

category. Back-to-back stemplots (page 10) and side-by-side boxplots (page 36) are useful tools for this purpose.

We will study methods for describing the association between two categorical variables in Section 2.5 (page 136).

## SECTION 2.1 Summary

To study relationships between variables, we must measure the variables on the same cases.

If we think that a variable  $x$  may explain or even cause changes in another variable  $y$ , we call  $x$  an **explanatory variable** and  $y$  a **response variable**.

A **scatterplot** displays the relationship between two quantitative variables. 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.

Always plot the explanatory variable, if there is one, on the  $x$  axis of a scatterplot. Plot the response variable on the  $y$  axis.

Plot points with different colors or symbols to see the effect of a categorical variable in a scatterplot.

In examining a scatterplot, look for an overall pattern showing the **form**, **direction**, and **strength** of the relationship, and then for **outliers** or other deviations from this pattern.

**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.

**Direction:** If the relationship has a clear direction, we speak of either **positive association** (high values of the two variables tend to occur together) or **negative association** (high values of one variable tend to occur with low values of the other variable).

**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.

To display the relationship between a categorical explanatory variable and a quantitative response variable, make a graph that compares the distributions of the response for each category of the explanatory variable.

## SECTION 2.1 Exercises

For Exercise 2.1, see page 80; for 2.2 and 2.4, see pages 81–82 for 2.5 and 2.6, see page 84; for 2.7 and 2.8, see page 85; for 2.9 to 2.11, see pages 88–89, and for 2.12, see page 92.


**2.13 What's wrong?** Explain what is wrong with each of the following:

- A boxplot can be used to examine the relationship between two variables.
- In a scatterplot we put the response variable on the  $y$  axis and the explanatory variable on the  $x$  axis.
- If two variables are positively associated, then high

values of one variable are associated with low values of the other variable.


**2.14 Make some sketches.** For each of the following situations, make a scatterplot that illustrates the given relationship between two variables.

- A strong negative linear relationship.
- No apparent relationship.
- A weak positive relationship.
- A more complicated relationship. Give the sketch and explain the relationship.


**2.15 Who does not have health insurance?** The lack of adequate health insurance coverage is a major problem for many Americans. The Current Population Survey collected data on the characteristics of the uninsured.<sup>10</sup> The numbers of uninsured and the total number of people classified by age for 2006 are as follows. The units are thousands of people.  HEALTHINSURANCE


Age group	Number uninsured	Total number
Under 18 years	8,661	74,101
18 to 24 years	8,323	28,405
25 to 34 years	10,713	39,868
35 to 44 years	8,018	42,762
45 to 64 years	10,738	75,653
65 years and older	541	36,035

- Plot the number of uninsured versus age group.
- Find the total number of uninsured persons and use this total to compute the percent of the uninsured who are in each age group.
- Plot the percents versus age group.
- Explain how the plot you produced in part (c) differs from the plot that you made in part (a).
- Summarize what you can conclude from these plots.

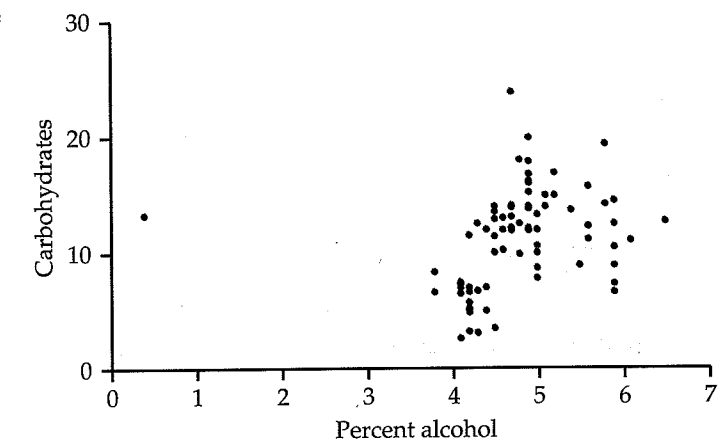
**2.16 Which age groups have the larger percent uninsured?** Refer to the previous exercise. Let's take a look at the data from a different point of view.  HEALTHINSURANCE

- For each age group calculate the percent that are uninsured using the number of uninsured persons and the total number of persons in each group.
- Make a plot of the percent uninsured versus age group.
- Summarize the information in your plot and write a short summary of what you conclude from your analysis.

**2.17 Compare the two percents.** In the previous two exercises, you computed percents in two different ways and generated plots versus age group. Describe the difference between the two ways with an emphasis on what kinds of conclusions can be drawn from each.  HEALTHINSURANCE


**2.18 What's in the beer?** Beer100.com advertises itself as "Your Place for All Things Beer." One of their things is a list of 86 domestic beer brands with the percent alcohol, calories per 12 ounces, and carbohydrates per 12 ounces (in grams).<sup>11</sup>  BEER

(a) Figure 2.10 gives a scatterplot of carbohydrates versus percent alcohol. Give a short summary of what can be learned from the plot.




**FIGURE 2.10** Scatterplot of carbohydrates versus percent alcohol for 86 brands of beer, for Exercise 2.18.

- One of the points is an outlier. Use the data file to find the outlier brand of beer. How is this brand of beer marketed as compared with the other brands?
- Remove the outlier from the data set and generate a scatterplot of the remaining data.
- Describe the relationship between carbohydrates and percent alcohol based on what you see in your scatterplot.

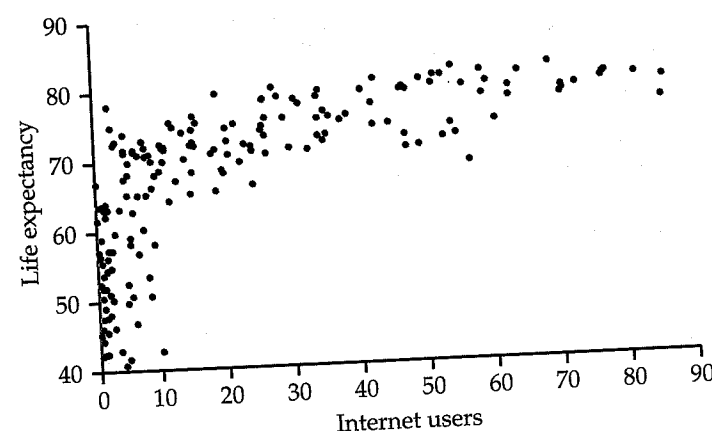
**2.19 More beer.** Refer to the previous exercise.  BEER

- Make a scatterplot of calories versus percent alcohol using the data set without the outlier.
- Describe the relationship between these two variables.

**2.20 Will you live longer if you use the Internet?** The World Bank collects data on many variables related to world development for countries throughout the world. Two of these are Internet use, in number of users per 100 people, and life expectancy, in years.<sup>12</sup> Figure 2.11 is a scatterplot of life expectancy versus Internet use.  INTERNETANDLIFE

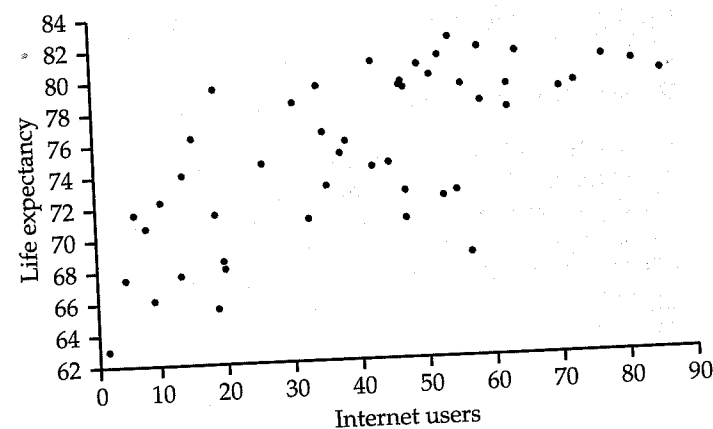
- Describe the relationship between these two variables.
- A friend looks at this plot and concludes that using the Internet will increase the length of your life. Write a short paragraph explaining why the association seen in the scatterplot does not provide a reason to draw this conclusion.

**2.21 Let's look at Europe.** Refer to the previous exercise. Figure 2.12 gives a scatterplot for the same data



**FIGURE 2.11** Scatterplot of life expectancy (in years) versus Internet users (per 100 people) for 181 countries, for Exercise 2.20.

for the 48 European countries in the data set. Compare this figure with Figure 2.11 which plots that data for all 181 countries in the data set. Write a paragraph summarizing the relationship between life expectancy and Internet use for European countries with an emphasis on how the European countries compare with the entire set of 181 countries. Be sure to take into account the fact the software used here automatically chooses the range of values for each axis so that the space in the plot is used efficiently. In this case, the range of values for Internet use is the same for both scatterplots but the range of values for life expectancy is quite different. INTERNETANDLIFE



**FIGURE 2.12** Scatterplot of life expectancy (in years) versus Internet users (per 100 people) for 48 European countries, for Exercise 2.21.

**2.22** How would you make a better plot? In the previous two exercises, we looked at the relationship between life expectancy and Internet use. First, we made a

scatterplot for all 181 countries in the data set. Then we made one for the subset of 48 European countries. Explain how you would construct a single plot to make a comparison between the European countries and the other countries in the data set. (Optional: Make the plot if you have software that can do what you need.) INTERNETANDLIFE

**2.23 Average temperatures.** Here are the average temperatures in degrees for Lafayette, Indiana, during the months of February through May: WLAFTemps

Month	February	March	April	May
Temperature (degrees F)	30	41	51	62

- (a) Explain why month should be the explanatory variable for examining this relationship.  
(b) Make a scatterplot and describe the relationship.

**2.24 Relationship between first test and final exam.** How strong is the relationship between the score on the first exam and the score on the final exam in an elementary statistics course? Here are data for eight students from such a course: STATCOURSE8

First-test score	153	144	162	149	127	118	158	153
Final-exam score	145	140	145	170	145	175	170	160

- (a) Which variable should play the role of the explanatory variable in describing this relationship?  
(b) Make a scatterplot and describe the relationship.  
(c) Give some possible reasons why this relationship is so weak.

**2.25 Relationship between second test and final exam.** Refer to the previous exercise. Here are the data for the second test and the final exam for the same students: STATCOURSE8

Second-test score	158	162	144	162	136	158	175	153
Final-exam score	145	140	145	170	145	175	170	160

- (a) Explain why you should use the second-test score as the explanatory variable.  
(b) Make a scatterplot and describe the relationship.  
(c) Why do you think the relationship between the second-test score and the final-exam score is stronger than the relationship between the first-test score and the final-exam score?

**2.26 Add an outlier to the plot.** Refer to the previous exercise. Add a ninth student whose scores on the second test and final exam would lead you to classify the additional data point as an outlier. Highlight the outlier on your scatterplot and describe the performance of the student on the second exam and final exam and why that leads to the conclusion that the result is an outlier. Give a possible reason for the performance of this student.

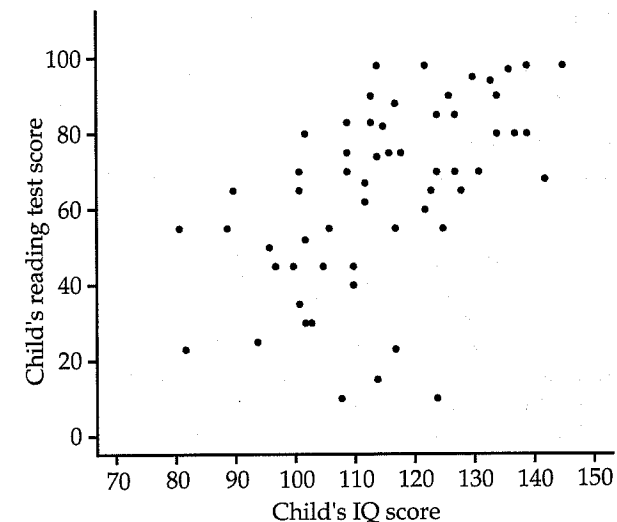
**2.27 Explanatory and response variables.** In each of the following situations, is it more reasonable to simply explore the relationship between the two variables or to view one of the variables as an explanatory variable and the other as a response variable? In the latter case, which is the explanatory variable and which is the response variable?

- (a) The weight of a child and the age of the child from birth to 10 years.  
(b) High school English grades and high school math grades.  
(c) The rental price of apartments and the number of bedrooms in the apartment.  
(d) The amount of sugar added to a cup of coffee and how sweet the coffee tastes.  
(e) The student evaluation scores for an instructor and the student evaluation scores for the course.

**2.28 Parents' income and student loans.** How well does the income of a college student's parents predict how much the student will borrow to pay for college? We have data on parents' income and college debt for a sample of 1200 recent college graduates. What are the explanatory and response variables? Are these variables categorical or quantitative? Do you expect a positive or negative association between these variables? Why?

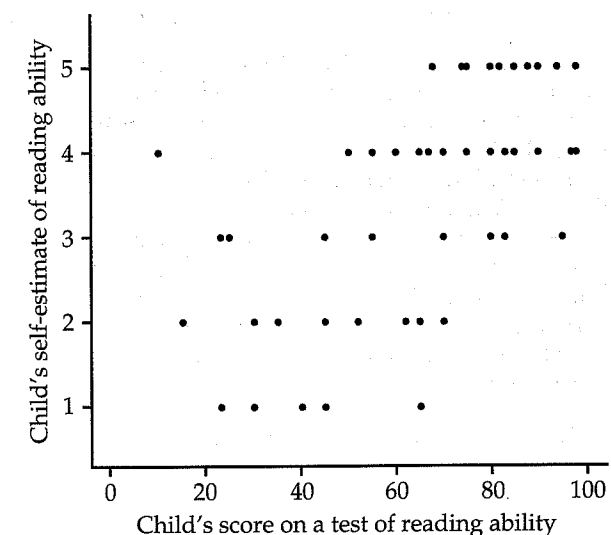
**2.29 Reading ability and IQ.** A study of reading ability in school children chose 60 fifth-grade children at random from a school. The researchers had the children's scores on an IQ test and on a test of reading ability.<sup>13</sup> Figure 2.13 plots reading test score (response) against IQ score (explanatory).

- (a) Explain why we should expect a positive association between IQ and reading score for children in the same grade. Does the scatterplot show a positive association?  
(b) A group of four points appear to be outliers. In what way do these children's IQ and reading scores deviate from the overall pattern?  
(c) Ignoring the outliers, is the association between IQ and reading scores roughly linear? Is it very strong? Explain your answers.




**FIGURE 2.13** IQ and reading test scores for 60 fifth-grade children, for Exercise 2.29.

**2.30 Can children estimate their reading ability?** The main purpose of the study cited in Exercise 2.29 was to ask whether school children can estimate their own reading ability. The researchers had the children's scores on a test of reading ability. They asked each child to estimate his or her reading level, on a scale from 1 (low) to 5 (high). Figure 2.14 is a scatterplot of the children's estimates (response) against their reading scores (explanatory).




**FIGURE 2.14** Reading test scores for 60 fifth-grade children and the children's estimates of their own reading levels, for Exercise 2.30.

- (a) What explains the “stair-step” pattern in the plot?
- (b) Is there an overall positive association between reading score and self-estimate?
- (c) There is one clear outlier. What is this child’s self-estimated reading level? Does this appear to over- or underestimate the level as measured by the test?

**2.31 Small falcons in Sweden.** Often the percent of an animal species in the wild that survive to breed again is lower following a successful breeding season. This is part of nature’s self-regulation, tending to keep population size stable. A study of merlins (small falcons) in northern Sweden observed the number of breeding pairs in an isolated area and the percent of males (banded for identification) who returned the next breeding season. Here are data for nine years.<sup>14</sup>  **FALCONS**

Pairs	28	29	29	29	30	32	33	38	38
Percent	82	83	70	61	69	58	43	50	47


- (a) Why is the response variable the *percent* of males that return rather than the *number* of males that return?
- (b) Make a scatterplot. To emphasize the pattern, also plot the mean response for years with 29 and 38 breeding pairs and draw lines connecting the mean responses for the six values of the explanatory variable.
- (c) Describe the pattern. Do the data support the theory that a smaller percent of birds survive following a successful breeding season?

**2.32 Biological clocks.** Many plants and animals have “biological clocks” that coordinate activities with the time of day. When researchers looked at the length of the biological cycles in the plant *Arabidopsis* by measuring leaf movements, they found that the length of the cycle is not always 24 hours. The researchers suspected that the plants adapt their clocks to their north-south position. Plants don’t know geography, but they do respond to light, so the researchers looked at the relationship between the plants’ cycle lengths and the length of the day on June 21 at their locations. The data file includes data on cycle length and day length, both in hours, for 146 plants.<sup>15</sup> Plot cycle length as the response variable against day length as the explanatory variable. Does there appear to be a positive association? Is it a strong association? Explain your answers.  **BIOCLOCKS**

**2.33 Social rejection and pain.** We often describe our emotional reaction to social rejection as “pain.” A clever

study asked whether social rejection causes activity in areas of the brain that are known to be activated by physical pain. If it does, we really do experience social and physical pain in similar ways. Subjects were first included and then deliberately excluded from a social activity while increases in blood flow in their brains were measured. After each activity, the subjects filled out questionnaires that assessed how excluded they felt.

Here are data for 13 subjects.<sup>16</sup> The explanatory variable is “social distress” measured by each subject’s questionnaire score after exclusion relative to the score after inclusion (values greater than 1 show the degree of distress caused by exclusion). The response variable is activity in the anterior cingulate cortex, a region of the brain that is activated by physical pain.

 **SOCIALREJECTION**

Subject	Social distress	Brain activity	Subject	Social distress	Brain activity
1	1.26	-0.055	8	2.18	0.025
2	1.85	-0.040	9	2.58	0.027
3	1.10	-0.026	10	2.75	0.033
4	2.50	-0.017	11	2.75	0.064
5	2.17	-0.017	12	3.33	0.077
6	2.67	0.017	13	3.65	0.124
7	2.01	0.021			

Plot brain activity against social distress. Describe the direction, form, and strength of the relationship, as well as any outliers. Do the data suggest that brain activity in the “pain” region is directly related to the distress from social exclusion?

**2.34 Business revenue and team value in the NBA.**

Management theory says that the value of a business should depend on its operating income, the income produced by the business after taxes. (Operating income excludes income from sales of assets and investments, which don’t reflect the actual business.) Total revenue, which ignores costs, should be less important. Debt includes borrowing for the construction of a new arena. Table 2.1 shows the value, operating income, debt, and revenue of the teams in the National Basketball Association (NBA).<sup>17</sup> Professional sports teams are generally privately owned, often by very wealthy individuals who may treat their team as a source of prestige rather than as a business.

 **NBA**

- (a) Plot team value against revenue. Describe the relationship.
- (b) Plot team value against debt. Describe the relationship.


**TABLE 2.1**

**NBA teams as businesses**


Team	Value (\$millions)	Revenue (\$millions)	Debt (\$millions)	Income (\$millions)
New York Knicks	613	0	208	29.6
Los Angeles Lakers	584	22	191	47.9
Chicago Bulls	504	11	165	55.4
Detroit Pistons	480	0	160	40.4
Cleveland Cavaliers	477	42	159	13.1
Houston Rockets	469	15	156	31.2
Dallas Mavericks	466	26	153	-13.6
Phoenix Suns	452	39	148	28.9
Boston Celtics	447	40	149	20.1
San Antonio Spurs	415	13	138	19.0
Toronto Raptors	400	40	138	27.7
Miami Heat	393	43	131	-1.1
Philadelphia 76ers	360	18	116	0.3
Utah Jazz	358	3	119	8.8
Washington Wizards	353	47	118	14.9
Sacramento Kings	350	24	117	7.0
Orlando Magic	349	29	100	6.2
Golden State Warriors	335	22	112	14.2
Denver Nuggets	329	14	112	-26.3
Portland Trail Blazers	307	34	114	-0.9
Atlanta Hawks	306	23	102	6.7
Indiana Pacers	303	16	101	-6.5
Minnesota Timberwolves	301	17	100	-5.7
Oklahoma City Thunder	300	47	82	-9.4
Los Angeles Clippers	297	0	99	10.7
New Jersey Nets	295	71	98	-0.9
Memphis Grizzlies	294	51	95	-3.2
New Orleans Hornets	285	35	95	3.2
Charlotte Bobcats	284	53	95	-4.9
Milwaukee Bucks	278	20	94	5.4

- (c) Plot team value against income. Describe the relationship. In your description be sure to pay attention to the teams that have negative income, that is, to the teams that lost money.

- (d) Write a short summary comparing the relationships that you described in parts (a), (b), and (c) of this exercise.

**2.35  Body mass and metabolic rate.** Metabolic rate, the rate at which the body consumes energy, is important in studies of weight gain, dieting, and exercise. The following table gives data on the lean body mass and resting metabolic rate for 12 women and 7 men who are subjects in a study of dieting. Lean body mass, given in kilograms, is a person’s weight leaving out all fat. Metabolic rate is measured in calories burned per 24 hours, the same calories used to describe the energy

content of foods. The researchers believe that lean body mass is an important influence on metabolic rate.

 **BODYMASS**

Subject	Sex	Mass	Rate	Subject	Sex	Mass	Rate
1	M	62.0	1792	11	F	40.3	1189
2	M	62.9	1666	12	F	33.1	913
3	F	36.1	995	13	M	51.9	1460
4	F	54.6	1425	14	F	42.4	1124
5	F	48.5	1396	15	F	34.5	1052
6	F	42.0	1418	16	F	51.1	1347
7	M	47.4	1362	17	F	41.2	1204
8	F	50.6	1502	18	M	51.9	1867
9	F	42.0	1256	19	M	46.9	1439
10	M	48.7	1614				

- (a) Make a scatterplot of the data, using different symbols or colors for men and women.

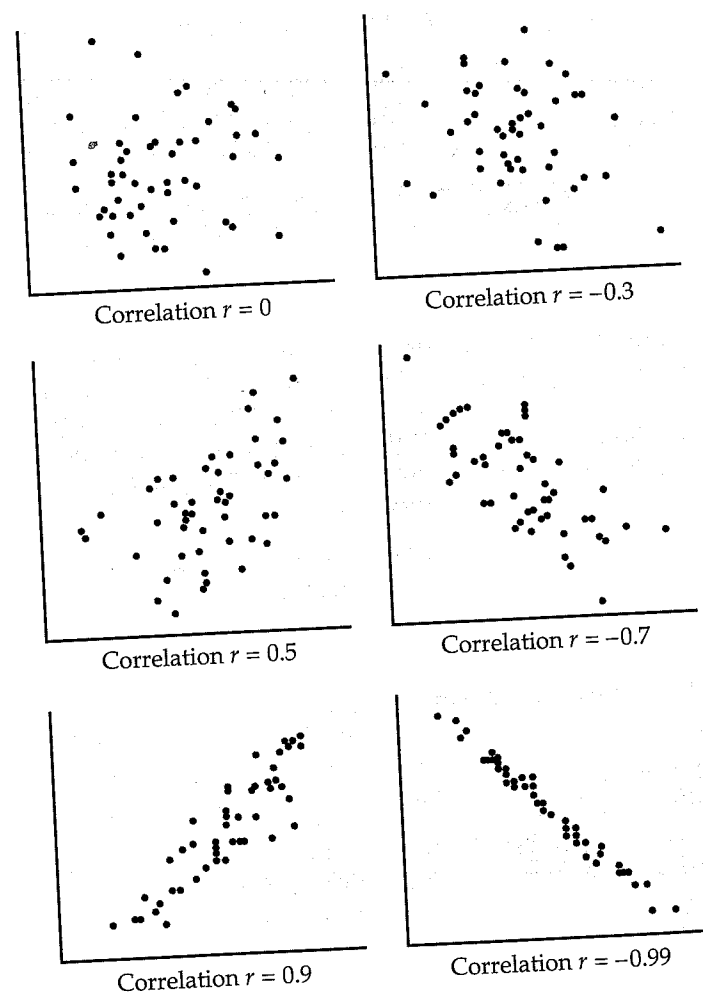


- Like the mean and standard deviation, the correlation is not resistant:  $r$  is strongly affected by a few outlying observations. Use  $r$  with caution when outliers appear in the scatterplot.



The scatterplots in Figure 2.16 illustrate how values of  $r$  closer to 1 or  $-1$  correspond to stronger linear relationships. To make the essential meaning of  $r$  clear, the standard deviations of both variables in these plots are equal and the horizontal and vertical scales are the same. In general, it is not so easy to guess the value of  $r$  from the appearance of a scatterplot. Remember that changing the plotting scales in a scatterplot may mislead our eyes, but it does not change the standardized values of the variables and therefore cannot change the correlation. To explore how extreme observations can influence  $r$ , use the *Correlation and Regression* applet available on the text CD and Web site.

Finally, remember that **correlation is not a complete description 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. (Because the formula for correlation uses the means and standard deviations, these measures are the proper choices to accompany a correlation.) Conclusions based on correlations alone may require rethinking in the light of a more complete description of the data.



**FIGURE 2.16** How the correlation  $r$  measures the direction and strength of a linear association.



### EXAMPLE

**2.17 Scoring of figure skating in the Olympics.** 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, to 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 in the example above do not contradict each other. They are simply 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, 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 must add 0.8 points to Pierre's scores to arrive at a fair comparison.

### SECTION 2.2 Summary

The **correlation  $r$**  measures the direction and strength of the linear (straight line) association between two quantitative variables  $x$  and  $y$ . Although you can calculate a correlation for any scatterplot,  $r$  measures only linear 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 relationship by how close it is to  $-1$  or  $1$ . Perfect correlation,  $r = \pm 1$ , occurs only when the points lie exactly on a straight line.

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  $r$ .

### SECTION 2.2 Exercises

**2.40 Thinking about correlation.** Figure 2.4 (page 88) is a scatterplot of 2007 debt versus 2006 debt for 24 countries. Is the correlation  $r$  for these data near  $-1$ , clearly negative but not near  $-1$ , near  $0$ , clearly positive but not near  $1$ , or near  $1$ ? Explain your answer. DEBT

#### 2.41 Brand names and generic products.

(a) If a store always prices its generic "store brand" products at 90% of the brand name products' prices, what would be the correlation between the prices of the brand name products and the store brand products? (Hint: Draw a scatterplot for several prices.)

(b) If the store always prices its generic products \$1 less than the corresponding brand name products, then what would be the correlation between the prices of the brand name products and the store brand products?


**2.42 Strong association but no correlation.** Here is a data set that illustrates an important point about correlation: CORRELATION


X	20	30	40	50	60
Y	10	30	50	30	10

- Make a scatterplot of  $Y$  versus  $X$ .
- Describe the relationship between  $Y$  and  $X$ . Is it weak or strong? Is it linear?
- Find that the correlation between  $Y$  and  $X$ .
- What important point about correlation does this exercise illustrate?





106 CHAPTER 2 Looking at Data—Relationships

**2.43 Alcohol and carbohydrates in beer.** Figure 2.10 (page 95) gives a scatterplot of carbohydrates versus percent alcohol in 86 brands of beer. Compute the correlation for these data. 


**2.44 Alcohol and carbohydrates in beer revisited.** Refer to the previous exercise. The data that you used to compute the correlation includes an outlier. 

- (a) Remove the outlier and recompute the correlation.
- (b) Write a short paragraph about the possible effects of outliers on a correlation using this example to illustrate your ideas.

**2.45 Will you live longer if you use the Internet?** Figure 2.11 (page 96) is a scatterplot of the number of life expectancy versus Internet users per 100 people for 181 countries. In Exercise 2.20 you described this relationship. Make a plot of the data similar to Figure 2.11 and report the correlation. 


**2.46 Let's look at Europe.** Refer to the previous exercise. Figure 2.12 (page 96) gives a scatterplot for the same data for the 48 European countries in the data set. 

- (a) Make a plot of the data similar to Figure 2.12.
- (b) Report the correlation.
- (c) Summarize the differences and similarities between the relationship for all 181 countries and the results that you found in this exercise for the European countries only.

**2.47 Second test and final exam.** In Exercise 2.25 you looked at the relationship between the score on the second test and the score on the final exam in an elementary statistics course. Here are the data: 

Second-test score	158	162	144	162	136	158	175	153
Final-exam score	145	140	145	170	145	175	170	160

- (a) Find the correlation between these two variables.
- (b) Do you think that the correlation between the first test and the final exam should be higher than, approximately equal to, or lower than the correlation between the second test and the final exam? Give a reason for your answer.


**2.48 First test and final exam.** Refer to the previous exercise. Here are the data for the first test and the final exam. 


First-exam score	153	144	162	149	127	118	158	153
Final-exam score	145	140	145	170	145	175	170	160


- (a) Find the correlation between these two variables.
- (b) In Exercise 2.24 we noted that the relationship between these two variables is weak. Does your calculation of the correlation support this statement? Explain your answer.
- (c) Examine part (b) of the previous exercise. Does your calculation agree with your prediction?

**2.49 The effect of a different point.** Examine the data in the Exercise 2.47 and add a ninth student who has low scores on the second test and the final exam and fits the overall pattern of the other scores in the data set. Calculate the correlation and compare it with the correlation that you calculated in Exercise 2.47. Write a short summary of your findings.

**2.50 The effect of an outlier.** Refer to the Exercise 2.47. Add a ninth student whose scores on the second test and final exam would lead you to classify the additional data point as an outlier. Recalculate the correlation with this additional case and summarize the effect it as on the value of the correlation.

**2.51 NBA teams.** Table 2.1 (page 99) gives the values of the 30 teams in the National Basketball Association, along with their total revenues, debt, and operating incomes. You made scatterplots of value against the three explanatory variables in Exercise 2.34. Find the correlations of team value with revenue, with debt, and with operating income. Do you think that the values of  $r$  provide a good first comparison of what the plots show about predicting value? 

**2.52 Correlations measure strong and weak linear associations.** Your scatterplots for Exercises 2.32 (page 98) and 2.36 (Table 2.2, page 100) illustrate a quite weak linear association and a very strong linear association. Find the correlations that go with these plots. It isn't surprising that a laboratory experiment on physical behavior (the icicles) gives a much stronger correlation than field data on living things (the biological clock). How strong a correlation must be to interest scientists depends on the field of study. 

**2.53 Heights of people who date each other.** A student wonders if tall women tend to date taller men than do short women. She measures herself, her dormitory roommate, and the women in the adjoining rooms; then she measures the next man each woman dates. Here are the data (heights in inches): 

Women ( $x$ )	66	64	66	65	70	65
Men ( $y$ )	72	68	70	68	71	65

(a) Make a scatterplot of these data. Based on the scatterplot, do you expect the correlation to be positive or negative? Near  $\pm 1$  or not?

(b) Find the correlation  $r$  between the heights of the men and women.


(c) How would  $r$  change if all the men were 6 inches shorter than the heights given in the table? Does the correlation tell us whether women tend to date men taller than themselves?

(d) If heights were measured in centimeters rather than inches, how would the correlation change? (There are 2.54 centimeters in an inch.)

(e) If every woman dated a man exactly 3 inches taller than herself, what would be the correlation between male and female heights?

**2.54 An interesting set of data.** Make a scatterplot of the following data:

$x$	1	2	3	4	10	10
$y$	1	3	3	5	1	11

Use your calculator to show that the correlation is about 0.5. 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? 

**2.55 Use the applet.** You are going to use the *Correlation and Regression* applet to make different scatterplots with 10 points that have correlation close to 0.8. *Many patterns can have the same correlation. Always plot your data before you trust a correlation.*

(a) Stop after adding the first 2 points. What is the value of the correlation? Why does it have this value no matter where the 2 points are located?


(b) Make a lower-left to upper-right pattern of 10 points with correlation about  $r = 0.8$ . (You can drag points up or down to adjust  $r$  after you have 10 points.) Make a rough sketch of your scatterplot.

(c) Make another scatterplot, this time with 9 points in a vertical stack at the left of the plot. Add one point far to the right and move it until the correlation is close to 0.8. Make a rough sketch of your scatterplot.

(d) Make yet another scatterplot, this time 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 quite smooth curve with correlation close to 0.8. Make a rough sketch of this scatterplot also.

**2.56 Use the applet.** Go to the *Correlation and Regression* applet. Click on the scatterplot to create a group of 10 points in the lower-right corner of the scatterplot with a strong straight-line negative pattern (correlation about  $-0.9$ ).

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

(b)  Drag this last point down until it is opposite the group of 10 points. How small can you make the correlation? Can you make the correlation positive? *A single outlier can greatly strengthen or weaken a correlation. Always plot your data to check for outlying points.*

**2.57 What is the correlation?** Suppose that women always married men 2 years older than themselves. Draw a scatterplot of the ages of 5 married couples, with the wife's age as the explanatory variable. What is the correlation  $r$  for your data? Why?

**2.58 High correlation does not mean that the values are the same.** Investment reports often include correlations. Following a table of correlations among mutual funds, a report adds, "Two funds can have perfect correlation, yet different levels of risk. For example, Fund A and Fund B may be perfectly correlated, yet Fund A moves 20% whenever Fund B moves 10%." Write a brief explanation, for someone who knows no statistics, of how this can happen. Include a sketch to illustrate your explanation.

**2.59 Student ratings of teachers.** 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.

**2.60 What's wrong?** Each of the following statements contains a blunder. Explain in each case what is wrong.

(a) "There is a high correlation between the age of American workers and their occupation."

(b) "We found a high correlation ( $r = 1.19$ ) between students' ratings of faculty teaching and ratings made by other faculty members."

(c) "The correlation between the gender of a group of students and the color of their cell phone was  $r = 0.23$ ."