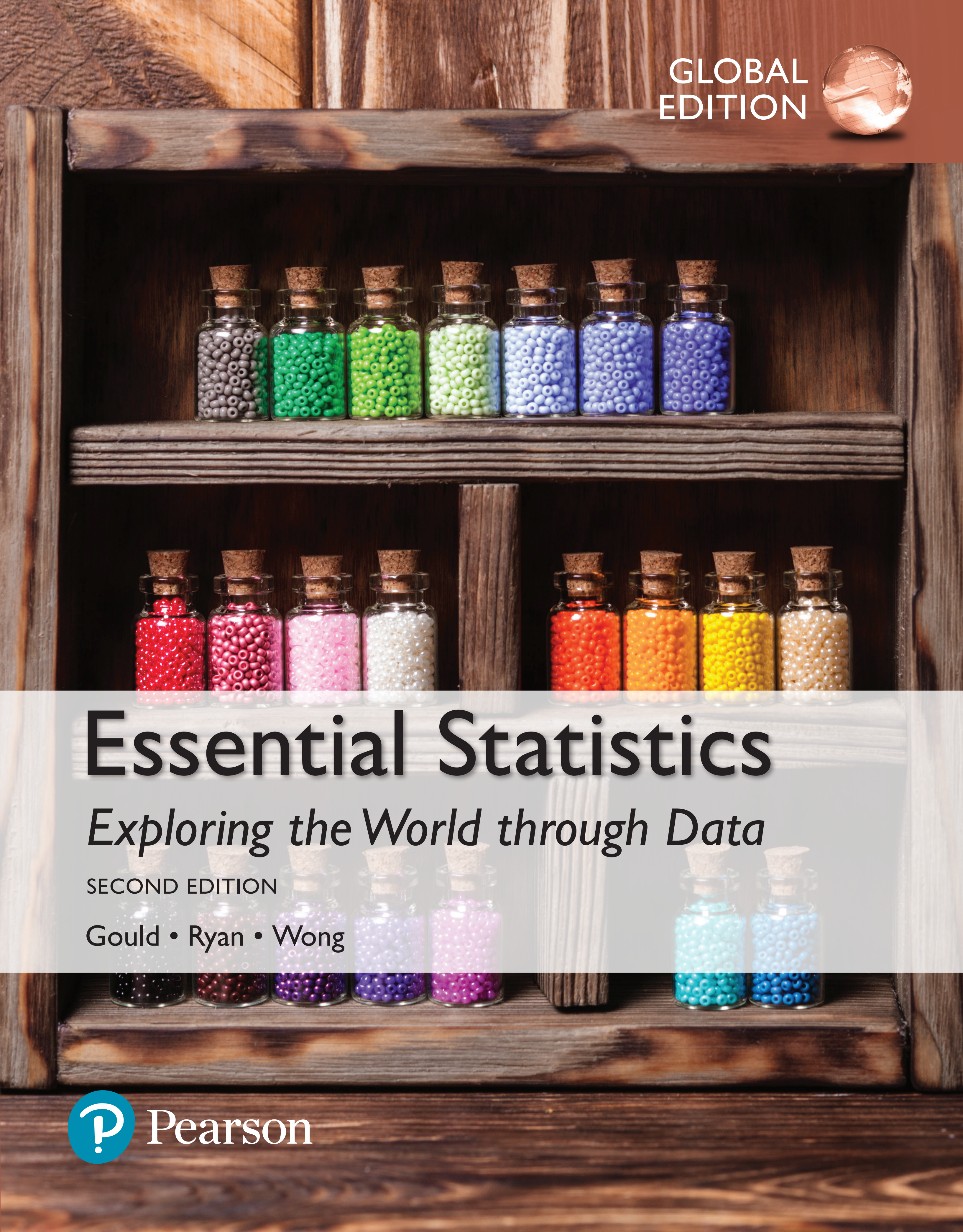


GLOBAL EDITION

Essential Statistics

Exploring the World through Data

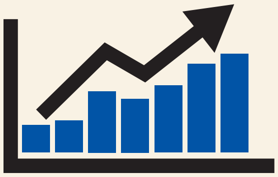
SECOND EDITION

Gould • Ryan • Wong

Pearson

BREAK THROUGH
To improving results

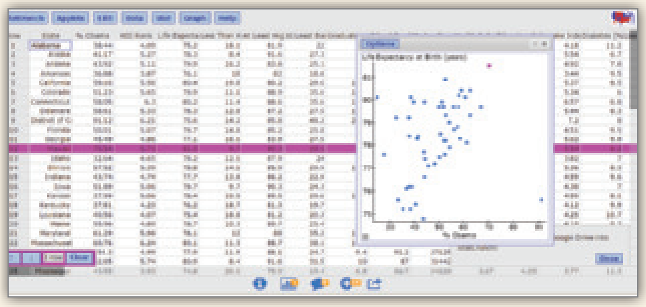
Available in MyStatLab™ for your Introductory Statistics Courses



MyStatLab is the market-leading online resource for learning and teaching statistics.

Leverage the Power of StatCrunch

MyStatLab leverages the power of StatCrunch—powerful, web-based statistics software. Integrated into MyStatLab, students can easily analyze data from their exercises and etext. In addition, access to the full online community allows users to take advantage of a wide variety of resources and applications at www.statcrunch.com.

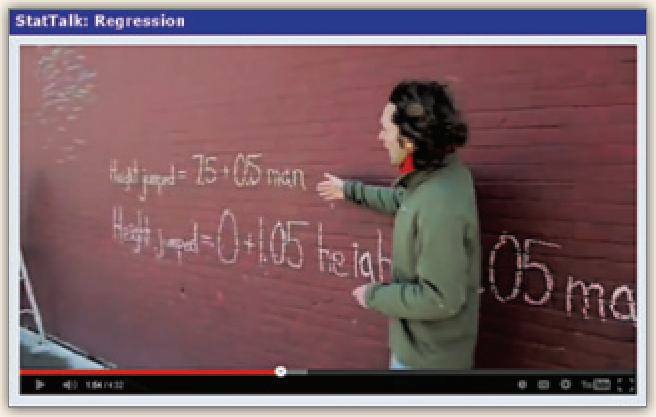


Bring Statistics to Life

Virtually flip coins, roll dice, draw cards, and interact with animations on your mobile device with the extensive menu of experiments and applets in StatCrunch. Offering a number of ways to practice resampling procedures, such as permutation tests and bootstrap confidence intervals, StatCrunch is a complete and modern solution.

Real-World Statistics

MyStatLab video resources help foster conceptual understanding. StatTalk Videos, hosted by fun-loving statistician Andrew Vickers, demonstrate important statistical concepts through interesting stories and real-life events. This series of 24 videos includes assignable questions built in MyStatLab and an instructor's guide.



www.mystatlab.com

College admission offices sometimes report correlations between students' Scholastic Aptitude Test (SAT) scores and their first-year GPAs. If the association is linear and the correlation is high, this justifies using the SAT to make admissions decisions, because a high correlation would indicate a strong association between SAT scores and academic performance. A positive correlation means that students who score above average on the SAT tend to get above-average grades. Conversely, those who score below average on the SAT tend to get below-average grades. Note that we're careful to say "tend to." Certainly, some students with low SAT scores do very well, and some with high SAT scores struggle to pass their classes. The correlation coefficient does not tell us about individual students; it tells us about the overall trend.

! Caution

Linearity Is Important
The correlation coefficient is interpretable only for linear associations.

More Context: Correlation Does Not Mean Causation!

Quite often, you'll hear someone use the correlation coefficient to support a claim of cause and effect. For example, one of the authors once read that a politician wanted to close liquor stores in a city because there was a positive correlation between the number of liquor stores in a neighborhood and the amount of crime.

As you learned in Chapter 1, we can't form cause-and-effect conclusions from observational studies. If your data came from an observational study, it doesn't matter how strong the correlation is. Even a correlation near 1 is not enough for us to conclude that changing one variable (closing down liquor stores) will lead to a change in the other variable (crime rate).

A positive correlation also exists between the number of blankets sold in Canada per week and the number of brush fires in Australia per week. Are brush fires in Australia caused by cold Canadians? Probably not. The correlation is likely to be the result of weather. When it is winter in Canada, people buy blankets. When winter is happening in Canada, summer is happening in Australia (which is located in the Southern Hemisphere), and summer is brush-fire season.

What, then, can we conclude from the fact that the number of liquor stores in a neighborhood is positively correlated with the crime rate in that neighborhood? Only that neighborhoods with a higher-than-average number of liquor stores typically (but not always) have a higher-than-average crime rate.

If you learn nothing else from this book, remember this: No matter how tempting, do *not* conclude that a cause-and-effect relationship between two variables exists just because there is a correlation, no matter how close to $+1$ or -1 that correlation might be!

KEY POINT Correlation does not imply causation.

Finding the Correlation Coefficient

The correlation coefficient is best determined through the use of technology. We calculate a correlation coefficient by first converting each observation to a *z*-score, using the appropriate variable's mean and standard deviation. For example, to find the correlation coefficient that measures the strength of the linear relation between weight and height, we first convert each person's weight and height to *z*-scores. The next step is to multiply the observations' *z*-scores together. If both are positive or both negative—meaning that both *z*-scores are above average or both are below average—then the product is a positive number. In a strong positive association, most of these products are positive values. In a strong negative association, however, observations above average on one variable tend to be below average on the other variable. In this case, one *z*-score is negative and one positive, so the product is negative. Thus, in a strong negative association, most *z*-score products are negative.

Looking Back

***z*-Scores**
Recall that *z*-scores show how many standard deviations a measurement is from the mean. To find a *z*-score from a measurement, first subtract the mean and then divide the difference by the standard deviation.

To find the correlation coefficient, add the products of *z*-scores together and divide by $n - 1$ (where n is the number of observed pairs in the sample). In mathematical terms, we get

$$\textbf{Formula 4.1: } r = \frac{\sum z_x z_y}{n - 1}$$

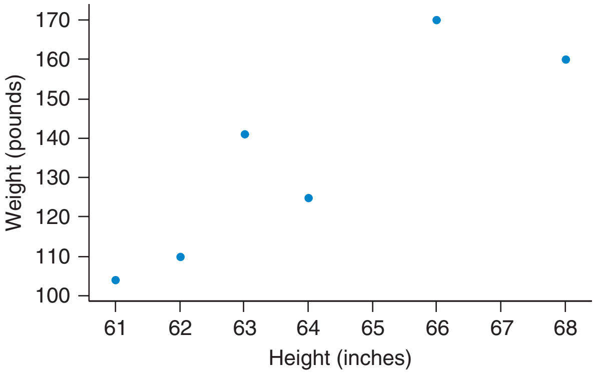
The following example illustrates how to use Formula 4.1 in a calculation.

EXAMPLE 2 Heights and Weights of Six Women

Figure 4.10a shows the scatterplot for heights and weights of six women.

QUESTION Using the data provided, find the correlation coefficient of the linear association between heights (inches) and weights (pounds) for these six women.

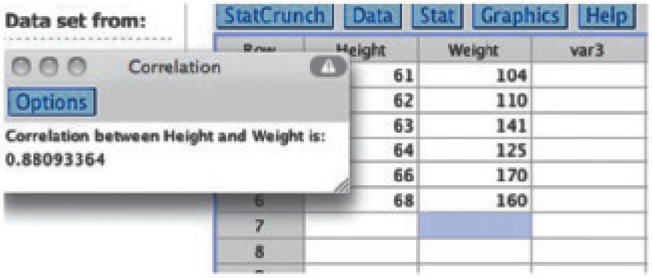
Heights	61	62	63	64	66	68
Weights	104	110	141	125	170	160



▶ **FIGURE 4.10a** Scatterplot showing heights and weights of six women.

SOLUTION Before proceeding, we verify that the conditions hold. Figure 4.10a suggests that a straight line is an acceptable model; a straight line through the data might summarize the trend, although this is hard to see with so few points.

Next, we calculate the correlation coefficient. Ordinarily, we use technology to do this, and Figure 4.10b shows the output from StatCrunch, which gives us the value $r = 0.88093364$.



Data set from:

Correlation

Correlation between Height and Weight is:
0.88093364

Row	Height	Weight	var3
	61	104	
	62	110	
	63	141	
	64	125	
	66	170	
6	68	160	
7			
8			

▶ **FIGURE 4.10b** StatCrunch, like all statistical software, lets you calculate the correlation between any two columns of numerical data you choose.

Because the sample size is small, we confirm this output using Formula 4.1. It is helpful to go through the steps of this calculation to better understand how the correlation coefficient measures linear relationships between variables.

The first step is to calculate average values of height and weight and then determine the standard deviation for each.

$$\text{For the height: } \bar{x} = 64 \text{ and } s_x = 2.608$$

$$\text{For the weight: } \bar{y} = 135 \text{ and } s_y = 26.73$$

Next we convert all of the points to pairs of z-scores and multiply them together. For example, for the woman who is 68 inches tall and weighs 160 pounds,

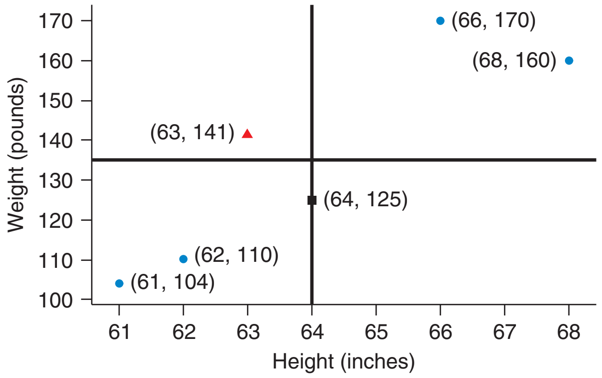
$$z_x = \frac{x - \bar{x}}{s_x} = \frac{68 - 64}{2.608} = \frac{4}{2.608} = 1.53$$

$$z_y = \frac{y - \bar{y}}{s_y} = \frac{160 - 135}{26.73} = \frac{25}{26.73} = 0.94$$

The product is

$$z_x \times z_y = 1.53 \times 0.94 = 1.44$$

Note that this product is positive and shows up in the upper right quadrant in Figure 4.10c.



◀ **FIGURE 4.10c** The same scatterplot as in 4.10a, but with the plot divided into quadrants based on average height and weight. Points represented with blue circles contribute a positive value to the correlation coefficient (positive times positive is positive, or negative times negative equals a positive). The red triangle represents an observation that contributes negatively (a negative z-score times a positive z-score is negative), and the black square contributes nothing because one of the z-scores is 0.

Figure 4.10c can help you visualize the rest of the process. The two blue circles in the upper right portion represent observations that are above average in both variables, so both z-scores are positive. The two blue circles in the lower left region represent observations that are below average in both variables; the products of the two negative z-scores are positive, so they add to the correlation. The red triangle has a positive z-score for weight (it is above average) but a negative z-score for height, so the product is negative. The black square is a point that makes no contribution to the correlation coefficient. This person is of average height, so her z-score for height is 0.

The correlation between height and weight for these six women comes out to be about 0.881.

CONCLUSION The correlation coefficient for the linear association of weights and heights of these six women is $r = 0.881$. Thus, there is a strong positive correlation between height and weight for these women. Taller women tend to weigh more.

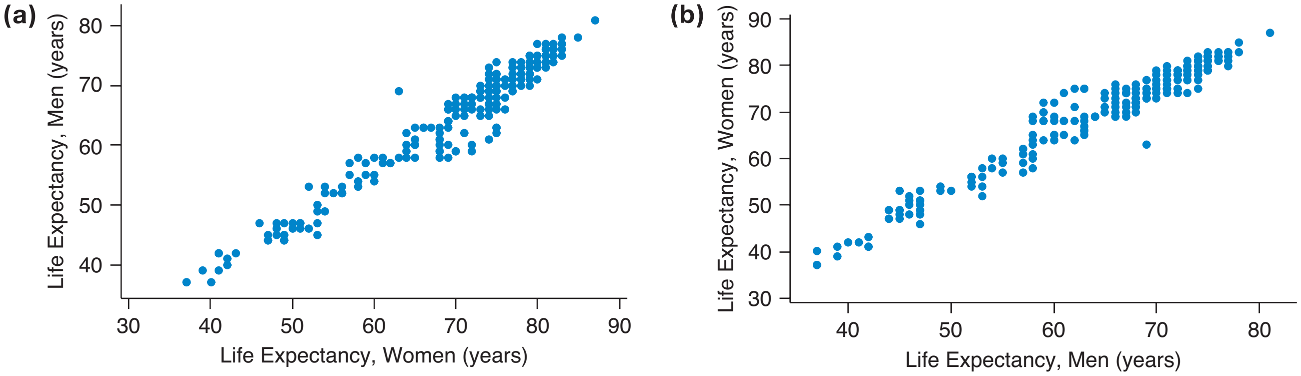
TRY THIS! Exercise 4.21a

Understanding the Correlation Coefficient

The correlation coefficient has a few features you should know about when interpreting a value of r or deciding whether you should compute the value.

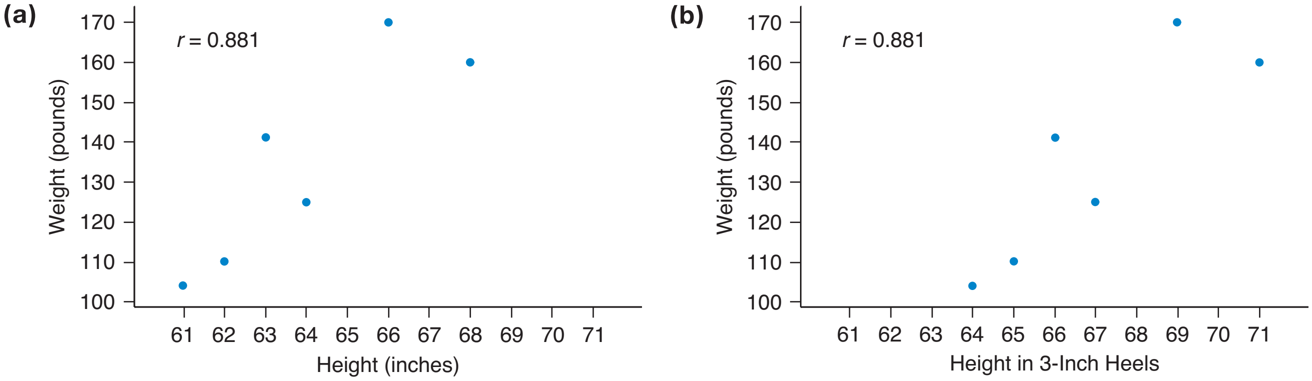
- *Changing the order of the variables does not change r.* This means that if the correlation between life expectancy for men and women is 0.977, then the correlation between life expectancy for women and men is also 0.977. This makes sense because the correlation measures the strength of the linear relationship between *x* and *y*, and that strength will be the same no matter which variable gets plotted on the horizontal axis and which on the vertical.

Figure 4.11a and b have the same correlation; we've just swapped axes.



▲ **FIGURE 4.11** Scatterplots showing the relationship between men's and women's life expectancies for various countries. **(a)** Women's life expectancy is plotted on the *x*-axis. **(b)** Men's life expectancy is plotted on the *x*-axis. (Sources: http://www.overpopulation.com and Fathom™ Sample Documents)

- *Adding a constant or multiplying by a positive constant does not affect r*. The correlation between the heights and weights of the six women in Example 2 was 0.881. What would happen if all six women in the sample had been asked to wear 3-inch platform heels when their heights were measured? Everyone would have been 3 inches taller. Would this have changed the value of *r*? Intuitively, you should sense that it wouldn't. Figure 4.12a shows a scatterplot of the original data, and Figure 4.12b shows the data with the women in 3-inch heels.



▲ **FIGURE 4.12** **(a)** A repeat of the scatterplot of height and weight for six women. **(b)** The same women in 3-inch heels. The correlation remains the same.

We haven't changed the strength of the relationship. All we've done is shift the points on the scatterplot 3 inches to the right. But shifting the points doesn't change the relationship between height and weight. We can verify that the correlation is unchanged by looking at the formula. The heights will have the same *z*-scores both before and after the women put on the shoes; since everyone "grows" by the same amount, everyone is still the same number of standard deviations away from the average, which also "grows" by 3 inches. As another example, if science found a way to add 5 years to the life expectancy of men in all countries in the world, the correlation between life expectancies for men and women would still be the same.

More generally, we can add a constant (a fixed value) to all of the values of one variable, or of both variables, and not affect the correlation coefficient.

For the very same reason, we can multiply either or both variables by positive constants without changing *r*. For example, to convert the women's heights from inches to feet, we multiply their heights by 1/12. Doing this does not change how strong the association is; it merely changes the units we're using to measure height. Because the strength of the association does not change, the correlation coefficient does not change.

- *The correlation coefficient is unitless.* Height is measured in inches and weight in pounds, but *r* has no units because the *z*-scores have no units. This means that we will get the same value for correlation whether we measure height in inches, meters, or fathoms.
- *Linear, linear, linear.* We've said it before, but we'll say it again: We're talking only about linear relationships here. The correlation can be misleading if you do not have a linear relationship. Figure 4.13a through d illustrate the fact that different nonlinear patterns can have the same correlation. All of these graphs have $r = 0.817$, but the graphs have very different shapes. The take-home message is that the correlation alone does not tell us much about the shape of a graph. We must also know that the relationship is linear to make sense of the correlation.

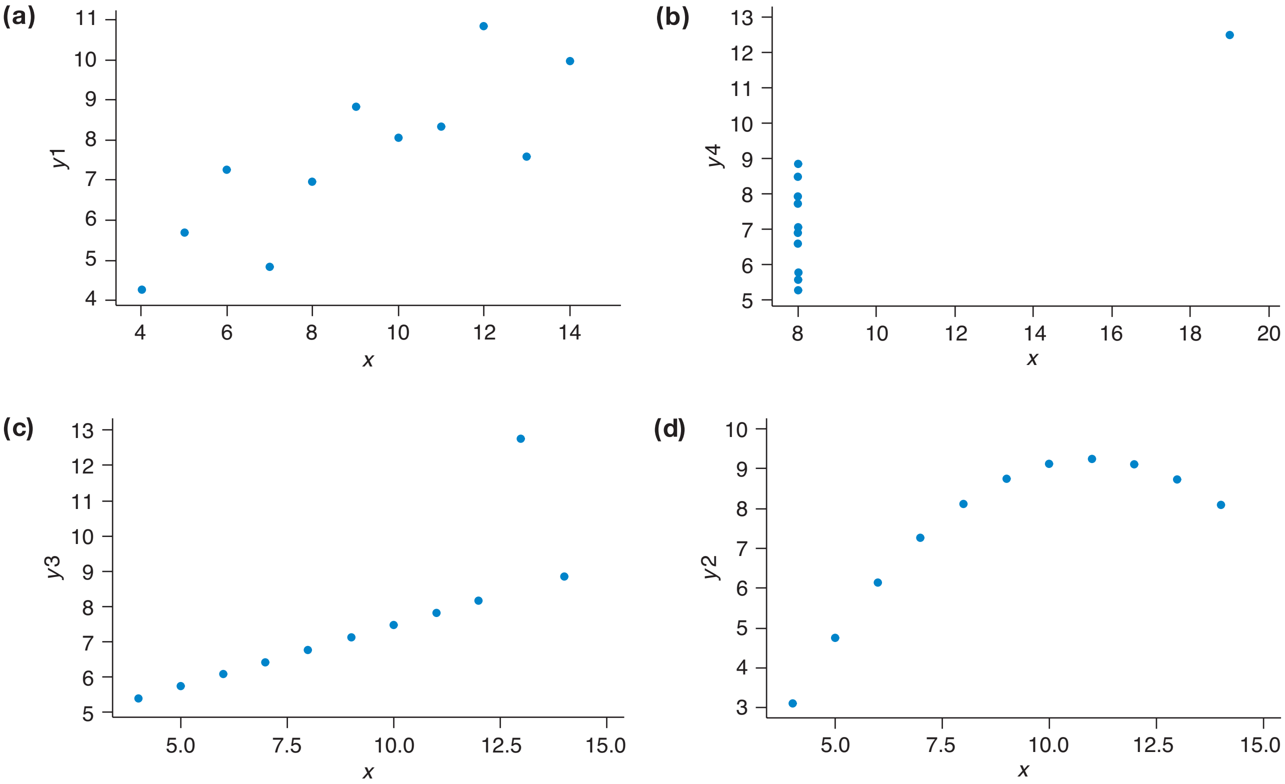
Remember: *Always* make a graph of your data. If the trend is nonlinear, the correlation (and, as you'll see in the next section, other statistics) can be very misleading.

! Caution

Correlation Coefficient and Linearity

A value of *r* close to 1 or −1 does *not* tell you that the relationship is linear. You must check visually; otherwise, your interpretation of the correlation coefficient might be wrong.

KEY POINT The correlation coefficient does not tell you whether an association is linear. However, if you already know that the association is linear, then the correlation coefficient tells you how strong the association is.



▲ **FIGURE 4.13 (a–d)** Four scatterplots with the same correlation coefficient of 0.817 have very different shapes. The correlation coefficient is meaningful only if the trend is linear. (Source: Anscombe, F. 1973)

SNAPSHOT THE CORRELATION COEFFICIENT

WHAT IS IT? ▶ Correlation coefficient.

WHAT DOES IT DO? ▶ Measures the strength of a linear association.

HOW DOES IT DO IT? ▶ By comparing *z*-scores of the two variables. The products of the two *z*-scores for each point are averaged.

HOW IS IT USED? ▶ The sign tells us whether the trend is positive (+) or negative (−). The value tells us the strength. If the value is close to 1 or −1, then the points are tightly clustered about a line; if the value is close to 0, then there is no linear association.

Note: The correlation coefficient can be interpreted only with linear associations.

SECTION 4.3

Modeling Linear Trends

How much more do people tend to weigh for each additional inch in height? How much value do cars lose each year as they age? Are home run hitters good for their teams? Can we predict how much space a book will take on a bookshelf just by knowing how many pages are in the book? It's not enough to remark that a trend exists. To make a prediction based on data, we need to measure the trend and the strength of the trend.

To measure the trend, we're going to perform a bit of statistical sleight of hand. Rather than interpret the data themselves, we will substitute a model of the data and interpret the model. The model consists of an equation and a set of conditions that describe when the model will be appropriate. Ideally, this equation is a very concise and accurate description of the data; if so, the model is a good fit. When the model is a good fit to the data, any understanding we gain about the model accurately applies to our understanding of the real world. If the model is a bad fit, however, then our understanding of real situations might be seriously flawed.

The Regression Line

The **regression line** is a tool for making predictions about future observed values. It also provides us with a useful way of summarizing a linear relationship. Recall from Chapter 3 that we could summarize a sample distribution with a mean and a standard deviation. The regression line works the same way: It reduces a linear relationship to its bare essentials and enables us to analyze a relationship without being distracted by small details.

Review: Equation of a Line The regression line is given by an equation for a straight line. Recall from algebra that equations for straight lines contain a ***y* intercept** and a **slope**. The equation for a straight line is

$$y = mx + b$$

The letter *m* represents the slope, which tells how steep the line is, and the letter *b* represents the *y*-intercept, which is the value of *y* when $x = 0$.

Statisticians write the equation of a line slightly differently and put the intercept first; they use the letter *a* for the intercept and *b* for the slope and write

$$y = a + bx$$