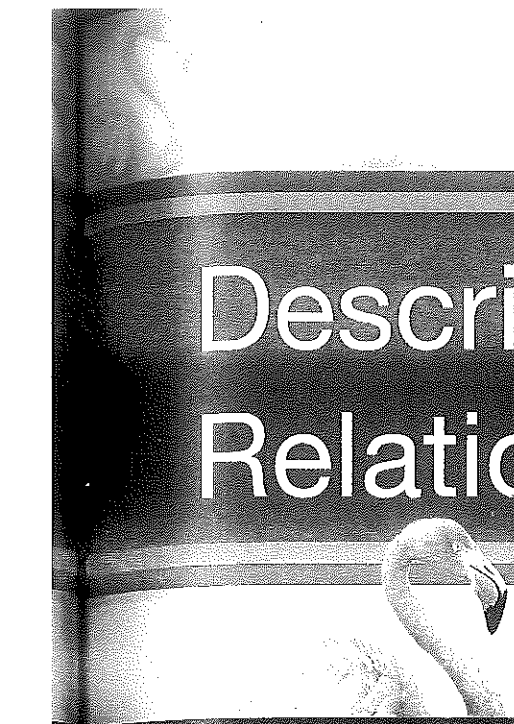
Describing Relationships



case study

How Faithful Is Old Faithful?

The Starnes family visited Yellowstone National Park in hopes of seeing the Old Faithful geyser erupt. They had only about four hours to spend in the park. When they pulled into the parking lot near Old Faithful, a large crowd of people was headed back to their cars from the geyser. Old Faithful had just finished erupting. How long would the Starnes family have to wait until the next eruption?

Let's look at some data. Figure 3.1 shows a histogram of times (in minutes) between consecutive eruptions of Old Faithful in the month before the Starnes family's visit. The shortest interval was 47 minutes, and the longest was 113 minutes. That's a lot of variability! The distribution has two clear peaks—one at about 60 minutes and the other at about 90 minutes.

If the Starnes family hopes for a 60-minute gap between eruptions, but the actual interval is closer to 90 minutes, the kids will get impatient. If they plan for a 90-minute interval and go somewhere else in the park, they won't get back in time to see the next eruption if the gap is only about 60 minutes. What should the Starnes family do?

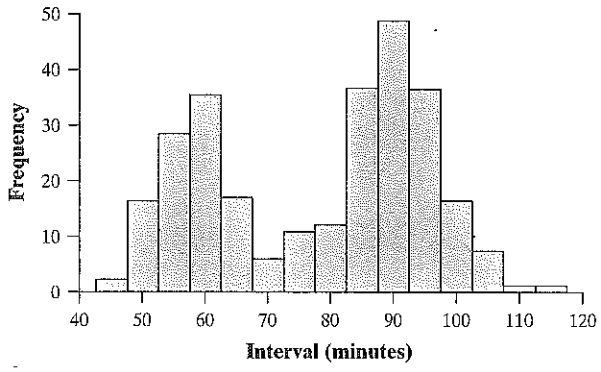


FIGURE 3.1 Histogram of the interval (in minutes) between eruptions of the Old Faithful geyser in the month prior to the Starnes family's visit.

Later in the chapter, you'll answer this question. For now, keep this in mind: to understand one variable (like eruption interval), you often have to look at how it is related to other variables.

Introduction

Investigating relationships between variables is central to what we do in statistics. When we understand the relationship between two variables, we can use the value of one variable to help us make predictions about the other variable. In Section 1.1, we explored relationships between categorical variables, such as the gender of a young person and his or her opinion about future income. The association between these two variables suggests that males are generally more optimistic about their future income than females.

In this chapter, we investigate relationships between two quantitative variables. Does knowing the number of points a football team scores per game tell us anything about how many wins it will have? What can we learn about the price of a used car from the number of miles it has been driven? Are there any variables that might help the Starnes family predict how long it will be until the next eruption of Old Faithful?

ACTIVITY | CSI Stats: The case of the missing cookies

MATERIALS:

Meterstick, handprint, and math department roster (from *Teacher's Resource Materials*) for each group of three to four students; one sheet of graph paper per student

Mrs. Hagen keeps a large jar full of cookies on her desk for her students. Over the past few days, a few cookies have disappeared. The only people with access to Mrs. Hagen's desk are the other math teachers at her school. She asks her colleagues whether they have been making withdrawals from the cookie jar. No one confesses to the crime.

But the next day, Mrs. Hagen catches a break—she finds a clear handprint on the cookie jar. The careless culprit has left behind crucial evidence! At this point, Mrs. Hagen calls in the CSI Stats team (your class) to help her identify the prime suspect in "The Case of the Missing Cookies."

1. Measure the height and hand span of each member of your group to the nearest centimeter (cm). (Hand span is the maximum distance from the tip of the thumb to the tip of the pinkie finger on a person's fully stretched-out hand.)

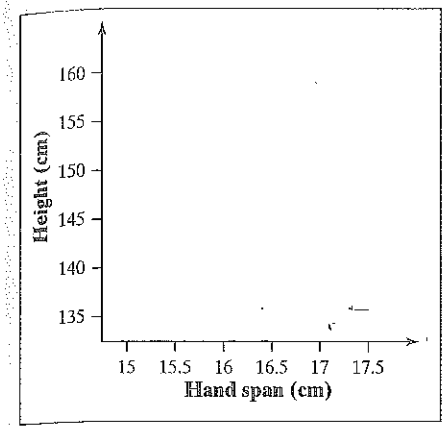
2. Your teacher will make a data table on the board with two columns, labeled as follows:

Hand span (cm)	Height (cm)

Send a representative to record the data for each member of your group in the table.

3. Copy the data table onto your graph paper very near the left margin of the page. Next, you will make a graph of these data. Begin by constructing a set of coordinate axes. Allow plenty of space on the page for your graph. Label the horizontal axis "Hand span (cm)" and the vertical axis "Height (cm)."

4. Since neither hand span nor height can be close to 0 cm, we want to start our horizontal and vertical scales at larger numbers. Scale the horizontal axis in 0.5-cm increments starting with 15 cm. Scale the vertical axis in 5-cm



increments starting with 135 cm. Refer to the sketch in the margin for comparison.

5. Plot each point from your class data table as accurately as you can on the graph. Compare your graph with those of your group members.

6. As a group, discuss what the graph tells you about the relationship between hand span and height. Summarize your observations in a sentence or two.

7. Ask your teacher for a copy of the handprint found at the scene and the math department roster. Which math teacher does your group believe is the "prime suspect"? Justify your answer with appropriate statistical evidence.

3.1 Scatterplots and Correlation

WHAT YOU WILL LEARN By the end of the section, you should be able to:

- Identify explanatory and response variables in situations where one variable helps to explain or influences the other.
- Make a scatterplot to display the relationship between two quantitative variables.
- Describe the direction, form, and strength of a relationship displayed in a scatterplot and identify outliers in a scatterplot.
- Interpret the correlation.
- Understand the basic properties of correlation, including how the correlation is influenced by outliers.
- Use technology to calculate correlation.
- Explain why association does not imply causation.

Most statistical studies examine data on more than one variable. Fortunately, 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:

- Plot the data, then add numerical summaries.
- Look for overall patterns and departures from those patterns.
- When there's a regular overall pattern, use a simplified model to describe it.

Explanatory and Response Variables

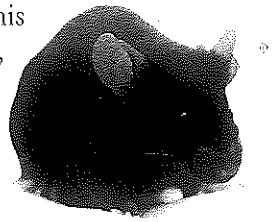
We think that car weight helps explain accident deaths and that smoking influences life expectancy. In these relationships, the two variables play different roles. Accident death rate and life expectancy are the **response variables** of interest. Car weight and number of cigarettes smoked are the **explanatory variables**.

DEFINITION: Response variable, explanatory variable

A **response variable** measures an outcome of a study. An **explanatory variable** may help explain or predict changes in a response variable.

You will often see explanatory variables called *independent variables* and response variables called *dependent variables*. Because the words "independent" and "dependent" have other meanings in statistics, we won't use them here.

It is easiest to identify explanatory and response variables when we actually specify values of one variable to see how it affects another variable. For instance, to study the effect of alcohol on body temperature, researchers gave several different amounts of alcohol to mice. Then they measured the change in each mouse's body temperature 15 minutes later. In this case, amount of alcohol is the explanatory variable, and change in body temperature is the response variable. When we don't specify the values of either variable but just observe both variables, there may or may not be explanatory and response variables. Whether there are depends on how you plan to use the data.



EXAMPLE Linking SAT Math and Critical Reading Scores

Explanatory or response?

Julie asks, "Can I predict a state's mean SAT Math score if I know its mean SAT Critical Reading score?" Jim wants to know how the mean SAT Math and Critical Reading scores this year in the 50 states are related to each other.

PROBLEM: For each student, identify the explanatory variable and the response variable if possible.

SOLUTION: Julie is treating the mean SAT Critical Reading score as the explanatory variable and the mean SAT Math score as the response variable. Jim is simply interested in exploring the relationship between the two variables. For him, there is no clear explanatory or response variable.

For Practice *Try Exercise* 1

In many studies, the goal is to show that changes in one or more explanatory variables actually *cause* changes in a response variable. However, other explanatory-response relationships don't involve direct causation. In the alcohol and mice study, alcohol actually *causes* a change in body temperature. But there is no cause-and-effect relationship between SAT Math and Critical Reading scores. Because the scores are closely related, we can still use a state's mean SAT Critical Reading score to predict its mean Math score. We will learn how to make such predictions in Section 3.2.

CHECK YOUR UNDERSTANDING

Identify the explanatory and response variables in each setting.

1. How does drinking beer affect the level of alcohol in people's blood? The legal limit for driving in all states is 0.08%. In a study, adult volunteers drank different numbers of cans of beer. Thirty minutes later, a police officer measured their blood alcohol levels.

2. The National Student Loan Survey provides data on the amount of debt for recent college graduates, their current income, and how stressed they feel about college debt. A sociologist looks at the data with the goal of using amount of debt and income to explain the stress caused by college debt.

Displaying Relationships: Scatterplots

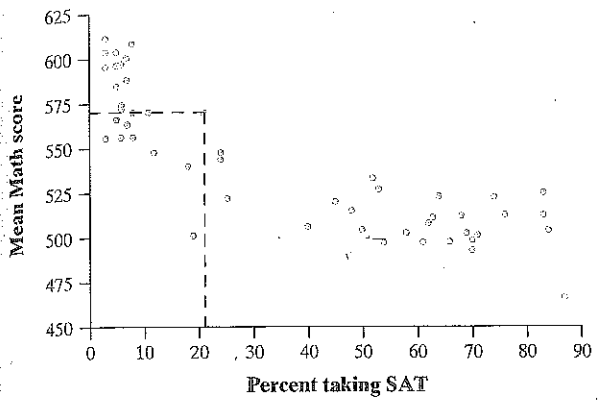


FIGURE 3.2 Scatterplot of the mean SAT Math score in each state against the percent of that state's high school graduates who took the SAT. The dotted lines intersect at the point (21, 570), the values for Colorado.

The most useful graph for displaying the relationship between two quantitative variables is a **scatterplot.** Figure 3.2 shows a scatterplot of the percent of high school graduates in each state who took the SAT and the state's mean SAT Math score in a recent year. We think that "percent taking" will help explain "mean score." So "percent taking" is the explanatory variable and "mean score" is the response variable. We want to see how mean score changes when percent taking changes, so we put percent taking (the explanatory variable) on the horizontal axis. Each point represents a single state. In Colorado, for example, 21% took the SAT, and their mean SAT Math score was 570. Find 21 on the x (horizontal) axis and 570 on the y (vertical) axis. Colorado appears as the point (21, 570).

DEFINITION: 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 a point in the graph.

Here's a helpful way to remember: the e**X**planatory variable goes on the x axis.

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.

We used computer software to produce Figure 3.2. For some problems, you'll be expected to make scatterplots by hand. Here's how to do it.

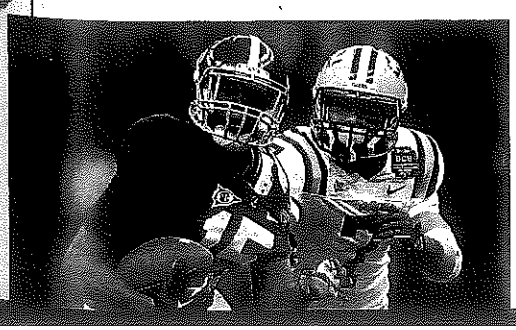
HOW TO MAKE A SCATTERPLOT

1. Decide which variable should go on each axis.
2. Label and scale your axes.
3. Plot individual data values.

The following example illustrates the process of constructing a scatterplot.

EXAMPLE SEC Football

Making a scatterplot



At the end of the 2011 college football season, the University of Alabama defeated Louisiana State University for the national championship. Interestingly, both of these teams were from the Southeastern Conference (SEC). Here are the average number of points scored per game and number of wins for each of the twelve teams in the SEC that season.[1]

Team	Alabama	Arkansas	Auburn	Florida	Georgia	Kentucky
Points per game	34.8	36.8	25.7	25.5	32.0	15.8
Wins	12	11	8	7	10	5
Team	**Louisiana State**	**Mississippi**	**Mississippi State**	**South Carolina**	**Tennessee**	**Vanderbilt**
Points per game	35.7	16.1	25.3	30.1	20.3	26.7
Wins	13	2	7	11	5	6

PROBLEM: Make a scatterplot of the relationship between points per game and wins.

SOLUTION: We follow the steps described earlier to make the scatterplot.

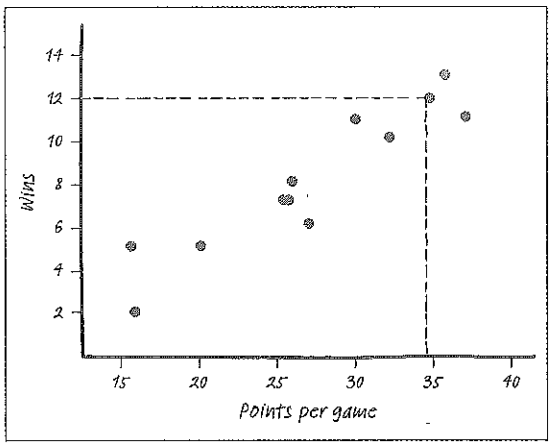


FIGURE 3.3 Completed scatterplot of points per game and wins for the teams in the SEC. The dotted lines intersect at the point (34.8, 12), the values for Alabama.

1. *Decide which variable should go on which axis.* The number of wins a football team has depends on the number of points they score. So we'll use points per game as the explanatory variable (*x* axis) and wins as the response variable (*y* axis).

2. *Label and scale your axes.* We labeled the *x* axis "Points per game" and the *y* axis "Wins." Because the teams' points per game vary from 15.8 to 36.8, we chose a horizontal scale starting at 15 points, with tick marks every 5 points. The teams' wins vary from 2 to 13, so we chose a vertical scale starting at 0 with tick marks every 2 wins.

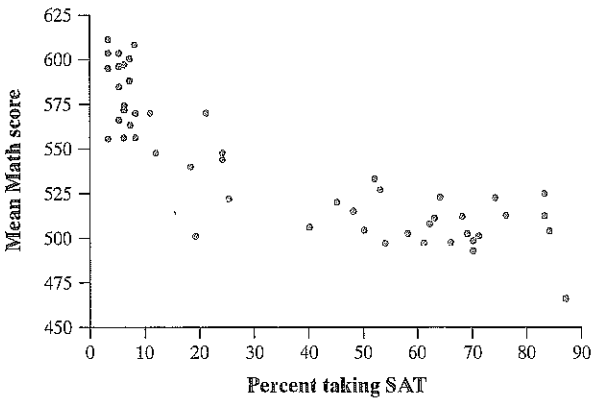
3. *Plot individual data values.* The first team in the table, Alabama, scored 34.8 points per game and had 12 wins. We plot this point directly above 34.8 on the horizontal axis and to the right of 12 on the vertical axis, as shown in Figure 3.3. For the second team in the list, Arkansas, we add the point (36.8, 11) to the graph. By adding the points for the remaining ten teams, we get the completed scatterplot in Figure 3.3.

For Practice *Try Exercise* 5

Describing Scatterplots

To describe a scatterplot, follow the basic strategy of data analysis from Chapters 1 and 2: look for patterns and important departures from those patterns. Let's take a closer look at the scatterplot from Figure 3.2. What do we see?

- The graph shows a clear **direction:** the overall pattern moves from upper left to lower right. That is, states in which higher percents of high school graduates take the SAT tend to have lower mean SAT Math scores. We call this a *negative association* between the two variables.
- The **form** of the relationship is slightly curved. More important, most states fall into one of two distinct *clusters*. In about half of the states, 25% or fewer graduates took the SAT. In the other half, more than 40% took the SAT.
- The **strength** of a relationship in a scatterplot is determined by how closely the points follow a clear form. The overall relationship in Figure 3.2 is moderately strong: states with similar percents taking the SAT tend to have roughly similar mean SAT Math scores.



- Two states stand out in the scatterplot: West Virginia at (19, 501) and Maine at (87, 466). These points can be described as **outliers** because they fall outside the overall pattern.

THINK ABOUT IT

What explains the clusters? There are two widely used college entrance exams, the SAT and the American College Testing (ACT) exam. Each state usually favors one or the other. The ACT states cluster at the left of Figure 3.2 and the SAT states at the right. In ACT states, most students who take the SAT are applying to a selective college that prefers SAT scores. This select group of students has a higher mean score than the much larger group of students who take the SAT in SAT states.

HOW TO EXAMINE A SCATTERPLOT

As in any graph of data, look for the *overall pattern* and for striking *departures* from that pattern.

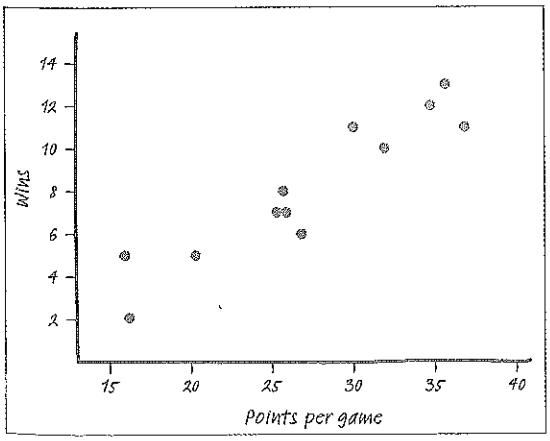
- You can describe the overall pattern of a scatterplot by the **direction**, **form**, and **strength** of the relationship.
- An important kind of departure is an **outlier**, an individual value that falls outside the overall pattern of the relationship.

Let's practice examining scatterplots using the SEC football data from the previous example.

EXAMPLE SEC Football

Describing a scatterplot

In the last example, we constructed the scatterplot shown below that displays the average number of points scored per game and the number of wins for college football teams in the Southeastern Conference.



PROBLEM: Describe what the scatterplot reveals about the relationship between points per game and wins.

SOLUTION: *Direction:* In general, it appears that teams that score more points per game have more wins and teams that score fewer points per game have fewer wins. We say that there is a *positive association* between points per game and wins.

Form: There seems to be a linear pattern in the graph (that is, the overall pattern follows a straight line).

Strength: Because the points do not vary much from the linear pattern, the relationship is fairly strong. There do not appear to be any values that depart from the linear pattern, so there are no *outliers*.

For Practice *Try Exercise* 7

Even when there is a clear association between two variables in a scatterplot, the direction of the relationship only describes the overall trend—not the relationship for each pair of points. For example, even though teams that score more points per game generally have more wins, Georgia and South Carolina are exceptions to the overall pattern. Georgia scored *more* points per game than South Carolina (32 versus 30.1) but had *fewer* wins (10 versus 11).

So far, we've seen relationships with two different directions. The number of wins generally increases as the points scored per game increases (**positive association**). The mean SAT score generally goes down as the percent of graduates taking the test increases (**negative association**). Let's give a careful definition for these terms.

DEFINITION: Positive association, negative association

Two variables have a **positive association** when above-average values of one tend to accompany above-average values of the other and when below-average values also tend to occur together.

Two variables have a **negative association** when above-average values of one tend to accompany below-average values of the other.

Of course, not all relationships have a clear direction that we can describe as a positive association or a negative association. Exercise 9 involves a relationship that doesn't have a single direction. This next example, however, illustrates a strong positive association with a simple and important form.

EXAMPLE

The Endangered Manatee

Pulling it all together

Manatees are large, gentle, slow-moving creatures found along the coast of Florida. Many manatees are injured or killed by boats. The table below contains data on the number of boats registered in Florida (in thousands) and the number of manatees killed by boats for the years 1977 to 2010.[2]

Florida boat registrations (thousands) and manatees killed by boats

YEAR	BOATS	MANATEES	YEAR	BOATS	MANATEES	YEAR	BOATS	MANATEES
1977	447	13	1989	711	50	2001	944	81
1978	460	21	1990	719	47	2002	962	95
1979	481	24	1991	681	53	2003	978	73
1980	498	16	1992	679	38	2004	983	69
1981	513	24	1993	678	35	2005	1010	79
1982	512	20	1994	696	49	2006	1024	92
1983	526	15	1995	713	42	2007	1027	73
1984	559	34	1996	732	60	2008	1010	90
1985	585	33	1997	755	54	2009	982	97
1986	614	33	1998	809	66	2010	942	83
1987	645	39	1999	830	82			
1988	675	43	2000	880	78			

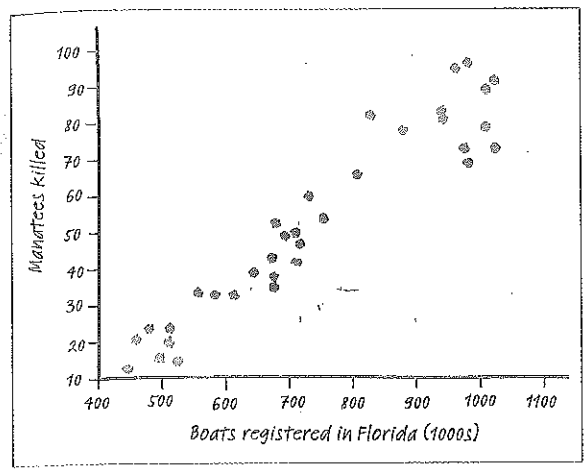
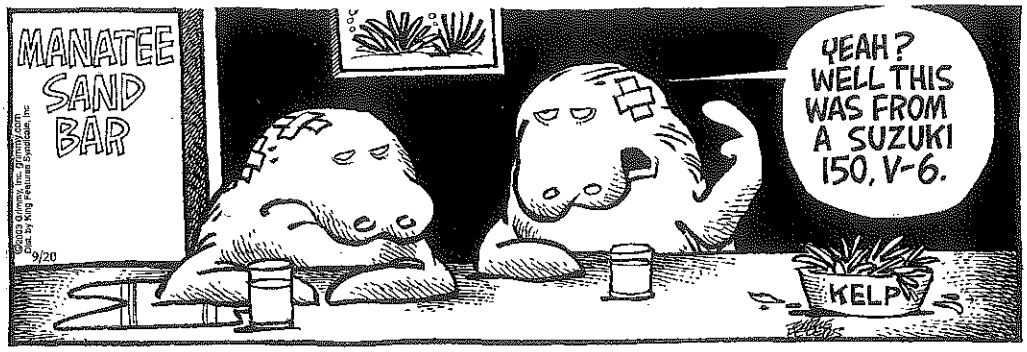


FIGURE 3.4 Scatterplot of the number of Florida manatees killed by boats from 1977 to 2010 against the number of boats registered in Florida that year.

PROBLEM: Make a scatterplot to show the relationship between the number of manatees killed and the number of registered boats. Describe what you see.

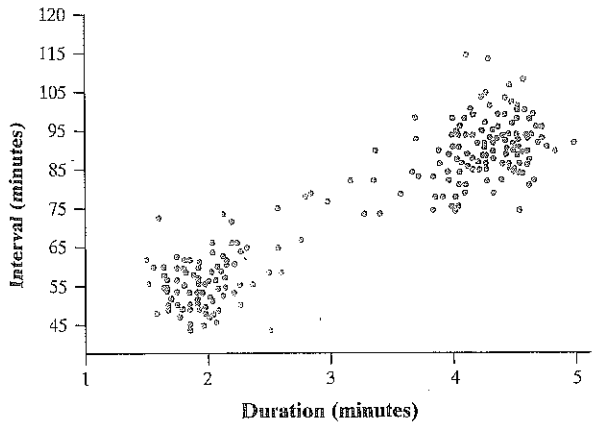
SOLUTION: For the scatterplot, we'll use "boats registered" as the explanatory variable and "manatees killed" as the response variable. Figure 3.4 is our completed scatterplot. There is a positive association—more boats registered goes with more manatees killed. The form of the relationship is linear. That is, the overall pattern follows a straight line from lower left to upper right. The relationship is strong because the points don't deviate greatly from a line, except for the 4 years that have a high number of boats registered, but fewer deaths than expected based on the linear pattern.

For Practice *Try Exercise* 13



CHECK YOUR UNDERSTANDING

In the chapter-opening Case Study (page 141), the Starnes family arrived at Old Faithful after it had erupted. They wondered how long it would be until the next eruption. Here is a scatterplot that plots the interval between consecutive eruptions of Old Faithful against the duration of the previous eruption, for the month prior to their visit.



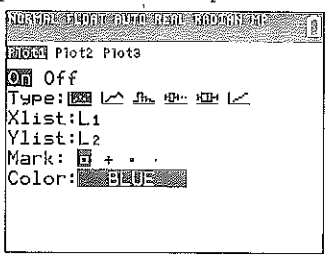
1. Describe the direction of the relationship. Explain why this makes sense.
2. What form does the relationship take? Why are there two clusters of points?
3. How strong is the relationship? Justify your answer.
4. Are there any outliers?
5. What information does the Starnes family need to predict when the next eruption will occur?

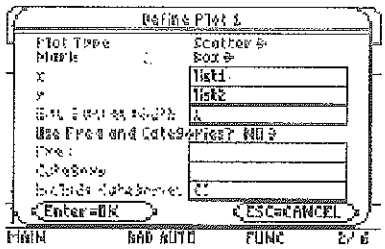
7. TECHNOLOGY CORNER SCATTERPLOTS ON THE CALCULATOR

TI-Nspire instructions in Appendix B; HP Prime instructions on the book's Web site.

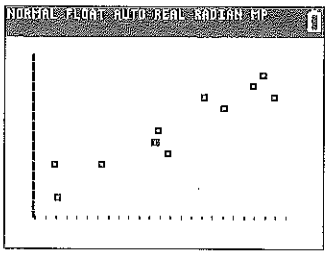
Making scatterplots with technology is much easier than constructing them by hand. We'll use the SEC football data from page 146 to show how to construct a scatterplot on a TI-83/84 or TI-89.

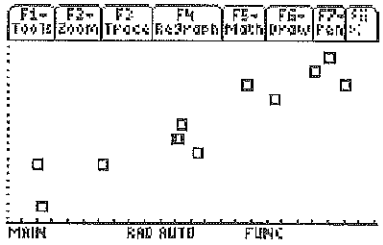
- Enter the data values into your lists. Put the points per game in L1/list1 and the number of wins in L2/list2.
- Define a scatterplot in the statistics plot menu (press F2 on the TI-89). Specify the settings shown below.





- Use ZoomStat (ZoomData on the TI-89) to obtain a graph. The calculator will set the window dimensions automatically by looking at the values in L1/list1 and L2/list2.





Notice that there are no scales on the axes and that the axes are not labeled. If you copy a scatterplot from your calculator onto your paper, make sure that you scale and label the axes.

AP® EXAM TIP If you are asked to make a scatterplot on a free-response question, be sure to label and scale both axes. *Don't* just copy an unlabeled calculator graph directly onto your paper.

Measuring Linear Association: Correlation

A scatterplot displays the direction, form, and strength of the relationship between two quantitative variables. Linear relationships are particularly important because a straight line is a simple pattern that is quite common. A linear relationship is strong if the points lie close to a straight line and weak if they are widely scattered about a line. Unfortunately, *our eyes are not good judges of how strong a linear relationship is.* The two scatterplots in Figure 3.5 (on the facing page) show the same data, but the graph on the right is drawn smaller in a large field. The right-hand graph seems to show a stronger linear relationship.

Because it's easy to be fooled by different scales or by the amount of space around the cloud of points in a scatterplot, we need to use a numerical measure to supplement the graph. **Correlation** is the measure we use.

Some people refer to *r* as the "correlation coefficient."

DEFINITION: Correlation *r*

The **correlation *r*** measures the direction and strength of the linear relationship between two quantitative variables.

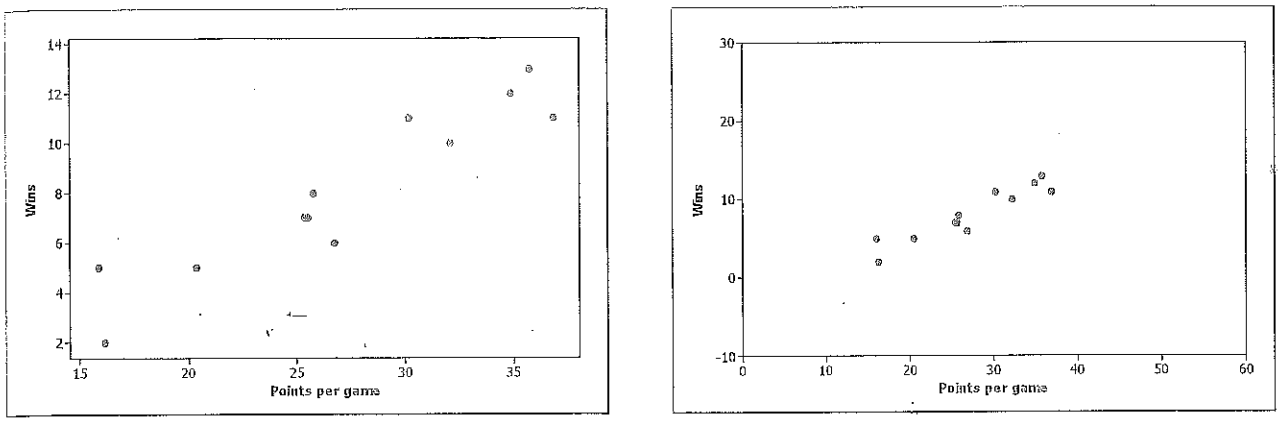


FIGURE 3.5 Two Minitab scatterplots of the same data. The straight-line pattern in the graph on the right appears stronger because of the surrounding space.

How good are you at estimating the correlation by eye from a scatterplot? To find out, try an online applet. Just search for "guess the correlation applets."

The correlation r is always a number between −1 and 1. Correlation indicates the direction of a linear relationship by its sign: $r > 0$ for a positive association and $r < 0$ for a negative association. Values of r near 0 indicate a very weak linear relationship. The strength of the linear relationship increases as r moves away from 0 toward either −1 or 1. The extreme values $r = -1$ and $r = 1$ occur *only* in the case of a perfect linear relationship, when the points lie exactly along a straight line.

Figure 3.6 shows scatterplots that correspond to various values of r. To make the meaning of r clearer, the standard deviations of both variables in these plots are equal, and the horizontal and vertical scales are the same. The correlation describes the direction and strength of the linear relationship in each graph.

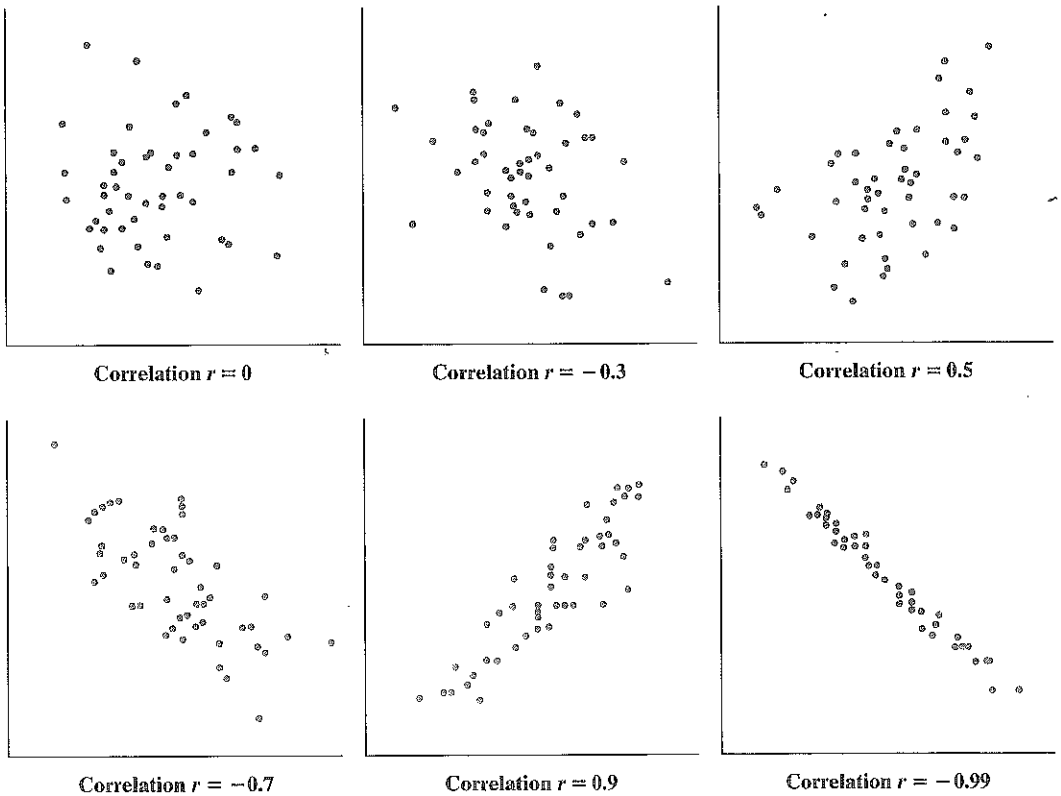


FIGURE 3.6 How correlation measures the strength of a linear relationship. Patterns closer to a straight line have correlations closer to 1 or −1.

The following Activity lets you explore some important properties of the correlation.

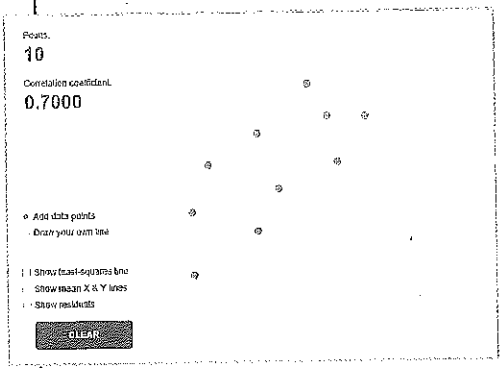
ACTIVITY | Correlation and Regression applet

MATERIALS:
Computer with Internet connection

Go to the book's Web site, www.whfreeman.com/tps5e, and launch the *Correlation and Regression* applet.

1. You are going to use the *Correlation and Regression* applet to make several scatterplots with 10 points that have correlation close to 0.7.

(a) Start by putting two points on the graph. What's the value of the correlation? Why does this make sense?



(b) Make a lower-left to upper-right pattern of 10 points with correlation about $r = 0.7$. (You can drag points up or down to adjust r after you have 10 points.)

(c) Make another scatterplot: this one should have 9 points in a vertical stack at the left of the plot. Add 1 point far to the right and move it until the correlation is close to 0.7.

(d) Make a third scatterplot: make this one with 10 points in a curved pattern that starts at the lower left, rises to the right, then falls again at the far right. Adjust the points up or down until you have a very smooth curve with correlation close to 0.7.

Summarize: If you know that the correlation between two variables is $r = 0.7$, what can you say about the form of the relationship?

2. Click on the scatterplot to create a group of 10 points in the lower-left corner of the scatterplot with a strong straight-line pattern (correlation about 0.9).

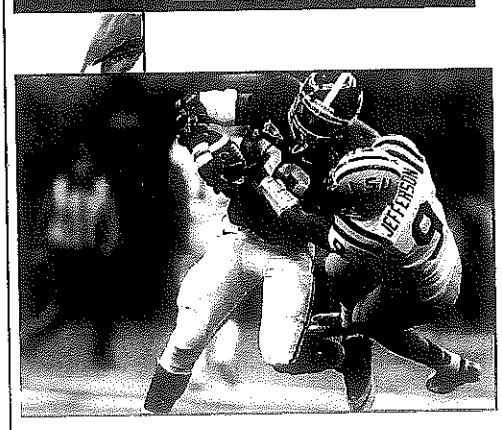
(a) Add 1 point at the upper right that is in line with the first 10. How does the correlation change?

(b) Drag this last point straight down. How small can you make the correlation? Can you make the correlation negative?

Summarize: What did you learn from Step 2 about the effect of a single point on the correlation?

Now that you know what information the correlation provides—and doesn't provide—let's look at an example that shows how to interpret it.

EXAMPLE



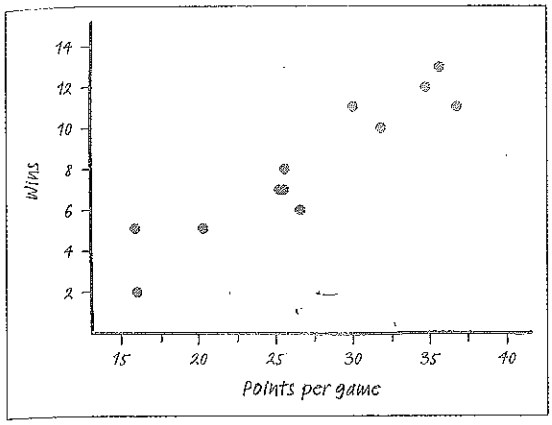
SEC Football

Interpreting correlation

PROBLEM: Our earlier scatterplot of the average points per game and number of wins for college football teams in the SEC is repeated at top right. For these data, $r = 0.936$.

(a) Interpret the value of r in context.

(b) The point highlighted in red on the scatterplot is Mississippi. What effect does Mississippi have on the correlation? Justify your answer.



SOLUTION:

(a) The correlation of 0.936 confirms what we see in the scatterplot: there is a strong, positive linear relationship between points per game and wins in the SEC.

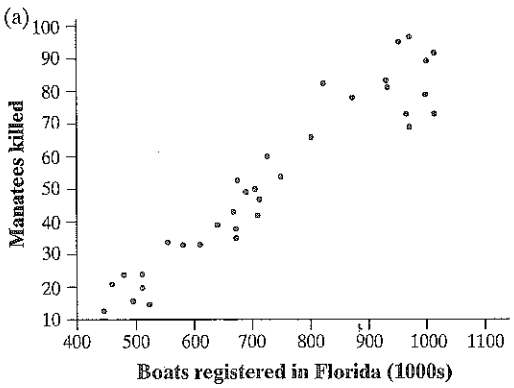
(b) Mississippi makes the correlation closer to 1 (stronger). If Mississippi were not included, the remaining points wouldn't be as tightly clustered in a linear pattern.

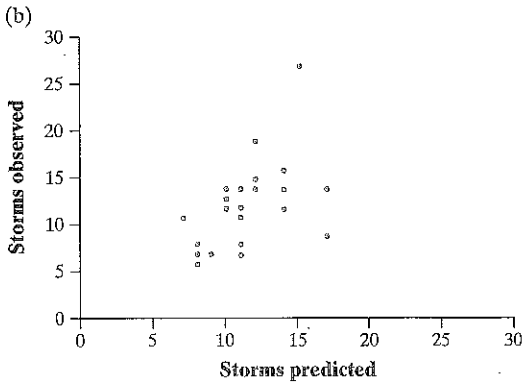
For Practice *Try Exercise* 21

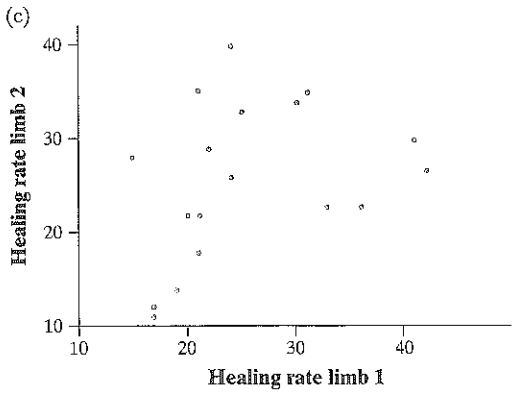
AP® EXAM TIP If you're asked to interpret a correlation, start by looking at a scatterplot of the data. Then be sure to address direction, form, strength, and outliers (sound familiar?) and put your answer in context.

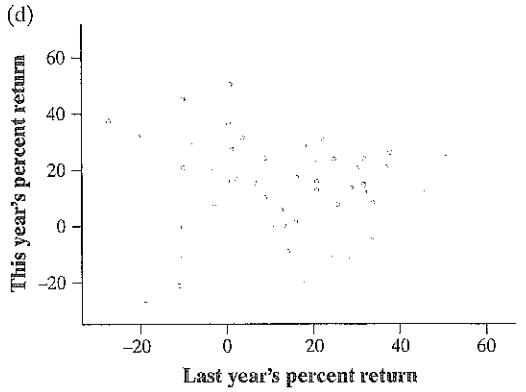
CHECK YOUR UNDERSTANDING

The scatterplots below show four sets of real data: (a) repeats the manatee plot in Figure 3.4 (page 149); (b) shows the number of named tropical storms and the number predicted before the start of hurricane season each year between 1984 and 2007 by William Gray of Colorado State University; (c) plots the healing rate in micrometers (millionths of a meter) per hour for the two front limbs of several newts in an experiment; and (d) shows stock market performance in consecutive years over a 56-year period. For each graph, estimate the correlation r. Then interpret the value of r in context.









Calculating Correlation Now that you have some idea of how to interpret the correlation, let's look at how it's calculated.

HOW TO CALCULATE THE CORRELATION r

Suppose that we have data on variables x and y for n individuals. The values for the first individual are x_1 and y_1, the values for the second individual are x_2 and y_2, and so on. The means and standard deviations of the two variables are $\bar{x}$ and s_x for the x-values, and $\bar{y}$ and s_y for the y-values. The correlation r between x and y is

$$r = \frac{1}{n-1}\left[\left(\frac{x_1-\bar{x}}{s_x}\right)\left(\frac{y_1-\bar{y}}{s_y}\right)+\left(\frac{x_2-\bar{x}}{s_x}\right)\left(\frac{y_2-\bar{y}}{s_y}\right)+\cdots+\left(\frac{x_n-\bar{x}}{s_x}\right)\left(\frac{y_n-\bar{y}}{s_y}\right)\right]$$

or, more compactly,

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

The formula for the correlation r is a bit complex. It helps us see what correlation is, but in practice, you should use your calculator or software to find r. Exercises 19 and 20 ask you to calculate a correlation step-by-step from the definition to solidify its meaning.

The formula for r begins by standardizing the observations. Let's use the familiar SEC football data to perform the required calculations. The table below shows the values of points per game x and number of wins y for the SEC college football teams. For these data, $\bar{x} = 27.07$ and $s_x = 7.16$.

Team	Alabama	Arkansas	Auburn	Florida	Georgia	Kentucky
Points per game	34.8	36.8	25.7	25.5	32.0	15.8
Wins	12	11	8	7	10	5
Team	Louisiana State	Mississippi	Mississippi State	South Carolina	Tennessee	Vanderbilt
Points per game	35.7	16.1	25.3	30.1	20.3	26.7
Wins	13	2	7	11	5	6

The value

$$\frac{x_i-\bar{x}}{s_x}$$

in the correlation formula is the standardized points per game (z-score) of the ith team. For the first team in the table (Alabama), the corresponding z-score is

$$z_x = \frac{34.8-27.07}{7.16} = 1.08$$

That is, Alabama's points per game total (34.8) is a little more than 1 standard deviation above the mean points per game for the SEC teams.

Standardized values have no units—in this example, they are no longer measured in points.

To standardize the number of wins, we use $\bar{y} = 8.08$ and $s_y = 3.34$. For Alabama, $z_y = \dfrac{12 - 8.08}{3.34} = 1.17$. Alabama's number of wins (12) is 1.17 standard deviations above the mean number of wins for SEC teams. When we multiply this team's two z-scores, we get a product of 1.2636. The correlation r is an "average" of the products of the standardized scores for all the teams. Just as in the case of the standard deviation s_x, the average here divides by one fewer than the number of individuals. Finishing the calculation reveals that $r = 0.936$ for the SEC teams.

Some people like to write the correlation formula as

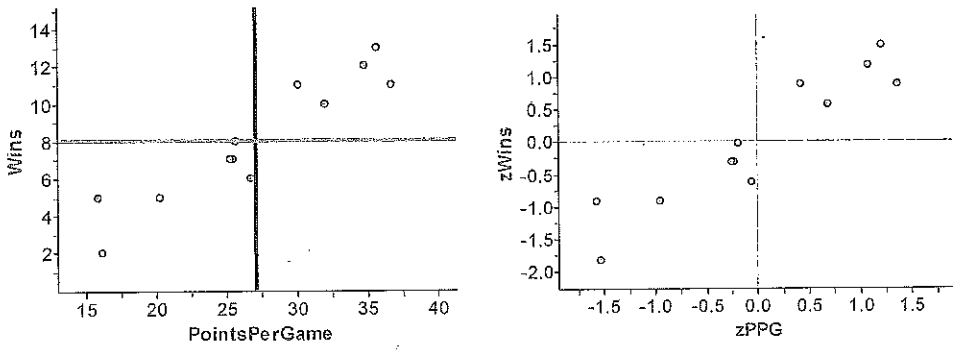
$$r = \frac{1}{n-1}\sum z_x z_y$$

to emphasize the product of standardized scores in the calculation.

THINK ABOUT IT

What does correlation measure? The Fathom screen shots below provide more detail. At the left is a scatterplot of the SEC football data with two lines added—a vertical line at the group's mean points per game and a horizontal line at the mean number of wins of the group. Most of the points fall in the upper-right or lower-left "quadrants" of the graph. That is, teams with above-average points per game tend to have above-average numbers of wins, and teams with below-average points per game tend to have numbers of wins that are below average. This confirms the positive association between the variables.

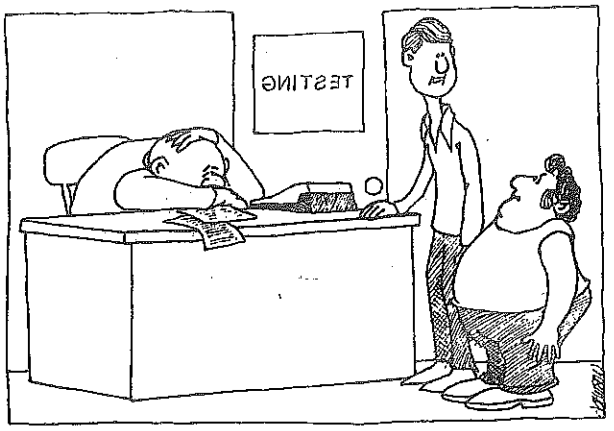
Below on the right is a scatterplot of the standardized scores. To get this graph, we transformed both the x- and the y-values by subtracting their mean and dividing by their standard deviation. As we saw in Chapter 2, standardizing a data set converts the mean to 0 and the standard deviation to 1. That's why the vertical and horizontal lines in the right-hand graph are both at 0.



Notice that all the products of the standardized values will be positive—not surprising, considering the strong positive association between the variables. What if there was a negative association between two variables? Most of the points would be in the upper-left and lower-right "quadrants" and their z-score products would be negative, resulting in a negative correlation.

Facts about Correlation

How correlation behaves is more important than the details of the formula. Here's what you need to know in order to interpret correlation correctly.



"He says we've ruined his positive correlation between height and weight."

1. *Correlation makes no distinction between explanatory and response variables.* It makes no difference which variable you call x and which you call y in calculating the correlation. Can you see why from the formula?

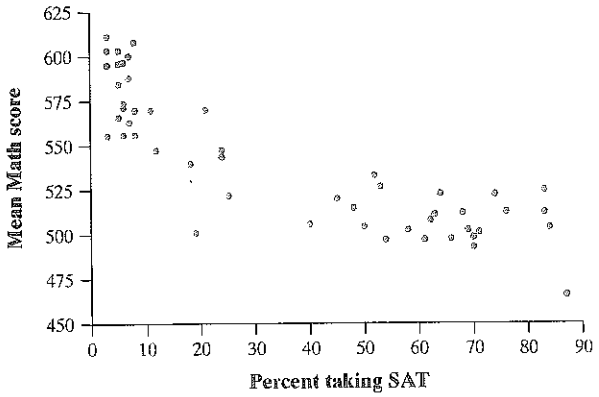
$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

2. Because r uses the standardized values of the observations, *r does not change when we change the units of measurement of x, y, or both.* Measuring height in centimeters rather than inches and weight in kilograms rather than pounds does not change the correlation between height and weight.

3. *The correlation r itself has no unit of measurement.* It is just a number.

Describing the relationship between two variables is more complex than describing the distribution of one variable. Here are some cautions to keep in mind when you use correlation.

- *Correlation does not imply causation.* Even when a scatterplot shows a strong linear relationship between two variables, we can't conclude that changes in one variable cause changes in the other. For example, looking at data from the last 10 years, there is a strong positive relationship between the number of high school students who own a cell phone and the number of students who pass the AP® Statistics exam. Does this mean that buying a cell phone will help you pass the AP® exam? Not likely. Instead, the correlation is positive because both of these variables are increasing over time.
- *Correlation requires that both variables be quantitative,* so that it makes sense to do the arithmetic indicated by the formula for r. We cannot calculate a correlation between the incomes of a group of people and what city they live in because city is a categorical variable.
- Correlation only measures the strength of a linear relationship between two variables. *Correlation does not describe curved relationships between variables, no matter how strong the relationship is.* A correlation of 0 doesn't guarantee that there's *no* relationship between two variables, just that there's no *linear* relationship.
- *A value of r close to 1 or -1 does not guarantee a linear relationship between two variables.* A scatterplot with a clear curved form can have a correlation that is close to 1 or -1. For example, the correlation between percent taking the SAT and mean Math score is close to -1, but the association is clearly curved. Always plot your data!



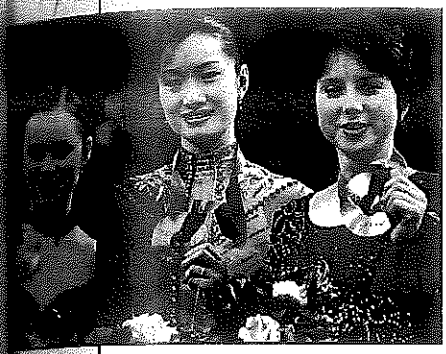
- *Like the mean and standard deviation, the correlation is not resistant: r is strongly affected by a few outlying observations.* Use r with caution when outliers appear in the scatterplot.
- *Correlation is not a complete summary of two-variable data,* even when the relationship between the variables is linear. You should give the means and standard deviations of both x and y along with the correlation.

Of course, even giving means, standard deviations, and the correlation for "state SAT Math scores" and "percent taking" will not point out the clusters in Figure 3.2. Numerical summaries complement plots of data, but they do not replace them.

EXAMPLE

Scoring Figure Skaters

Why correlation doesn't tell the whole story



Until a scandal at the 2002 Olympics brought change, figure skating was scored by judges on a scale from 0.0 to 6.0. The scores were often controversial. We have the scores awarded by two judges, Pierre and Elena, for many skaters. How well do they agree? We calculate that the correlation between their scores is $r = 0.9$. But the mean of Pierre's scores is 0.8 point lower than Elena's mean.

These facts don't contradict each other. They simply give different kinds of information. The mean scores show that Pierre awards lower scores than Elena. But because Pierre gives *every* skater a score about 0.8 point lower than Elena does, the correlation remains high. Adding the same number to all values of either x or y does not change the correlation. If both judges score the same skaters, the competition is scored consistently because Pierre and Elena agree on which performances are better than others. The high r shows their agreement. But if Pierre scores some skaters and Elena others, we should add 0.8 point to Pierre's scores to arrive at a fair comparison.

DATA EXPLORATION

The SAT essay: Is longer better?

Following the debut of the new SAT Writing test in March 2005, Dr. Les Perelman from the Massachusetts Institute of Technology stirred controversy by reporting, "It appeared to me that regardless of what a student wrote, the longer the essay, the higher the score." He went on to say, "I have never found a quantifiable predictor in 25 years of grading that was anywhere as strong as this one. If you just graded them based on length without ever reading them, you'd be right over 90 percent of the time."[3] The table below shows the data that Dr. Perelman used to draw his conclusions.[4]

Length of essay and score for a sample of SAT essays

Words:	460	422	402	365	357	278	236	201	168	156	133
Score:	6	6	5	5	6	5	4	4	4	3	2
Words:	114	108	100	403	401	388	320	258	236	189	128
Score:	2	1	1	5	6	6	5	4	4	3	2
Words:	67	697	387	355	337	325	272	150	135		
Score:	1	6	6	5	5	4	4	2	3		

Does this mean that if students write a lot, they are guaranteed high scores? Carry out your own analysis of the data. How would you respond to each of Dr. Perelman's claims?

Section 3.1 Summary

- A **scatterplot** displays the relationship between two quantitative variables measured on the same individuals. Mark values of one variable on the horizontal axis (x axis) and values of the other variable on the vertical axis (y axis). Plot each individual's data as a point on the graph.
- If we think that a variable x may help explain, predict, or even cause changes in another variable y, we call x an **explanatory variable** and y a **response variable**. Always plot the explanatory variable, if there is one, on the x axis of a scatterplot. Plot the response variable on the y axis.
- In examining a scatterplot, look for an overall pattern showing the **direction, form,** and **strength** of the relationship and then look for **outliers** or other departures from this pattern.
- **Direction:** If the relationship has a clear direction, we speak of either **positive association** (above-average values of the two variables tend to occur together) or **negative association** (above-average values of one variable tend to occur with below-average values of the other variable).
- **Form:** Linear relationships, where the points show a straight-line pattern, are an important form of relationship between two variables. Curved relationships and clusters are other forms to watch for.
- **Strength:** The strength of a relationship is determined by how close the points in the scatterplot lie to a simple form such as a line.
- The **correlation** r measures the strength and direction of the linear association between two quantitative variables x and y. Although you can calculate a correlation for any scatterplot, r measures strength for only straight-line relationships.
- Correlation indicates the direction of a linear relationship by its sign: $r > 0$ for a positive association and $r < 0$ for a negative association. Correlation always satisfies $-1 \leq r \leq 1$ and indicates the strength of a linear relationship by how close it is to -1 or 1. Perfect correlation, $r = \pm 1$, occurs only when the points on a scatterplot lie exactly on a straight line.
- Remember these important facts about r: Correlation does not imply causation. Correlation ignores the distinction between explanatory and response variables. The value of r is not affected by changes in the unit of measurement of either variable. Correlation is not resistant, so outliers can greatly change the value of r.

3.1 TECHNOLOGY CORNER

TI-Nspire instructions in Appendix B; HP Prime instructions on the book's Web site.

7. Scatterplots on the calculator

page 150