alaska (AK), Hawaii (HI), and Texas (TX) are marked on the scatterplot.

(a) About what are the highest daily and yearly precipitation values for Alaska?

(b) Alaska and Hawaii have very high yearly maximums relative to their daily maximums. Omit these two states as outliers. Describe the nature of the relationship for the other states. Would knowing a state's highest daily

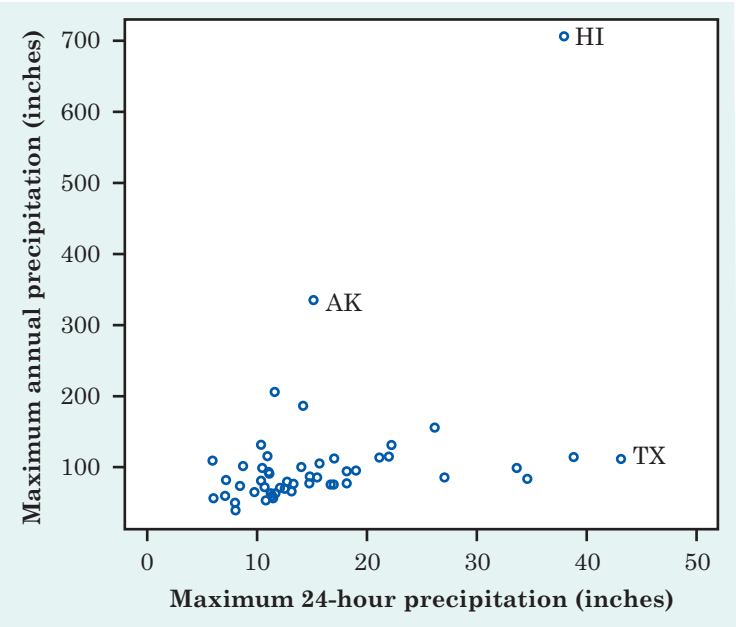


Figure 14.12 Record-high yearly precipitation recorded at any weather station in each state plotted against record-high daily precipitation for the state, Exercise 14.33.

precipitation be a great help in predicting that state's highest yearly precipitation?

14.34 How many corn plants is too many? How much corn per acre should a farmer plant to obtain the highest yield? To find the best planting rate, do an experiment: plant at different rates on several plots of ground and measure the harvest. Here are data from such an experiment:

Plants per acre	Yield (bushels per acre)			
12,000	150.1	113.0	118.4	142.6
16,000	166.9	120.7	135.2	149.8
20,000	165.3	130.1	139.6	149.9
24,000	134.7	138.4	156.1	
28,000	119.0	150.5		

(a) Is yield or planting rate the explanatory variable? Why?

(b) Make a scatterplot of yield and planting rate.

(c) Describe the overall pattern of the relationship. Is it a straight line? Is there a positive or negative association, or neither? Explain why increasing the number of plants per acre of ground has the effect that your graph shows.

14.35 Why so small? Make a scatterplot of the following data:

x	1	2	3	4	9	10
y	12	2	3	5	9	11

Use your calculator to show that the correlation is about 0.4. What feature of the data is responsible for reducing the correlation to this value

despite a strong straight-line association between x and y in most of the observations?

14.36 Ecological correlation. Many studies reveal a positive correlation between income and number of years of education. To investigate this, a researcher makes two plots.

Plot 1: Plot the number of years of education (the explanatory variable) versus the average annual income of all adults having that many years of education (the response variable).

Plot 2: Plot the number of years of education (the explanatory variable) versus the individual annual incomes of all adults (the response variable).

Which plot will display a stronger correlation? (*Hint*: Which plot will display a greater amount of scatter? In particular, will the variation from individual to individual having the same number of years of education create more or less scatter in Plot 2 compared with plotting the average incomes in Plot 1? What effect will increased scatter have on the strength of the association we observe?)

Note: A correlation based on averages rather than on individuals is called an **ecological correlation**. Correlations based on averages can be misleading if they are interpreted to be about individuals.

14.37 Ecological correlation again. In Exercise 14.11 (page 330), would the association be stronger, weaker, or the same if the data given listed the GPAs of individual students (rather than averages) and the distance they lived from campus?

EXPLORING THE WEB

Follow the QR code to access exercises.