The Granger Collection, New York

CARL FRIEDRICH GAUSS

The Gaussian Distributions

By age 18, *Carl Friedrich Gauss (1777–1855)* had independently discovered the binomial theorem, the arithmetic-geometric mean, the law of quadratic reciprocity, and the prime-number theorem. By age 21, he had made one of his most important discoveries: the construction of a regular 17-sided polygon by ruler and compasses, the first advance in the field since the early Greeks.

Gauss's contributions to the field of statistics include the method of least squares and the normal distribution, frequently called a Gaussian distribution in his honor. The normal distribution arose as a result of his attempts to account for the variation in individual observations of stellar locations. In 1801, Gauss predicted the position of a newly discovered asteroid, Ceres. Although he did not disclose his methods at the time, Gauss had used his least-squares approximation method. When the French mathematician Legendre published his version of the method of least-squares in 1805, Gauss's response was that he had known the method for years but had never felt the need to publish. This was his frequent response to the discoveries of fellow scientists. Gauss was not being boastful; rather, he cared little for fame.

In 1807, Gauss was appointed director of the University of Göttingen Observatory, where he worked for the rest of his life. He made important discoveries in number theory, algebra, conic sections and elliptic orbits, hypergeometric functions, infinite series, differential equations, differential geometry, physics, and astronomy. Five years before Samuel Morse, Gauss built a primitive telegraph device that could send messages up to a mile away. It is probably fair to say that Archimedes, Newton, and Gauss are in a league of their own among the great mathematicians.

Gauss's contributions to the field of statistics include the method of least-squares and the normal distribution, frequently called a Gaussian distribution in his honor.

chapter 4

More on Two-Variable Data

- 4.1 Transforming Relationships
- 4.2 Cautions about Correlation and Regression
- 4.3 Relations in Categorical Data
- Chapter Review

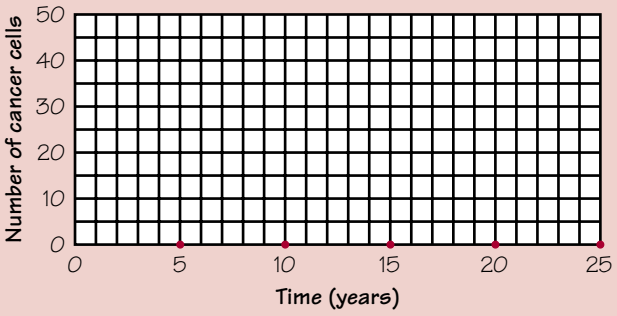
ACTIVITY 4 Modeling the Spread of Cancer in the Body

Materials: a regular six-sided die for each student; transparency grid; copy of grid for each student

Cancer begins with one cell, which divides into two cells.[1] Then these two cells divide and produce four cells. All the cancer cells produced are exactly like the original cell. This process continues until there is some intervention such as radiation or chemotherapy to interrupt the spread of the disease or until the patient dies. In this activity you will simulate the spread of cancer cells in the body.

1. Select one student to represent the original bad cell. That person rolls the die repeatedly, each roll representing a year. The number 5 will signal a cell division. When a 5 is rolled, a new student from the class will receive a die and join the original student (bad cell), so that there are now two cancer cells. These two students should be physically separated from the rest of the class, perhaps in a corner of the room.

2. As the die is rolled, another student will plot points on a transparency grid on the overhead projector. "Time," from 0 to 25 years, is marked on the horizontal axis, and the "Number of cancer cells," from 0 to 50, is on the vertical axis. The points on the grid will form a scatterplot.



3. At a signal from the teacher, each "cancer cell" will roll his or her die. If anyone rolls the number 5, a new student from the class receives a die and joins the circle of cancer cells. The total number of cancer cells is counted, and the next point on the grid is plotted. The simulation continues until all students in the class have become cancer cells.

Questions:
Do the points show a pattern? If so, is the pattern linear? Is it a curved pattern? What mathematical function would best describe the pattern of points?

Each student should keep a copy of the transparency grid with the plotted points. We will analyze the results later in the chapter, after establishing some principles.

4.1 TRANSFORMING RELATIONSHIPS

How is the weight of an animal's brain related to the weight of its body? Figure 4.1 is a scatterplot of brain weight against body weight for 96 species of mammals.[2] This line is the least-squares regression line for predicting brain weight from body weight. The outliers are interesting. We might say that dolphins and humans are smart, hippos are dumb, and elephants are just big. That's because dolphins and humans have larger brains than their body weights suggest, hippos have smaller brains, and the elephant is much heavier than any other mammal in both body and brain.

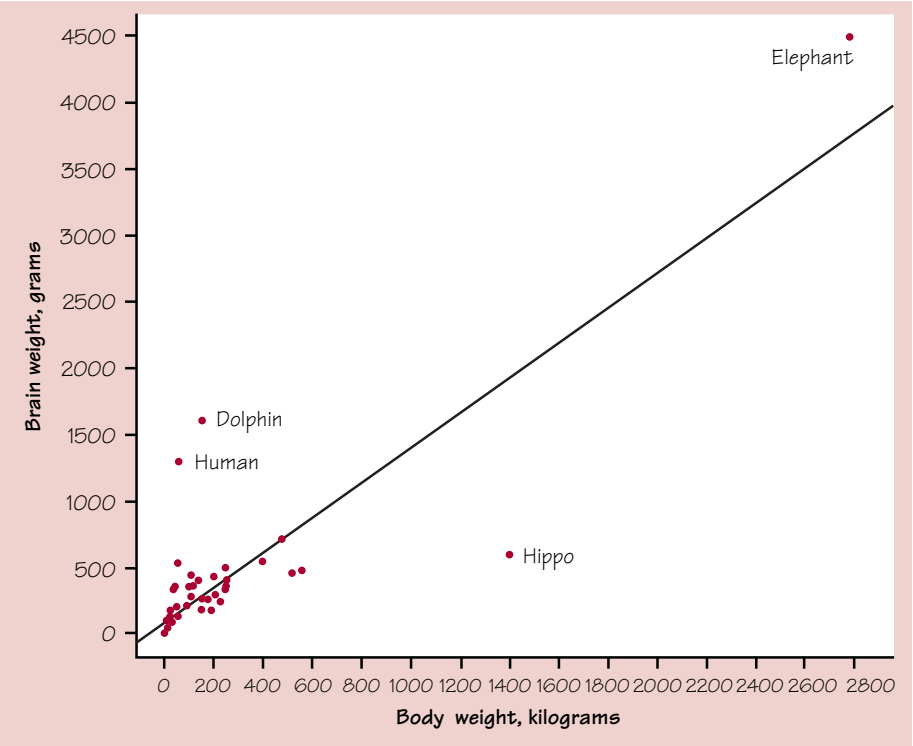


FIGURE 4.1 Scatterplot of brain weight against body weigth for 96 species of mammals.

EXAMPLE 4.1 MODELING MAMMAL BRAIN WEIGHT VERSUS BODY WEIGHT

The plot in Figure 4.1 is not very satisfactory. Most mammals are so small relative to elephants and hippos that their points overlap to form a blob in the lower-left corner of the plot. The correlation between brain weight and body weight is $r = 0.86$, but this is misleading. If we remove the elephant, the correlation for the other 95 species is $r = 0.50$. Figure 4.2 is a scatterplot of the data with the four outliers removed to allow a closer look at the other 92 observations. We can now see that the relationship is not linear. It bends to the right as body weight increases.

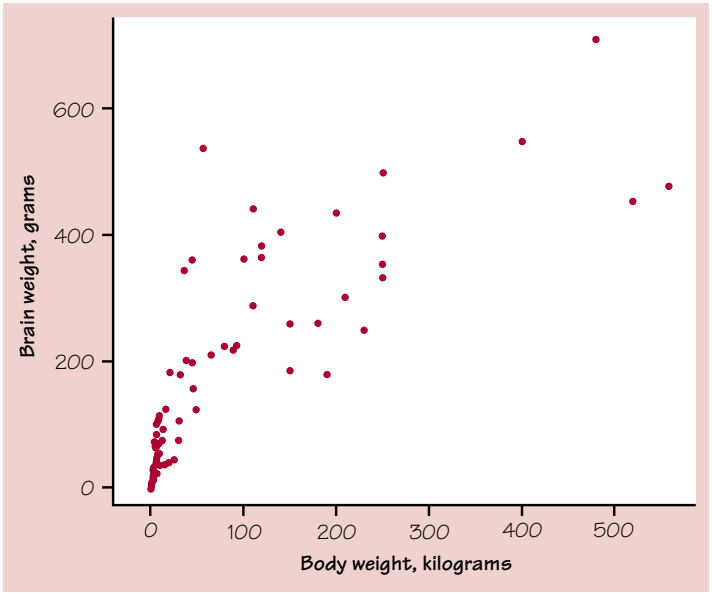


FIGURE 4.2 Brain weight against body weight for mammals, with outliers removed.

Biologists know that data on sizes often behave better if we take logarithms before doing more analysis. Figure 4.3 plots the logarithm of brain weight against the logarithm of body weight for all 96 species. The effect is almost magical. There are no longer any extreme outliers or very influential observations. The pattern is very linear, with correlation $r = 0.96$. The vertical spread about the least-squares line is similar everywhere, so that predictions of brain weight from body weight will be about equally precise for any body weight (in the log scale).

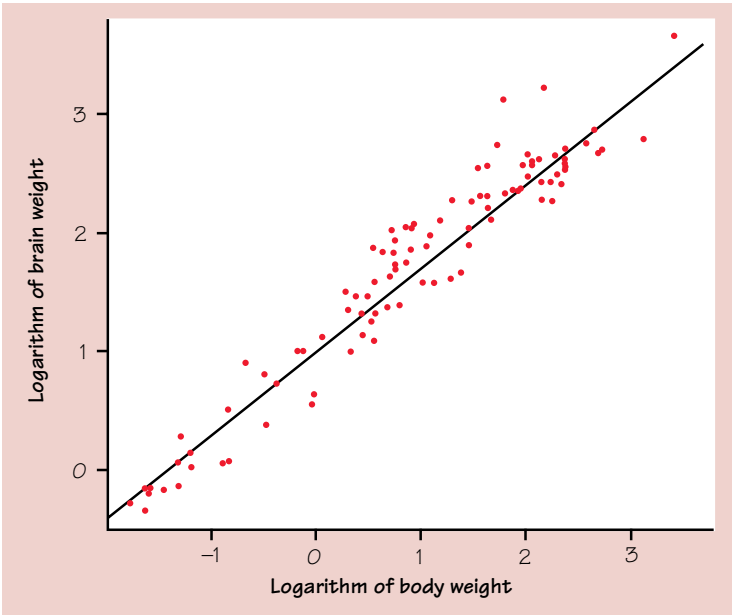


FIGURE 4.3 Scatterplot of the logarithm of brain weight against the logarithm of body weight for 96 species of mammals.

Example 4.1 shows that working with a *function* of our original measurements can greatly simplify statistical analysis. Applying a function such as the logarithm or square root to a quantitative transforming variable is called ***transforming*** or ***reexpressing*** the data. We will see in this section that understanding how simple functions work helps us choose and use transformations. Because we may want to transform either the explanatory variable x or the response variable y in a scatterplot, or both, we will call the variable t when talking about transforming in general.

transforming
reexpressing

First steps in transforming

Transforming data amounts to changing the scale of measurement that was used when the data were collected. We can choose to measure temperature in degrees Fahrenheit or in degrees Celsius, distance in miles or in kilometers. These changes of units are *linear transformations*, discussed on pages 53 to 55. **Linear transformations cannot straighten a curved relationship between two variables.** To do that, we resort to functions that are not linear. The logarithm, applied in Example 4.1, is a nonlinear function. Here are some others.

- How shall we measure the size of a sphere or of such roughly spherical objects as grains of sand or bubbles in a liquid? The size of a sphere can be expressed in terms of the diameter t, in terms of surface area (proportional to t^2), or in terms of volume (proportional to t^3). Any one of these *powers* of the diameter may be natural in a particular application.

- We commonly measure the fuel consumption of a car in miles per gallon, which is how many miles the car travels on 1 gallon of fuel. Engineers prefer to measure in gallons per mile, which is how many gallons of fuel the car needs to travel 1 mile. This is a *reciprocal* transformation. A car that gets 25 miles per gallon uses

$$\frac{1}{\text{miles per gallon}} = \frac{1}{25} = 0.04 \text{ gallons per mile}$$

The reciprocal is a *negative power* $1/t = t^{-1}$.

The transformations we have mentioned—linear, positive and negative powers, and logarithms—are those used in most statistical problems. They are all *monotonic*.

MONOTONIC FUNCTIONS

A **monotonic function** $f(t)$ moves in one direction as its argument t increases.

A **monotonic increasing function** preserves the order of data. That is, if $a > b$, then $f(a) > f(b)$.

A **monotonic decreasing function** reverses the order of data. That is, if $a > b$, then $f(a) < f(b)$.

The graph of a linear function is a straight line. The graph of a monotonic increasing function is increasing everywhere. A monotonic decreasing function has a graph that is decreasing everywhere. A function can be monotonic over some range of t without being everywhere monotonic. For example, the square function t^2 is monotonic increasing for $t \geq 0$. If the range of t includes both positive and negative values, the square is not monotonic—it decreases as t increases for negative values of t and increases as t increases for positive values.

Figure 4.4 compares three monotonic increasing functions and three monotonic decreasing functions for positive values of the argument t. Many variables take only 0 or positive values, so we are particularly interested in how functions behave for positive values of t. The increasing functions for $t > 0$ are

Linear	$a + bt$, slope $b > 0$
Square	t^2
Logarithm	$\log t$

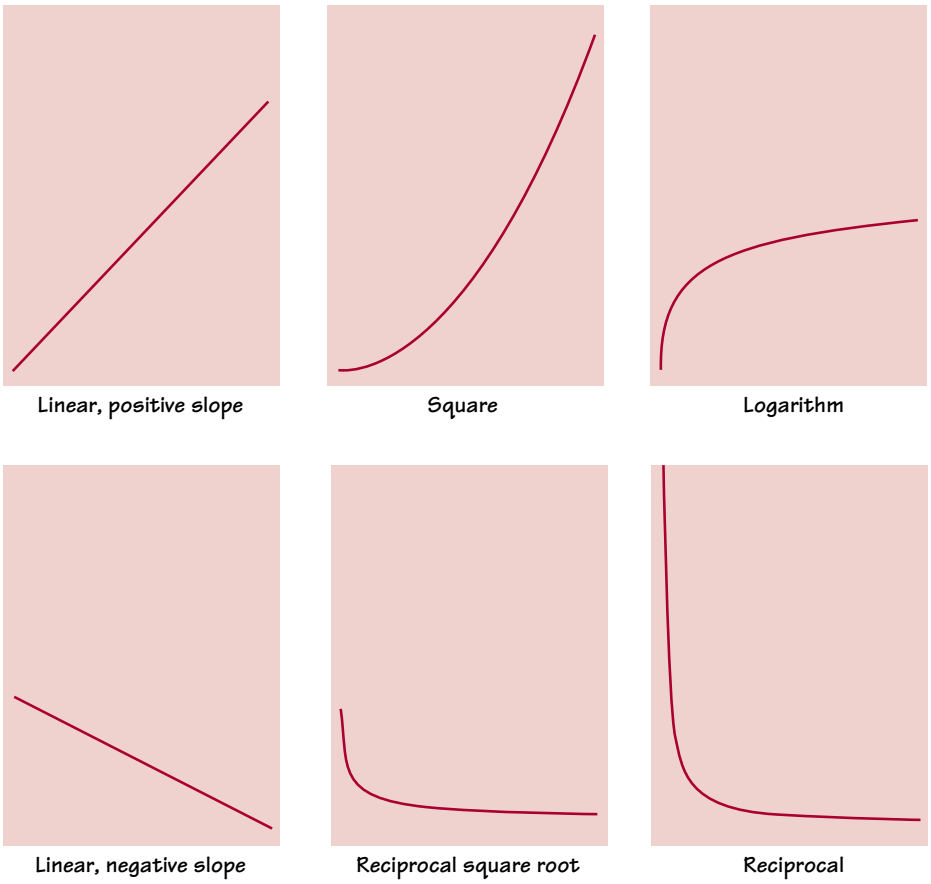


FIGURE 4.4 Three monotonic increasing functions and three monotonic decreasing functions.

The decreasing functions for $t > 0$ in the lower panel of Figure 4.4 are

Linear	$a + bt$, slope $b < 0$
Reciprocal square root	$1/\sqrt{t}$, or $t^{-1/2}$
Reciprocal	$1/t$, or t^{-1}

Nonlinear monotonic transformations change data enough to alter the shape of distributions and the form of relations between two variables, yet are simple enough to preserve order and allow recovery of the original data. We will concentrate on powers and logarithms. The even-numbered powers t^2, t^4, and so on are monotonic increasing for $t \geq 0$, but not when t can take both negative and positive values. The logarithm is not even defined unless $t > 0$. Our strategy for transforming data is therefore as follows:

1. If the variable to be transformed takes values that are 0 or negative, first apply a linear transformation to make the values all positive. Often we just add a constant to all the observations.

2. Then choose a power or logarithmic transformation that simplifies the data, for example, one that approximately straightens a scatterplot.

EXERCISES

4.1 Which of these transformations are monotonic increasing? Monotonic decreasing? Not monotonic? Give an equation for each transformation.

(a) You transform height in inches to height in centimeters.

(b) You transform typing speed in words per minute into seconds needed to type a word.

(c) You transform the diameter of a coin to its circumference.

(d) A composer insists that her new piece of music should take exactly 5 minutes to play. You time several performances, then transform the time in minutes into squared error, the square of the difference between 5 minutes and the actual time.

4.2 Suppose that t is an angle, measured in degrees between 0° and 180°. On what part of this range is the function sin t monotonic increasing? Monotonic decreasing?

The ladder of power transformations

Though simple in algebraic form and easy to compute with a calculator, the power and logarithm functions are varied in their behavior. It is natural to think of powers such as

$$\ldots, t^{-1}, t^{-1/2}, t^{1/2}, t, t^2, \ldots$$

as a hierarchy or ladder. Some facts about this ladder will help us choose transformations. In all cases, we look only at positive values of the argument t.

MONOTONICITY OF POWER FUNCTIONS

Power functions t^p for positive powers p are monotonic increasing for values $t > 0$. They preserve the order of observations. This is also true of the logarithm.

Power functions t^p for negative powers p are monotonic decreasing for values $t > 0$. They reverse the order of the observations.

It is hard to interpret graphs when the order of the original observations has been reversed. We can make a negative power such as the reciprocal $1/t$ monotonic increasing rather than monotonic decreasing by using $-1/t$ instead. Figure 4.5 takes this idea a step farther. This graph compares the ladder of power functions in the form

$$\frac{t^p - 1}{p}$$

The reciprocal (power $p = -1$), for example, is graphed as

$$\frac{1/x - 1}{-1} = 1 - \frac{1}{x}$$

This linear transformation does not change the nature of the power functions t^p, except that all are now monotonic increasing. It is chosen so that every power has the value 0 at $t = 1$ and also has slope 1 at that point. So the graphs in Figure 4.5 all touch at $t = 1$ and go through that point at the same slope.

Look at the $p = 0$ graph in Figure 4.5. The 0th power t^0 is just the constant 1, which is not very useful. The $p = 0$ entry in the figure is not constant. In fact, it is the logarithm, log t. That is, **the logarithm fits into the ladder of power transformations at $p = 0$.**[3]

Figure 4.5 displays another key fact about these functions. The graph of a linear function (power $p = 1$) is a straight line. Powers greater than 1 give graphs that bend upward. That is, the transformed variable grows ever faster as t gets larger. Powers less than 1 give graphs that bend downward. The transformed values continue to grow with t, but at a rate that decreases as t increases. What is more, the sharpness of the bend increases as we move away from $p = 1$ in either direction.

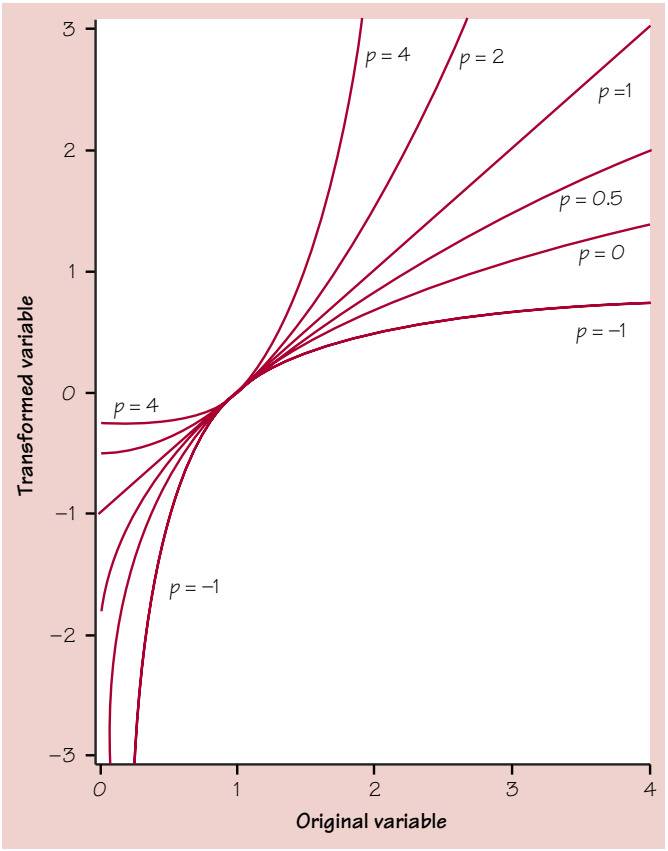


FIGURE 4.5 The ladder of power functions in the form $(t^p - 1)/p$.

CONCAVITY OF POWER FUNCTIONS

Power transformations t^p for powers p greater than 1 are **concave up**; that is, they have the shape ⌣. These transformations push out the right tail of a distribution and pull in the left tail. This effect gets stronger as the power p moves up away from 1.

Power transformations t^p for powers p less than 1 (and the logarithm for $p = 0$) are **concave down**; that is, they have the shape ⌢. These transformations pull in the right tail of a distribution and push out the left tail. This effect gets stronger as the power p moves down away from 1.

EXAMPLE 4.2 A COUNTRY'S GDP AND LIFE EXPECTANCY

Figure 4.6(a) is a scatterplot of data from the World Bank.[4] The individuals are all the world's nations for which data are available. The explanatory variable x is a measure of

how rich a country is: the gross domestic product (GDP) per person. GDP is the total value of the goods and services produced in a country, converted into dollars. The response variable y is life expectancy at birth.

Life expectancy increases in richer nations, but only up to a point. The pattern in Figure 4.6(a) at first rises rapidly as GDP increases but then levels out. Three African nations (Botswana, Gabon, and Namibia) are outliers with much lower life expectancy than the overall pattern suggests. Can we straighten the overall pattern by transforming?

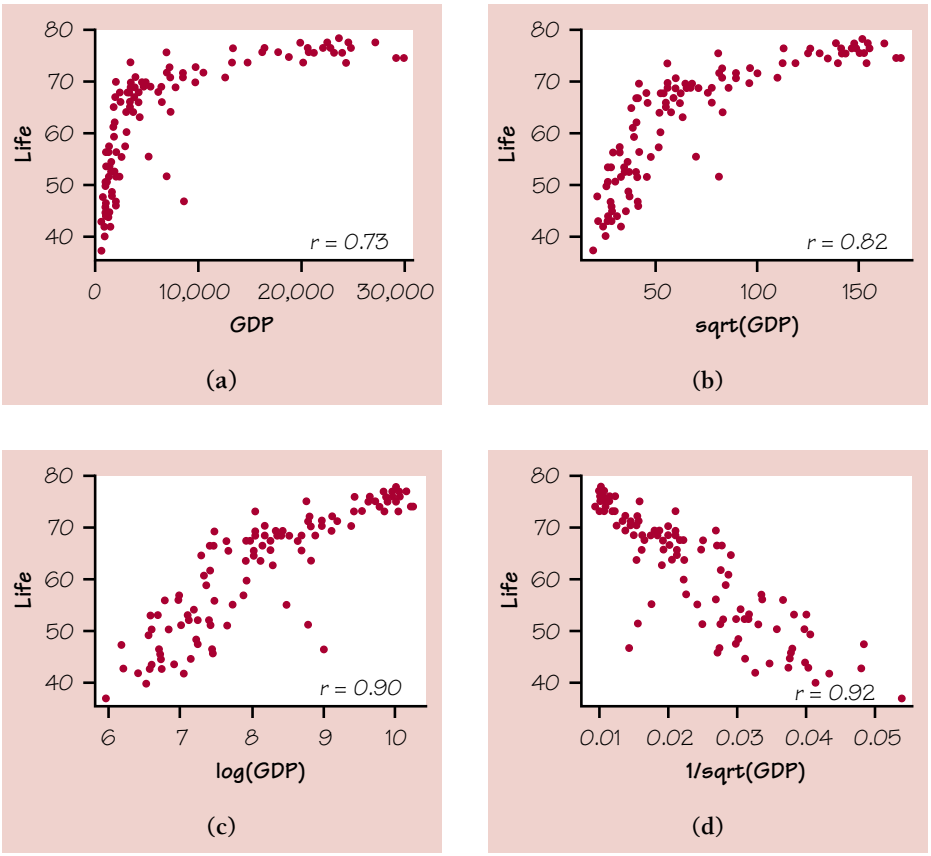


FIGURE 4.6 The ladder of transformations at work. The data are life expectancy and gross domestic product (GDP) for 115 nations. Panel (a) displays the original data. Panels (b), (c), and (d) transform GDP, moving down the ladder away from linear functions.

Life expectancy does not have a large range, but we can see that the distribution of GDP is right-skewed and very spread out. So GDP is a better candidate for transformation. We want to pull in the long right tail, so we try transformations with $p < 1$. Figures 4.6(b), (c), and (d) show the results of three transformations of GDP. The r-value in each figure is the correlation when the three outliers are omitted.

The square root $\sqrt{x}$, with $p = 1/2$, reduces the curvature of the scatterplot, but not enough. The logarithm $\log x$ ($p = 0$) straightens the pattern more, but it still bends to the right. The reciprocal square root $1/\sqrt{x}$, with $p = -1/2$, gives a pattern that is quite straight except for the outliers. To avoid reversing the order of the observations, we actually used $-1/\sqrt{x}$.

EXERCISES

4.3 MUSCLE STRENGTH AND WEIGHT, I Bigger people are generally stronger than smaller people, though there's a lot of individual variation. Let's find a theoretical model. Body weight increases as the cube of height. The strength of a muscle increases with its cross-sectional area, which we expect to go up as the square of height. Put these together: What power law should describe how muscle strength increases with weight?

4.4 MUSCLE STRENGTH AND WEIGHT, II Let's apply your result from the previous problem. Graph the power law relation between strength and body weight for weights from (say) 1 to 1000. (Constants in the power law just reflect the units of measurement used, so we can ignore them.) Use the graph to explain why a person 1 million times as heavy as an ant can't lift a million times as much as an ant can lift.

4.5 HEART RATE AND BODY RATE Physiologists say that resting heart rate of humans is related to our body weight by a power law. Specifically, average heart rate y (beats per minute) is found from body weight x (kilograms) by[5]

$$y = 241 \times x^{-1/4}$$

Let's try to make sense of this. Kleiber's law says that energy use in animals, including humans, increases as the 3/4 power of body weight. But the weight of human hearts and lungs and the volume of blood in the body are directly proportional to body weight. Given these facts, you should not be surprised that heart rate is proportional to the –1/4 power of body weight. Why not?

Example 4.2 shows the ladder of powers at work. As we move down the ladder from linear transformations (power $p = 1$), the scatterplot gets straighter. Moving farther down the ladder, to the reciprocal $1/x = x^{-1}$, begins to bend the plot in the other direction. But this "try it and see" approach isn't very satisfactory. That life expectancy depends linearly on $1/\sqrt{\text{GDP}}$ does not increase our understanding of the relationship between the health and wealth of nations. We don't recommend just pushing buttons on your calculator to try to straighten a scatterplot.

It is much more satisfactory to begin with a theory or mathematical model that we expect to describe a relationship. The transformation needed to make the relationship linear is then a consequence of the model. One of the most common models is *exponential growth*.

Exponential growth

A variable grows linearly over time if it *adds* a fixed increment in each equal time period. Exponential growth occurs when a variable is *multiplied* by a fixed number in each time period. To grasp the effect of multiplicative growth, consider a population of bacteria in which each bacterium splits into two each hour. Beginning with a single bacterium, we have 2 after one hour, 4 at the end of two hours, 8 after three hours, then 16, 32, 64, 128, and so on. These first few numbers are deceiving. After 1 day of doubling each hour, there are 2^{24} (16,777,216) bacteria in the population. That number then doubles the next hour! Try successive multiplications by 2 on your calculator to see for yourself the very rapid increase after a slow start. Figure 4.7 shows the growth of the bacteria population over 24

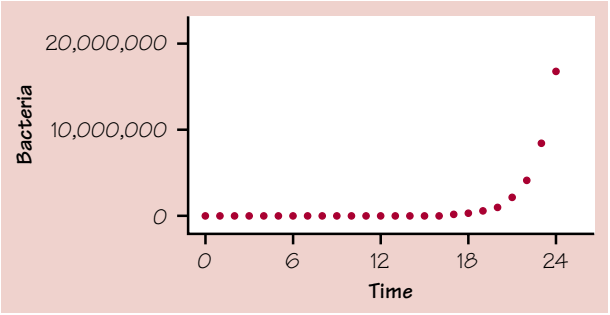


FIGURE 4.7 Growth of a bacteria population over a 24-hour period.

hours. For the first 15 hours, the population is too small to rise visibly above the zero level on the graph. It is characteristic of exponential growth that the increase appears slow for a long period, then seems to explode.

LINEAR VERSUS EXPONENTIAL GROWTH

Linear growth increases by a fixed *amount* in each equal time period.
Exponential growth increases by a fixed *percentage* of the previous total.

Populations of living things—like bacteria and the malignant cancer cells in Activity 4—tend to grow exponentially if not restrained by outside limits such as lack of food or space. More pleasantly, money also displays exponential growth when returns to an investment are compounded. Compounding means that last period's income earns income this period.

EXAMPLE 4.3 THE GROWTH OF MONEY

A dollar invested at an annual rate of 6% turns into \$1.06 in a year. The original dollar remains and has earned \$0.06 in interest. That is, 6% annual interest means that any amount on deposit for the entire year is multiplied by 1.06. If the \$1.06 remains invested for a second year, the new amount is therefore 1.06×1.06, or 1.06^2. That is only \$1.12, but this in turn is multiplied by 1.06 during the third year, and so on. After x years, the dollar has become 1.06^x dollars.

If the Native Americans who sold Manhattan Island for \$24 in 1626 had deposited the \$24 in a savings account at 6% annual interest, they would now have almost \$80 billion. Our savings accounts don't make us billionaires, because we don't stay around long enough. A century of growth at 6% per year turns \$24 into \$8143. That's 1.06^{100} times \$24. By 1826, two centuries after the sale, the account would hold a bit over \$2.7 million. Only after a patient 302 years do we finally reach \$1 billion. That's real money, but 302 years is a long time.

exponential growth model

The count of bacteria after x hours is 2^x. The value of \$24 invested for x years at 6% interest is 24×1.06^x. Both are examples of the ***exponential growth model*** $y = a \times b^x$ for different constants a and b. In this model, the response y is multiplied by b in each time period.

EXAMPLE 4.4 GROWTH OF CELL PHONE USE

Does the exponential growth model sometimes describe real data that don't arise from any obvious process of multiplying by a fixed number over and over again? Let's look at the cell phone phenomenon in the United States. Cell phones have revolutionized the communications industry, the way we do business, and the way we stay in touch with friends and family. The industry enjoyed substantial growth in the 1990s. One way to measure cell phone growth in the 1990s is to look at the number of subscribers. Table 4.1 and Figure 4.8 show the growth of cell phone subscribers from 1990 to 1999.

TABLE 4.1 The number of cell phone subscribers in the United States, 1990–1999

Year	1990	1993	1994	1995	1996	1997	1998	1999
Subscribers (thousands)	5283	16,009	24,134	33,786	44,043	55,312	69,209	86,047

Source: Statistical Abstract of the United States, 2000 and the Cellular Telecommunications Industry Association, Washington, D.C.

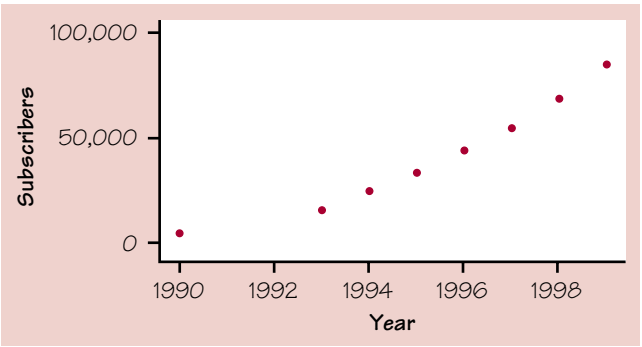


FIGURE 4.8 Scatterplot of cell phone growth versus year, 1990–1999.

There is an increasing trend, but the overall pattern is not linear. The number of cell phone subscribers has increased much faster than linear growth. The pattern of growth follows a smooth curve, and it looks a lot like an exponential curve. Is this exponential growth?

The logarithm transformation

The growth curve for the number of cell phone subscribers does look somewhat like the exponential curve in Figure 4.7, but our eyes are not very good at comparing curves of roughly similar shape. We need a better way to check whether growth is exponential. If you suspect exponential growth, you should first calculate ratios of consecutive terms. In Table 4.2, we have divided each entry in the "Subscribers" column (the y variable) by its predecessor, leaving out both the first value of y, because it doesn't have a predecessor, and the second value, because the x increment is not 1. Notice that the ratios are not *exactly* the same, but they are *approximately* the same.

TABLE 4.2 Ratios of consecutive y-values and the logarithms of the y-values for the cell phone data of Example 4.4

Year	Subscribers	Ratios	log(y)
1990	5,283	—	3.72288
1993	16,009	—	4.20436
1994	24,134	1.51	4.38263
1995	33,786	1.40	4.52874
1996	44,043	1.30	4.64388
1997	55,312	1.26	4.74282
1998	69,209	1.25	4.84016
1999	86,047	1.24	4.93474

The next step is to apply a mathematical transformation that changes exponential growth into linear growth—and patterns of growth that are not exponential into something other than linear. But before we do the transformation, we need to review the properties of logarithms. The basic idea of a logarithm is this: $\log_2 8 = 3$ because 3 is the exponent to which the base 2 must be raised to yield 8. Here is a quick summary of algebraic properties of logarithms:

ALGEBRAIC PROPERTIES OF LOGARITHMS

$$\log_b x = y \quad \text{if and only if} \quad b^y = x$$

The rules for logarithms are

1. $\log(AB) = \log A + \log B$
2. $\log(A/B) = \log A - \log B$
3. $\log X^p = p \log X$

EXAMPLE 4.5 TRANSFORMING CELL PHONE GROWTH

Returning to the cell phone growth model, we hypothesize an exponential model of the form $y = ab^x$ where a and b represent constants. The necessary transformation is carried out by taking the logarithm of both sides of this equation:

$$\begin{aligned} \log y &= \log(ab^x) \\ &= \log a + \log b^x && \text{using Rule 1} \\ &= \log a + (\log b)x && \text{using Rule 3} \end{aligned}$$

Notice that log a and log b are constants because a and b are constants. So the right side of the equation looks like the form for a straight line. That is, if our data really are growing exponentially and we plot log y versus x, we should observe a straight line for the transformed data. Table 4.2 includes the logarithms of the y-values. Figure 4.9 plots points in the form (x, log y).

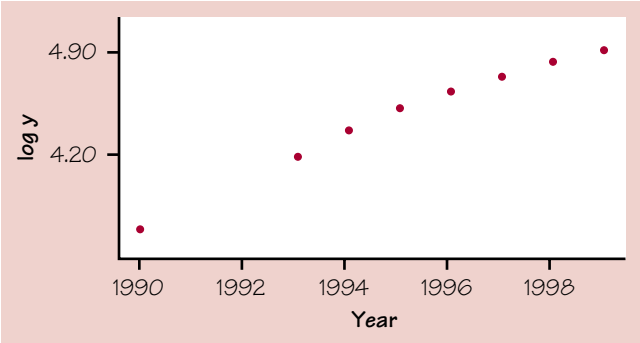


FIGURE 4.9 Scatterplot of log(subscribers) versus year.

The plot appears to be slightly concave down, but it is more linear than our original scatterplot. Applying least-squares regression to the transformed data, Minitab reports:

```
LOG(Y) = - 263 + 0.134 YEAR

Predictor       Coef       Stdev   t-ratio        p
Constant     -263.20       14.63    -17.99    0.000
YEAR        0.134170    0.007331     18.30    0.000

s = 0.05655   R-sq = 98.2%    R-sq(adj) = 97.9%
```

As is usually the case, Minitab tells us more than we want to know, but observe that the value of r^2 is 0.982. That means that 98.2% of the variation in log y is explained by least-squares regression of log y on x. That's pretty impressive. Let's continue. Figure 4.10 is a plot of the transformed data along with the fitted line.

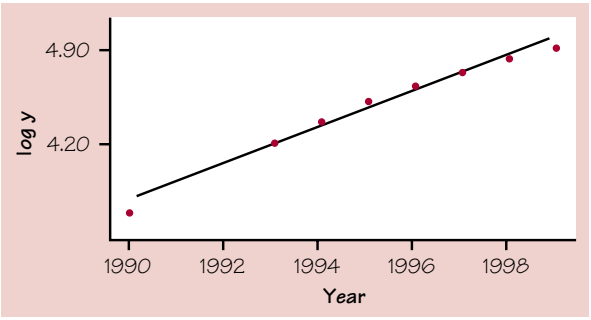


FIGURE 4.10 Plot of transformed data with least-squares line.

This appears to be a useful model for prediction purposes. Although the r^2-value is high, one should always inspect the residual plot to further assess the quality of the model. Figure 4.11 is a residual plot.

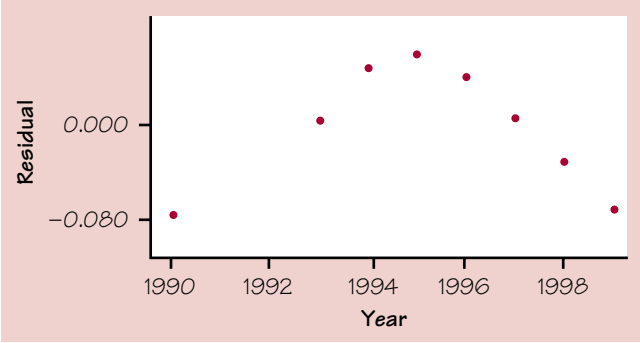


FIGURE 4.11 Residual plot for transformed cell phone growth data.

This is a surprise. But it also suggests an adjustment. The very regular pattern of the last four points really does look linear. So if the purpose is to be able to predict the number of subscribers in the year 2000, then one approach would be to discard the first four points, because they are the oldest and furthest removed from the year 2000, and retain the last four points. If you do this, the least-squares line for the four transformed points (years 1996 through 1999) is

```
log NewY = -189 + 0.0970 NewX
```

and the r^2-value improves to 1. The actual r^2-value is 0.999897 to six decimal places. The residual plot is shown in Figure 4.12.

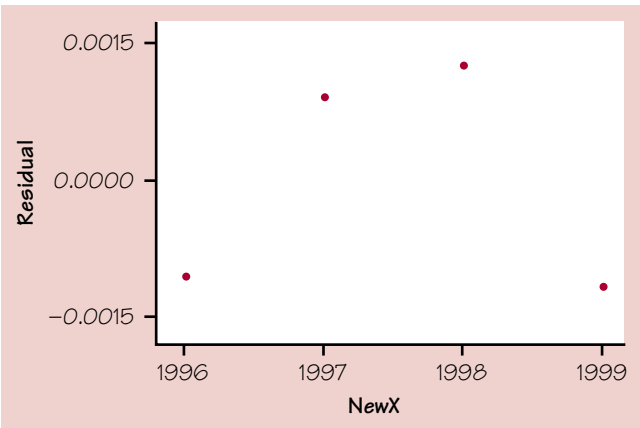


FIGURE 4.12 Residual plot for reduced transformed data set.

Although there is still a slight pattern in the residual plot, the residuals are very small in magnitude, and the r^2 value is nearly 1.

Prediction in the exponential growth model

Regression is often used for prediction. When we fit a least-squares regression line, we find the predicted response y for any value of the explanatory variable x by substituting our x-value into the equation of the line. In the case of exponential growth, the logarithms rather than the actual responses follow a linear pattern. To do prediction, we need to "undo" the logarithm transformation to return to the original units of measurement. The same idea works for any monotonic transformation. There is always exactly one original value behind any transformed value, so we can always go back to our original scale.

EXAMPLE 4.6 PREDICTING CELL PHONE GROWTH FOR 2000

Our examination of cell phone growth left us with four transformed data points and a least-squares line with equation

```
log(subscribers) = -189 + 0.0970(year)
```

To perform the back-transformation, we need to do the inverse operation. The inverse operation of the logarithmic function is raising 10 to a power. If we raise 10 to the left side of the equation, and set that equal to 10 raised to the right side of the equation, we will eliminate the log() on the left;

$$10^{\log(\text{subscribers})} = 10^{-189 + 0.0970(\text{year})}$$

Then

$$\text{subscribers} = (10^{-189})(10^{0.0970(\text{year})})$$

To then predict the number of subscribers in the year 2000, we substitute 2000 for year and solve for number of subscribers. The problem is that the first factor is too small a quantity for the calculator, and it will evaluate to 0. To get around this machine difficulty, if you have installed the equation of the least-squares line in the calculator as Y1, then define Y2 to be 10^Y1. Doing this, we find that the predicted number of subscribers for the year 2000 is Y2(2000) = 10,7864.5. Alternatively, we could have coded the years to avoid the overflow problem.

Postscript: The stock market tumbled in 2000, the economy floundered, unemployment increased, and the cell phone industry in particular had a very poor year. So predicting the number of cell phone subscribers in 2000 is risky indeed.

Make sure that you understand the big idea here. The necessary transformation is carried out by taking the logarithm of the response variable. Your calculator and most statistical software will calculate the logarithms of all the values of a variable with a single command. The essential property of the logarithm for our purposes is that it straightens an exponential growth curve. **If a variable grows exponentially, its logarithm grows linearly.**

EXAMPLE 4.7 TRANSFORMING BACTERIA COUNTS

Figure 4.13 plots the logarithms of the bacteria counts in Figure 4.7 (page 204). Sure enough, exact exponential growth turns into an exact straight line when we plot the logarithms. After 15 hours, for example, the population contains $2^{15} = 32{,}768$ bacteria. The logarithm of 32,768 is 4.515, and this point appears above the 15-hour mark in Figure 4.13.

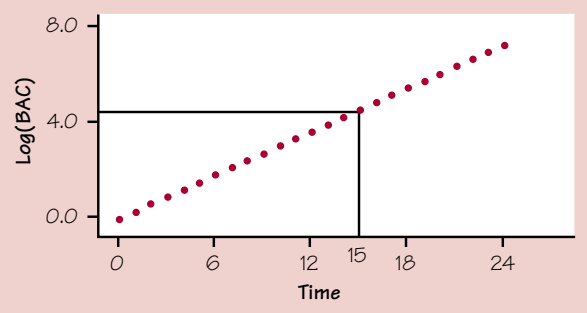


FIGURE 4.13 Logarithms of the bacteria counts.

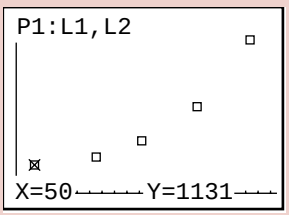
TECHNOLOGY TOOLBOX *Modeling exponential growth with the TI-83/89*

The Census Bureau classifies residents of the United States as being either white; black; Hispanic origin; American Indian, Eskimo, Aleut; or Asian, Pacific Islander. The population totals for these last two categories, from 1950 to 1990, are[6]

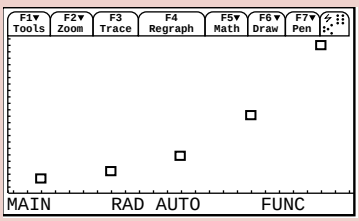
Year:	1950	1960	1970	1980	1990
Population (thousands):	1131	1620	2557	5150	9534

- Code the years using 1900 as the reference year, 0. Then 1950 is coded as 50, and so forth. Enter the coded years and population, in thousands, in L_1/list1 and L_2/list2. Then plot the scatterplot.

TI-83

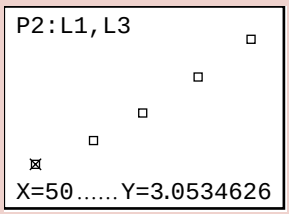


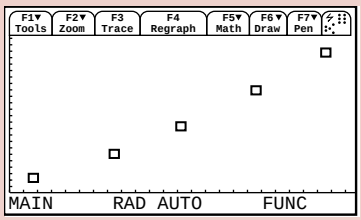
TI-89



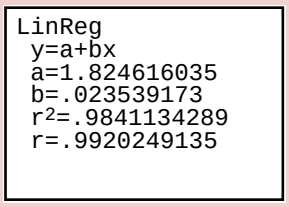
- Assuming an exponential model, here is a plot of log(POP), in L_3, versus YEAR on the TI-83. We'll plot ln(POP) versus YEAR on the TI-89 since the natural logarithm key is more accessible on the TI-89. The pattern is the same, but the regression equation numbers will be different.

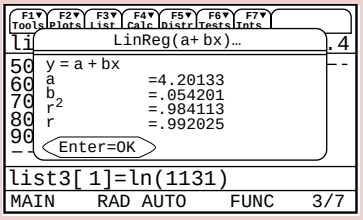
TECHNOLOGY TOOLBOX *Modeling exponential growth with the TI-83/89 (continued)*



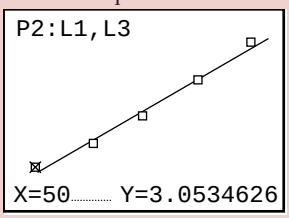


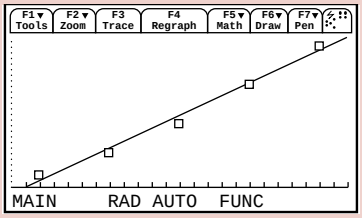
- The plot still shows a little upward concavity, and the residual plot will confirm this. Next, we perform least-squares regression on the transformed data.



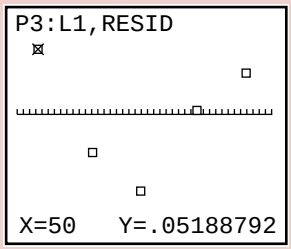


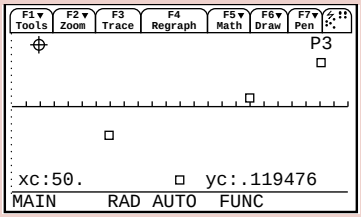
- Notice that the values of a and b in the equation of the least-squares line are different for the two calculators. That's because we use base 10 (log) on the TI-83 and we used base e (ln) on the TI-89. The final predicted values will be the same regardless of which route we take. Here are the scatterplots with the least-squares lines:





- Despite the high r^2-value, you should always inspect the residual plot. Here it is:





Ideally, the residual plot should show random scatter about the y = 0 reference line. The fact that the residual plot still shows a clearly curved pattern tells us that some improvement is still possible. For now, though, we will accept the exponential model on the basis of the high r^2-value ($r^2 = 0.992$).

- Now we're ready to predict the population of American Indians, Eskimos, Aleuts, Asians, and Pacific Islanders for the year 2000. With the regression equation installed as Y1, define Y2 = 10^Y1 on the TI-83, and Y2 = e^Y1 on the TI-89. The predicted population in year 2000 is then Y2(100) = 15,084.584 on the TI-83, and 15,084.7 on the TI-89. The difference is due to roundoff error. Since the table entries are in thousands, the actual predicted population is approximately 15,085,000. Looking at the plots, do you think this prediction will be too high or too low? Why?

EXERCISES

4.6 GYPSY MOTHS Biological populations can grow exponentially if not restrained by predators or lack of food. The gypsy moth outbreaks that occasionally devastate the forests of the Northeast illustrate approximate exponential growth. It is easier to count the number of acres defoliated by the moths than to count the moths themselves. Here are data on an outbreak in Massachusetts:[7]

Year	Acres
1978	63,042
1979	226,260
1980	907,075
1981	2,826,095

(a) Plot the number of acres defoliated y against the year x. The pattern of growth appears exponential.

(b) Verify that y is being multiplied by about 4 each year by calculating the ratio of acres defoliated each year to the previous year. (Start with 1979 to 1978, when the ratio is 226,260/63,042 = 3.6.)

(c) Take the logarithm of each number y and plot the logarithms against the year x. The linear pattern confirms that the growth is exponential.

(d) Verify that the least-squares line fitted to the transformed data is

$$\log \hat{y} = -1094.51 + 0.5558 \times \text{year}$$

(e) Construct and interpret a residual plot for $\log \hat{y}$ on year.

(f) Perform the inverse transformation to express $\hat{y}$ as an exponential equation. Display a scatterplot of the original data with the exponential curve model superimposed. Is your exponential function a satisfactory model for the data?

(g) Use your model to predict the number of acres defoliated in 1982.

(*Postscript:* A viral disease reduced the gypsy moth population between the readings in 1981 and 1982. The actual count of defoliated acres in 1982 was 1,383,265.)

4.7 MOORE'S LAW, I Gordon Moore, one of the founders of Intel Corporation, predicted in 1965 that the number of transistors on an integrated circuit chip would double every 18 months. This is "Moore's law," one way to measure the revolution in computing. Here are data on the dates and number of transistors for Intel microprocessors:[8]

Processor	Date	Transistors	Processor	Date	Transistors
4004	1971	2,250	486 DX	1989	1,180,000
8008	1972	2,500	Pentium	1993	3,100,000
8080	1974	5,000	Pentium II	1997	7,500,000
8086	1978	29,000	Pentium III	1999	24,000,000
286	1982	120,000	Pentium 4	2000	42,000,000
386	1985	275,000			

(a) Explain why Moore's law says that the number of transistors grows exponentially over time.

(b) Make a plot suitable to check for exponential growth. Does it appear that the number of transistors on a chip has in fact grown approximately exponentially?

4.8 MOORE'S LAW, II Return to Moore's law, described in Exercise 4.7.

(a) Find the least-squares regression line for predicting the logarithm of the number of transistors on a chip from the date. Before calculating your line, subtract 1970 from all the dates so that 1971 becomes year 1, 1972 is year 2, and so on.

(b) Suppose that Moore's law is exactly correct. That is, the number of transistors is 2250 in year 1 (1971) and doubles every 18 months (1.5 years) thereafter. Write the model for predicting transistors in year x after 1970. What is the equation of the line that, according to your model, connects the logarithm of transistors with x? Explain why a comparison of this line with your regression line from **(a)** shows that although transistor counts have grown exponentially, they have grown a bit more slowly than Moore's law predicts.

4.9 *E. COLI* (Exact exponential growth) The common intestinal bacterium *E. coli* is one of the fastest-growing bacteria. Under ideal conditions, the number of *E. coli* in a colony doubles about every 15 minutes until restrained by lack of resources. Starting from a single bacterium, how many *E. coli* will there be in 1 hour? In 5 hours?

4.10 GUN VIOLENCE (Exact exponential growth) A paper in a scholarly journal once claimed (I am not making this up), "Every year since 1950, the number of American children gunned down has doubled."[9] To see that this is silly, suppose that in 1950 just 1 child was "gunned down" and suppose that the paper's claim is exactly right.

(a) Make a table of the number of children killed in each of the next 10 years, 1951 to 1960.

(b) Plot the number of deaths against the year and connect the points with a smooth curve. This is an exponential curve.

(c) The paper appeared in 1995, 45 years after 1950. How many children were killed in 1995, according to the paper?

(d) Take the logarithm of each of your counts from **(a)**. Plot these logarithms against the year. You should get a straight line.

(e) From your graph in **(d)** find the approximate values of the slope b and the intercept a for the line. Use the equation $y = a + bx$ to predict the logarithm of the count for the 45th year. Check your result by taking the logarithm of the count you found in **(c)**.

4.11 U.S. POPULATION The following table gives the resident population of the United States from 1790 to 2000, in millions of persons:

Date	Pop.	Date	Pop.	Date	Pop.	Date	Pop.
1790	3.9	1850	23.2	1910	92.0	1970	203.3
1800	5.3	1860	31.4	1920	105.7	1980	226.5
1810	7.2	1870	39.8	1930	122.8	1990	248.7
1820	9.6	1880	50.2	1940	131.7	2000	281.4
1830	12.9	1890	62.9	1950	151.3		
1840	17.1	1900	76.0	1960	179.3		

(a) Plot population against time. The growth of the American population appears roughly exponential.

(b) Plot the logarithms of population against time. The pattern of growth is now clear. An expert says that "the population of the United States increased exponentially from 1790 to about 1880. After 1880 growth was still approximately exponential, but at a slower rate." Explain how this description is obtained from the graph.

(c) Use part or all the data to construct an exponential model for the purpose of predicting the population in 2010. Justify your modeling decision. Then predict the population in the year 2010. Do you think your prediction will be too low or too high? Explain.

(d) Construct a residual plot for the transformed data. What is the value of r^2 for the transformed data?

(e) Comment on the quality of your model.

Power law models

When you visit a pizza parlor, you order a pizza by its diameter, say 10 inches, 12 inches, or 14 inches. But the amount you get to eat depends on the *area* of the pizza. The area of a circle is π times the square of its radius. So the area of a round pizza with diameter x is

$$\text{area} = \pi r^2 = \pi(x/2)^2 = \pi(x^2/4) = (\pi/4)x^2$$

power law model

This is a ***power law model*** of the form

$$y = a \times x^p$$

When we are dealing with things of the same general form, whether circles or fish or people, we expect area to go up with the square of a dimension such as diameter or height. Volume should go up with the cube of a linear dimension. That is, geometry tells us to expect power laws in some settings.

Biologists have found that many characteristics of living things are described quite closely by power laws. There are more mice than elephants, and more flies than mice—the abundance of species follows a power law with body weight as the explanatory variable. So do pulse rate, length of life, the number of eggs a bird lays, and so on. Sometimes the powers can be predicted from geometry, but sometimes they are mysterious. Why, for example, does the rate at which animals use energy go up as the 3/4 power of their body weight? Biologists call this relationship *Kleiber's law*. It has been found to work all the way from bacteria to whales. The search goes on for some physical or geometrical explanation for why life follows power laws. There is as yet no general explanation, but power laws are a good place to start in simplifying relationships for living things.

Exponential growth models become linear when we apply the logarithm transformation to the response variable y. **Power law models become linear when we apply the logarithm transformation to both variables**. Here are the details:

1. The power law model is

$$y = a \times x^p$$

2. Take the logarithm of both sides of this equation. You see that

$$\log y = \log a + p \log x$$

That is, taking the logarithm of both variables straightens the scatterplot of y against x.

3. Look carefully: The *power* p in the power law becomes the *slope* of the straight line that links $\log y$ to $\log x$.

Prediction in power law models

If taking the logarithms of both variables makes a scatterplot linear, a power law is a reasonable model for the original data. We can even roughly estimate what power p the law involves by regressing $\log y$ on $\log x$ and using the slope of the regression line as an estimate of the power. Remember that the slope is only an estimate of the p in an underlying power model. The greater the scatter of the points in the scatterplot about the fitted line, the smaller our confidence that this estimate is accurate.

EXAMPLE 4.8 PREDICTING BRAIN WEIGHT

The magical success of the logarithm transformation in Example 4.1 on page 195 would not surprise a biologist. We suspect that a power law governs this relationship. Least-squares regression for the scatterplot in Figure 4.3 on page 196 gives the line

$$\log \hat{y} = 1.01 + 0.72 \times \log x$$

for predicting the logarithm of brain weight from the logarithm of body weight. To undo the logarithm transformation, remember that for common logarithms with base 10, $y = 10^{\log y}$. We see that

$$\begin{aligned} \hat{y} &= 10^{1.01 + 0.72 \log x} \\ &= 10^{1.01} \times 10^{0.72 \log x} \\ &= 10.2 \times (10^{\log x})^{0.72} \end{aligned}$$

Because $10^{\log x} = x$, the estimated power model connecting predicted brain weight $\hat{y}$ with body weight x for mammals is

$$\hat{y} = 10.2 \times x^{0.72}$$

Based on footprints and some other sketchy evidence, some people think that a large apelike animal, called Sasquatch or Bigfoot, lives in the Pacific Northwest. His weight is estimated to be about 280 pounds, or 127 kilograms. How big is Bigfoot's brain? Based on the power law estimated from data on other mammals, we predict

$$\begin{aligned}\hat{y} &= 10.2 \times 127^{0.72} \\ &= 10.2 \times 32.7 \\ &= 333.7 \text{ grams}\end{aligned}$$

For comparison, gorillas have an average body weight of about 140 kilograms and an average brain weight of about 406 grams. Of course, Bigfoot may have a larger brain than his weight predicts—after all, he has avoided being captured, shot, or videotaped for many years.

EXAMPLE 4.9 FISHING TOURNAMENT

Imagine that you have been put in charge of organizing a fishing tournament in which prizes will be given for the heaviest fish caught. You know that many of the fish caught during the tournament will be measured and released. You are also aware that trying to weigh a fish that is flipping around, in a boat that is rolling with the swells, using delicate scales will probably not yield very reliable results.

It would be much easier to measure the *length* of the fish on the boat. What you need is a way to convert the length of the fish to its weight. You reason that since length is one-dimensional and weight is three-dimensional, and since a fish 0 units long would weigh 0 pounds, the weight of a fish should be proportional to the cube of its length. Thus, a model of the form weight = $a \times \text{length}^3$ should work. You contact the nearby marine research laboratory and they provide the average length and weight catch data for the Atlantic Ocean rockfish *Sebastes mentella* (Table 4.3).[10] The lab also advises you that the model relationship between body length and weight has been found to be accurate for most fish species growing under normal feeding conditions.

TABLE 4.3 Average length and weight at different ages for Atlantic Ocean rockfish, *Sebastes mentella*

Age (yr)	Length (cm)	Weight (g)	Age (yr)	Length (cm)	Weight (g)
1	5.2	2	11	28.2	318
2	8.5	8	12	29.6	371
3	11.5	21	13	30.8	455
4	14.3	38	14	32.0	504
5	16.8	69	15	33.0	518
6	19.2	117	16	34.0	537
7	21.3	148	17	34.9	651
8	23.3	190	18	36.4	719
9	25.0	264	19	37.1	726
10	26.7	293	20	37.7	810

Figure 4.14 is a scatterplot of weight in grams versus height in centimeters. Although the growth might appear to be exponential, we know that it is frequently misleading to trust too much to the eye. Moreover, we have already decided on a model that makes sense in this context: weight = $a \times \text{length}^3$.

If we take the $\log_{10}$ of both sides, we obtain

$$\log(\text{weight}) = \log a + [3 \times \log(\text{length})]$$

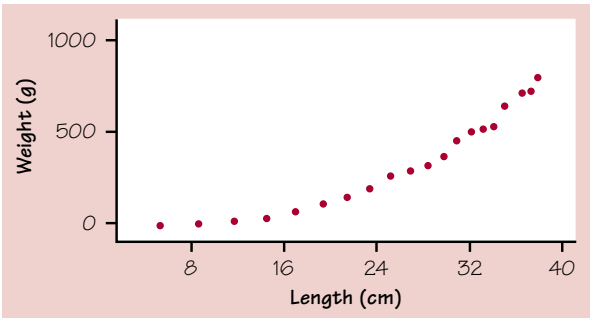


FIGURE 4.14 Scatterplots of Atlantic Ocean rockfish weight versus length.

This equation looks like a linear equation

$$Y = A + BX$$

so we plot log(weight) against log(length). See Figure 4.15.

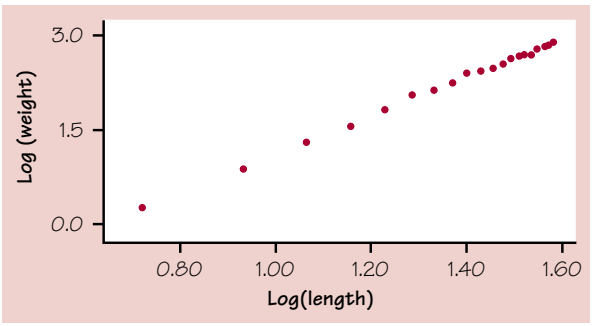


FIGURE 4.15 Scatterplot of log(weight) versus log(length).

We visually confirm that the relationship appears very linear. We perform a least-squares regression on the transformed points [log(length), log(weight)].

The least-squares regression line equation is

$$\log(\text{weight}) = -1.8994 + 3.0494 \log(\text{length})$$

$r = 0.99926$ and $r^2 = 0.9985$. We see that the correlation r of the logarithms of length and weight is virtually 1. (Remember, however, that correlation was defined only for

linear fits.) Despite the very high *r*-value, it's still important to look at a residual plot. The random scatter of the points in Figure 4.16 tells us that the line is a good model for the logs of length and weight.

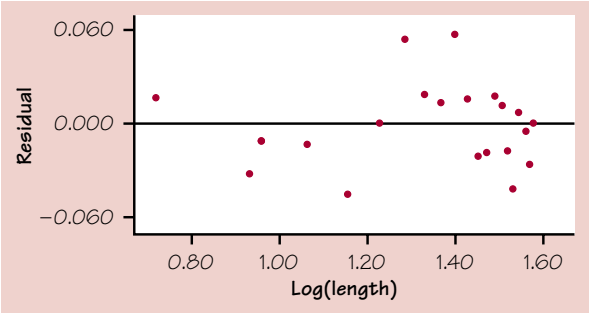


FIGURE 4.16 Plot of residuals versus log(length).

The last step is to perform an inverse transformation on the linear regression equation:

$$\begin{aligned}\log(\text{weight}) &= -1.8994 + [3.0494 \log(\text{length})] \\ &= -1.8994 + \log(\text{length})^{3.0494}\end{aligned}$$

This is the critical step: to remember to use a property of logarithms to write the multiplicative constant 3.0494 as an exponent. Let's continue. Raise 10 to the left side of the equation and set this equal to 10 raised to the right side:

$$\begin{aligned}10^{\log(\text{weight})} &= 10^{-1.8994+\log(\text{length})^{3.0494}} \\ \text{weight} &= 10^{-1.8994} \times \text{length}^{3.0494}\end{aligned}$$

This is the final power equation for the original data.

The scatterplot of the original data along with the power law model appears in Figure 4.17. The fit of this model has visual appeal. We will leave it as an exercise to calculate the sum of the squares of the deviations. It should be noted that the power of *x* that we obtained for the model, 3.0494, is very close to the value 3 that we conjectured when we proposed the form for our model.

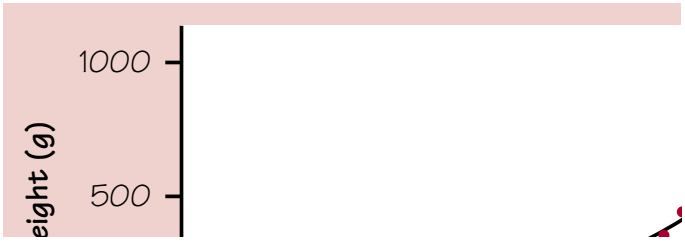


FIGURE 4.17 Atlantic Ocean rockfish data with power law model.

The original purpose for developing this model was to approximate the weight of a fish given its length. Suppose your catch measured 36 centimeters. Our model predicts a weight of Y2(36) = 702.0836281, or about 702 grams. If you entered a fishing contest, would you be comfortable with this procedure for determining the weights of the fish caught, and hence for determining the winner of the contest?

TECHNOLOGY TOOLBOX *Power law modeling*

- Enter the x data (explanatory) into L_1/list1 and the y data (response) into L_2/list2.
- Produce a scatterplot of y versus x. Confirm a nonlinear trend that could be modeled by a power function in the form $y = ax^b$.
- Define L_3/list3 to be log(L_1) or log(list1), and define L_4/list4 to be log(L_2) or log(list2).
- Plot log y versus log x. Verify that the pattern is approximately linear.
- Regress log y on log x. The command line should read LinReg a+bx,L3,L4,Y1. This stores the regression equation as Y1. Remember that Y1 is really log y. Check the r^2-value.
- Construct a residual plot, in the form of either RESID versus x or RESID versus predicted values (fits). Ideally, the points in a residual plot should be randomly scattered above and below the $y = 0$ reference line.
- Perform the back-transformation to find the power function $y = ax^b$ that models the original data. Define Y2 to be (10^a)(x^b). The calculator has stored the values of a and b for the most recent regression performed. Deselect Y1 and plot Y2 and the scatterplot for the original data together.
- To make a prediction for the value $x = k$, evaluate Y2(k) in the Home screen.

EXERCISES

4.12 FISH WEIGHTS

(a) Use the model we derived for approximating the weight of *Sebastes mentella*, $\hat{y} = 10^{-1.8994}x^{3.0494}$, to determine the sum of the squares of the deviations between the observed weights (in grams) and the predicted values. Did we minimize this quantity in the process of constructing our model? If not, what quantity was minimized?

(b) When we performed least-squares regression of log(weight) on log(length) on the calculator, residuals were calculated and stored in a list named RESID. Use this list and the 1-Var Stats command to calculate the sum of the squares of the residuals. Compare this sum of squares with the sum of squares you calculated in **(a)**.

(c) Would you expect the answers in **(a)** and **(b)** to be the same or different? Explain.

4.13 BODY WEIGHT AND LIFETIME Table 4.4 gives the average weight and average life span in captivity for several species of mammals. Some writers on power laws in biology claim that life span depends on body weight according to a power law with power

TABLE 4.4 Body weight and lifetime for several species of mammals

Species	Weight (kg)	Life span (years)	Species	Weight (kg)	Life span (years)
Baboon	32	20	Guinea pig	1	4
Beaver	25	5	Hippopotamus	1400	41
Cat, domestic	2.5	12	Horse	480	20
Chimpanzee	45	20	Lion	180	15
Dog	8.5	12	Mouse, house	0.024	3
Elephant	2800	35	Pig, domestic	190	10
Goat, domestic	30	8	Red fox	6	7
Gorilla	140	20	Sheep, domestic	30	12
Grizzly bear	250	25			

Source: G. A. Sacher and E. F. Staffelt, "Relation of gestation time to brain weight for placental mammals: implications for the theory of vertebrate growth," *American Naturalist*, 108 (1974), pp. 593–613. We found these data in F. L. Ramsey and D. W. Schafer, *The Statistical Sleuth: A Course in Methods of Data Analysis*, Duxbury, 1997.

$p = 0.2$. Fit a power law model to these data (using logarithms). Does this small set of data appear to follow a power law with power close to 0.2? Use your fitted model to predict the average life span for humans (average weight 143 kilograms). Humans are an exception to the rule.

4.14 HEART WEIGHTS OF MAMMALS Use the methods discussed in this section to analyze the following data on the hearts of various mammals.[11] Write your findings and conclusions in a short narrative.

Mammal	Heart weight (grams)	Length of cavity of left ventricle (centimeters)
Mouse	0.13	0.55
Rat	0.64	1.0
Rabbit	5.8	2.2
Dog	102	4.0
Sheep	210	6.5
Ox	2030	12.0
Horse	3900	16.0

4.15 The U.S. Department of Health and Human Services characterizes adults as "seriously overweight" if they meet certain criterion for their height as shown in the table below (only a portion of the chart is reproduced here).

Height (ft, in)	Height (in)	Severely overweight (lb)	Height (ft, in)	Height (in)	Severely overweight (lb)
4′10″	58	138	5′8″	68	190
5′0″	60	148	6′0″	72	213
5′2″	62	158	6′2″	74	225
5′4″	64	169	6′4″	76	238
5′6″	66	179	6′6″	78	250

Weights are given in pounds, without clothes. Height is measured without shoes. There is no distinction between men and women; a note accompanying the table states, "The higher weights apply to people with more muscle and bone, such as many men." Despite any reservations you may have about the department's common standards for both genders, do the following:

(a) Without looking at the data, hypothesize a relationship between height and weight of U.S. adults. That is, write a general form of an equation that you believe will model the relationship.

(b) Which variable would you select as explanatory and which would be the response? Plot the data from the table.

(c) Perform a transformation to linearize the data. Do a least-squares regression on the transformed data and check the correlation coefficient.

(d) Construct a residual plot of the transformed data. Interpret the residual plot.

(e) Perform the inverse transformation and write the equation for your model. Use your model to predict how many pounds a 5′10″ adult would have to weigh in order to be classified by the department as "seriously overweight." Do the same for a 7-foot tall individual.

4.16 THE PRICE OF PIZZAS The new manager of a pizza restaurant wants to add variety to the pizza offerings at the restaurant. She also wants to determine if the prices for existing sizes of pizzas are consistent. Prices for plain (cheese only) pizzas are shown below:

Size	Diameter (inches)	Cost
Small	10	$4.00
Medium	12	$6.00
Large	14	$8.00
Giant	18	$10.00

(a) Construct an appropriate model for these data. Comment on your choice of model.

(b) Based on your analysis, would you advise the manager to adjust the price on any of the pizza sizes? If so, explain briefly.

(c) Use your model to suggest a price for a new "personal pizza," with a 6-inch diameter.

(d) Use your model to suggest a price for a new "soccer team" size, with a 24-inch diameter (assuming the oven is large enough to hold it).

SUMMARY

Nonlinear relationships between two quantitative variables can sometimes be changed into linear relationships by **transforming** one or both of the variables.

The most common transformations belong to the family of **power transformations** t^p. The logarithm $\log t$ fits into the power family at position $p = 0$.

When the variable being transformed takes only positive values, the power transformations are all **monotonic.** This implies that there is an

inverse transformation that returns to the original data from the transformed values. The effect of the power transformations on data becomes stronger as we move away from linear transformations ($p = 1$) in either direction.

Transformation is particularly effective when there is reason to think that the data are governed by some mathematical model. The **exponential growth model** $y = ab^x$ becomes linear when we plot log y against x. The **power law model** $y = ax^p$ becomes linear when we plot log y against log x.

We can fit exponential growth and power models to data by finding the least-squares regression line for the transformed data, then doing the inverse transformation.

SECTION 4.1 EXERCISES

4.17 EXACT EXPONENTIAL GROWTH, I Maria is given a savings bond at birth. The bond is initially worth \$500 and earns interest at 7.5% each year. This means that the value is multiplied by 1.075 each year.

(a) Find the value of the bond at the end of 1 year, 2 years, and so on up to 10 years.

(b) Plot the value y against years x. Connect the points with a smooth curve. This is an exponential curve.

(c) Take the logarithm of each of the values y that you found in **(a)**. Plot the logarithm log y against years x. You should obtain a straight line.

4.18 EXACT EXPONENTIAL GROWTH, II Fred and Alice were born the same year, and each began life with \$500. Fred added \$100 each year, but earned no interest. Alice added nothing, but earned interest at 7.5% annually. After 25 years, Fred and Alice are getting married. Who has more money?

4.19 FISH IN FINLAND, I Here are data for 12 perch caught in a lake in Finland:[12]

Weight (grams)	Length (cm)	Width (cm)	Weight (grams)	Length (cm)	Width (cm)
5.9	8.8	1.4	300.0	28.7	5.1
100.0	19.2	3.3	300.0	30.1	4.6
110.0	22.5	3.6	685.0	39.0	6.9
120.0	23.5	3.5	650.0	41.4	6.0
150.0	24.0	3.6	820.0	42.5	6.6
145.0	25.5	3.8	1000.0	46.6	7.6

(a) Make a scatterplot of weight against length. Describe the pattern you see.

(b) How do you expect the weight of animals of the same species to change as their length increases? Make a transformation of weight that should straighten the plot if

your expectation is correct. Plot the transformed weights against length. Is the plot now roughly linear?

4.20 FISH IN FINLAND, II Plot the widths of the 12 perch in the previous problem against their lengths. What is the pattern of the plot? Explain why we should expect this pattern.

4.21 HOW MOLD GROWS, I Do mold colonies grow exponentially? In an investigation of the growth of molds, biologists inoculated flasks containing a growth medium with equal amounts of spores of the mold *Aspergillus nidulans*. They measured the size of a colony by analyzing how much remains of a radioactive tracer substance that is consumed by the mold as it grows. Each size measurement requires destroying that colony, so that the data below refer to 30 separate colonies. To smooth the pattern, we take the mean size of the three colonies measured at each time.[13]

Hours	Colony sizes			Mean
0	1.25	1.60	0.85	1.23
3	1.18	1.05	1.32	1.18
6	0.80	1.01	1.02	0.94
9	1.28	1.46	2.37	1.70
12	2.12	2.09	2.17	2.13
15	4.18	3.94	3.85	3.99
18	9.95	7.42	9.68	9.02
21	16.36	13.66	12.78	14.27
24	25.01	36.82	39.83	33.89
36	138.34	116.84	111.60	122.26

(a) Graph the mean colony size against time. Then graph the logarithm of the mean colony size against time.

(b) On the basis of data such as these, microbiologists divide the growth of mold colonies into three phases that follow each other in time. Exponential growth occurs during only one of these phases. Briefly describe the three phases, making specific reference to the graphs to support your description.

(c) The exponential growth phase for these data lasts from about 6 hours to about 24 hours. Find the least-squares regression line of the logarithms of mean size on hours for only the data between 6 and 24 hours. Use this line to predict the size of a colony 10 hours after inoculation. (The line predicts the logarithm. You must obtain the size from its logarithm.)

4.22 DETERMINING TREE BIOMASS It is easy to measure the "diameter at breast height" of a tree. It's hard to measure the total "aboveground biomass" of a tree, because to do this you must cut and weigh the tree. The biomass is important for studies of ecology, so ecologists commonly estimate it using a power law. Combining data on 378 trees in tropical rain forests gives this relationship between biomass y measured in kilograms and diameter x measured in centimeters:[14]

$$\log_e y = -2.00 + 2.42 \log_e x$$

Note that the investigators chose to use *natural logarithms*, with base $e = 2.71828$, rather than common logarithms with base 10.

(a) Translate the line given into a power model. Use the fact that for natural logarithms,

$$y = e^{\log_e y}$$

(b) Estimate the biomass of a tropical tree 30 centimeters in diameter.

4.23 HOW MOLD GROWS, II Find the correlation between the logarithm of mean size and hours for the data between 6 and 24 hours in Exercise 4.21. Make a scatterplot of the logarithms of the individual size measurements against hours for this same period and find the correlation. Why do we expect the second r to be smaller? Is it in fact smaller?

4.24 BE LIKE GALILEO Galileo studied motion by rolling balls down ramps. Newton later showed how Galileo's data fit his general laws of motion. Imagine that you are Galileo, without Newton's laws to guide you. He rolled a ball down a ramp at different heights above the floor and measured the horizontal distance the ball traveled before it hit the floor. Here are Galileo's data when he placed a horizontal shelf at the end of the ramp so that the ball is moving horizontally when it starts to fall. (We won't try to describe the obscure seventeenth-century units Galileo used to measure distance.)[15]

Distance	Height
1500	1000
1340	828
1328	800
1172	600
800	300

Plot distance y against height x. The pattern is very regular, as befits data described by a physical law. We want to find distance as a function of height. That is, we want to transform x to straighten the graph.

(a) Think before you calculate: Will powers x^p for $p < 1$ or $p > 1$ tend to straighten the graph. Why?

(b) Move along the ladder of transformations in the direction you have chosen until the graph is nearly straight. What transformation do you suggest?

4.25 SEED PRODUCTION Table 4.5 gives data on the mean number of seeds produced in a year by several common tree species and the mean weight (in milligrams) of the seeds produced. (Some species appear twice because their seeds were counted in two locations.) We might expect that trees with heavy seeds produce fewer of them, but what is the form of the relationship?

TABLE 4.5 Count and weight of seeds produced by common tree species

Tree species	Seed count	Seed weight (mg)	Tree species	Seed count	Seed weight (mg)
Paper birch	27,239	0.6	American beech	463	247
Yellow birch	12,158	1.6	American beech	1,892	247
White spruce	7,202	2.0	Black oak	93	1,851
Engelmann spruce	3,671	3.3	Scarlet oak	525	1,930
Red spruce	5,051	3.4	Red oak	411	2,475
Tulip tree	13,509	9.1	Red oak	253	2,475
Ponderosa pine	2,667	37.7	Pignut hickory	40	3,423
White fir	5,196	40.0	White oak	184	3,669
Sugar maple	1,751	48.0	Chestnut oak	107	4,535
Sugar pine	1,159	216.0			

Source: Data from many studies compiled in D. F. Greene and E. A. Johnson, "Estimating the mean annual seed production of trees," *Ecology*, 75 (1994), pp. 642–647.

(a) Make a scatterplot showing how the weight of tree seeds helps explain how many seeds the tree produces. Describe the form, direction, and strength of the relationship.

(b) If a power law holds for this relationship, the logarithms of the original data will display a linear pattern. Use your calculator or software to obtain the logarithms of both the seed weights and the seed counts in Table 4.5. Make a new scatterplot using these new variables. Now what are the form, direction, and strength of the relationship?

4.26 ACTIVITY 4: THE SPREAD OF CANCER CELLS

(a) Using the data you and your class collected in the chapter-opening activity, use transformation methods to construct an appropriate model. Show the important numerical and graphical steps you go through to develop your model, and tie these together with explanatory narrative to support your choice of a model.

(b) A theoretical analysis might begin as follows: The probability that an individual malignant cell reproduces is 1/6 each year. Let P = population of cancer cells at time t and let P_0 = population of cancer cells at time $t = 0$. At the end of Year 1, the population is $P = P_0 + (1/6)P_0 = P_0(7/6)$. At the end of Year 2, the population is $P = P_0(7/6) + P_0(1/6)(7/6) = P_0(7/6)^2$. Continue this line of reasoning to show that the growth equation after n years is $P = P_0(7/6)^n$.

(c) Enter the growth equation into your calculator as Y3, and plot it along with your exponential model calculated in **(a)**. Specify a thick plotting line for one of the curves. How do the two exponential curves compare?

4.2 CAUTIONS ABOUT CORRELATION AND REGRESSION

Correlation and regression are powerful tools for describing the relationship between two variables. When you use these tools, you must be aware of their limitations, beginning with the fact that **correlation and regression describe only linear relationships.** Also remember that **the correlation r and the least-squares**

regression line are not resistant. One influential observation or incorrectly entered data point can greatly change these measures. Always plot your data before interpreting regression or correlation. Here are some other cautions to keep in mind when you apply correlation and regression or read accounts of their use.

Extrapolation

Suppose that you have data on a child's growth between 3 and 8 years of age. You find a strong linear relationship between age x and height y. If you fit a regression line to these data and use it to predict height at age 25 years, you will predict that the child will be 8 feet tall. Growth slows down and stops at maturity, so extending the straight line to adult ages is foolish. Few relationships are linear for all values of x. So don't stray far from the domain of x that actually appears in your data.

EXTRAPOLATION

Extrapolation is the use of a regression line for prediction far outside the domain of values of the explanatory variable x that you used to obtain the line or curve. Such predictions are often not accurate.

Lurking variables

In our study of correlation and regression we looked at just two variables at a time. Often the relationship between two variables is strongly influenced by other variables. More advanced statistical methods allow the study of many variables together, so that we can take other variables into account. But sometimes the relationship between two variables is influenced by other variables that we did not measure or even think about. Because these variables are lurking in the background, we call them *lurking variables*.

LURKING VARIABLE

A **lurking variable** is a variable that is not among the explanatory or response variables in a study and yet may influence the interpretation of relationships among those variables.

A lurking variable can falsely suggest a strong relationship between x and y, or it can hide a relationship that is really there. Here are examples of each of these effects.

EXAMPLE 4.10 DISCRIMINATION IN MEDICAL TREATMENT?

Studies show that men who complain of chest pain are more likely to get detailed tests and aggressive treatment such as bypass surgery than are women with similar complaints. Is this association between gender and treatment due to discrimination?

Perhaps not. Men and women develop heart problems at different ages—women are on the average between 10 and 15 years older than men. Aggressive treatments are more risky for older patients, so doctors may hesitate to advise them. Lurking variables—the patient's age and condition—may explain the relationship between gender and doctors' decisions. As the author of one study of the issue said, "When men and women are otherwise the same and the only difference is gender, you find that treatments are very similar."[16]

EXAMPLE 4.11 MEASURING INADEQUATE HOUSING

A study of housing conditions in the city of Hull, England, measured a large number of variables for each of the wards in the city. Two of the variables were a measure x of overcrowding and a measure y of the lack of indoor toilets. Because x and y are both measures of inadequate housing, we expect a high correlation. In fact the correlation was only $r = 0.08$. How can this be?

Investigation found that some poor wards had a lot of public housing. These wards had high values of x but low values of y because public housing always includes indoor toilets. Other poor wards lacked public housing, and these wards had high values of both x and y. Within wards of each type, there was a strong positive association between x and y. Analyzing all wards together ignored the lurking variable—amount of public housing—and hid the nature of the relationship between x and y.[17]

Figure 4.18 shows in simplified form how groups formed by a lurking variable can make correlation and regression misleading. The groups appear as clusters of points in the scatterplot. There is a strong relationship between x and y within each of the clusters. In fact, $r = 0.85$ and $r = 0.91$ in the two clusters. However, because similar values of x correspond to quite different values of y in the two clusters, x alone is of little value for predicting y. The correlation for all the points together is only $r = 0.14$.

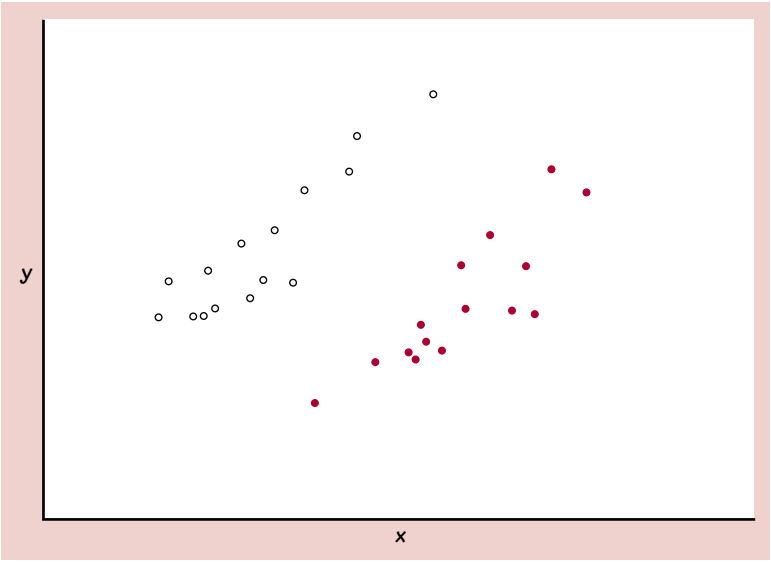


FIGURE 4.18 The variables in this scatterplot have a small correlation even though there is a strong correlation within each of the clusters.

Never forget that the relationship between two variables can be strongly influenced by other variables that are lurking in the background. Lurking variables can dramatically change the conclusions of a regression study. Because lurking variables are often unrecognized and unmeasured, detecting their effect is a challenge. Many lurking variables change systematically over time. One useful method for detecting lurking variables is therefore to *plot both the response variable and the regression residuals against the time order of the observations* whenever the time order is available. An understanding of the background of the data then allows you to guess what lurking variables might be present. Here is an example of plotting and interpreting residuals that uncovered a lurking variable.

EXAMPLE 4.12 PREDICTING ENROLLMENT

The mathematics department of a large state university must plan the number of sections and instructors required for its elementary courses. The department hopes that the number of students in these courses can be predicted from the number of first-year students, which is known before the new students actually choose courses. The table below contains data for several years.[18] The explanatory variable x is the number of first-year students. The response variable y is the number of students who enroll in elementary mathematics courses.

Year	1993	1994	1995	1996	1997	1998	1999	2000
x	4595	4827	4427	4258	3995	4330	4265	4351
y	7364	7547	7099	6894	6572	7156	7232	7450

A scatterplot (Figure 4.19) shows a reasonably linear pattern with a cluster of points near the center. We use regression software to obtain the equation of the least-squares regression line:

$$\hat{y} = 2492.69 + 1.0663x$$

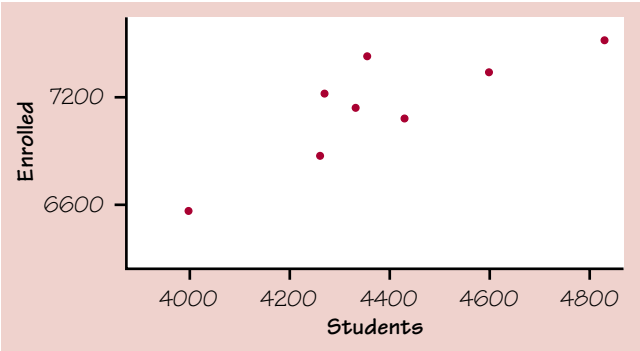


FIGURE 4.19 Enrollment in elementary math classes.

The software also tells us that $r^2 = 0.694$. That is, linear dependence on x explains about 70% of the variation in y. The line appears to fit reasonably well.

A plot of the residuals against x (Figure 4.20) magnifies the vertical deviations of the points from the line. We can see that a somewhat different line would fit the five lower points well. The three points above the line represent a different relation between the number of first-year students x and mathematics enrollments y.

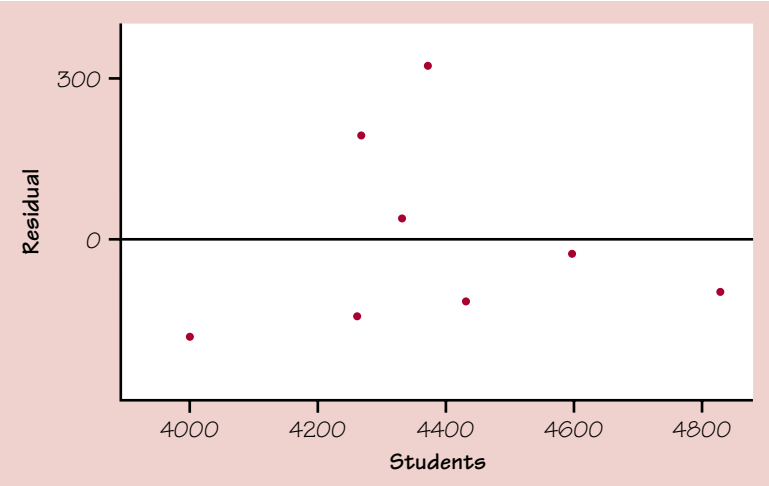


FIGURE 4.20 Residual plot.

A second plot of the residuals clarifies the situation. Figure 4.21 is a plot of the residuals against year. We now see that the five negative residuals are from the years 1993 to 1997, and the three positive residuals represent the years 1998 to 2000. This plot suggests that a change took place between 1997 and 1998 that caused a higher proportion of students to take mathematics courses beginning in 1998. In fact, one of the schools in the university changed its program to require that entering students take another mathematics course. This change is the lurking variable that explains the pattern we observed. The mathematics department should not use data from years before 1998 for predicting future enrollment.

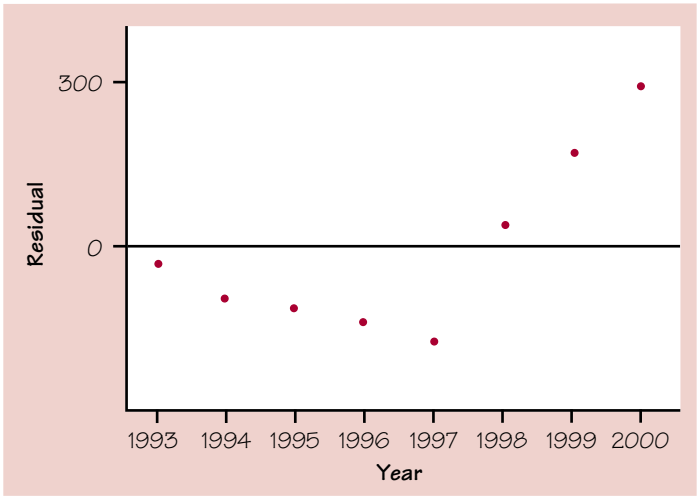


FIGURE 4.21 Plot of residuals versus year.

Using averaged data

Many regression or correlation studies work with averages or other measures that combine information from many individuals. You should note this carefully and resist the temptation to apply the results of such studies to individuals. We have seen, starting with Figure 3.2 (page 128), a strong relationship between outside temperature and the Sanchez household's natural gas consumption. Each point on the scatterplot represents a month. Both degree-days and gas consumed are averages over all the days in the month. Data for individual days would show more scatter about the regression line and lower correlation. Averaging over an entire month smooths out the day-to-day variation due to doors left open, houseguests using more gas to heat water, and so on. *Correlations based on averages are usually too high when applied to individuals.* This is another reminder that it is important to note exactly what variables were measured in a statistical study.

EXERCISES

4.27 THE SIZE OF AMERICAN FARMS The number of people living on American farms has declined steadily during this century. Here are data on the farm population (millions of persons) from 1935 to 1980.

Year:	1935	1940	1945	1950	1955	1960	1965	1970	1975	1980
Population:	32.1	30.5	24.4	23.0	19.1	15.6	12.4	9.7	8.9	7.2

(a) Make a scatterplot of these data and find the least-squares regression line of farm population on year.

(b) According to the regression line, how much did the farm population decline each year on the average during this period? What percent of the observed variation in farm population is accounted for by linear change over time?

(c) Use the regression equation to predict the number of people living on farms in 1990. Is this result reasonable? Why?

4.28 THE POWER OF HERBAL TEA A group of college students believes that herbal tea has remarkable powers. To test this belief, they make weekly visits to a local nursing home, where they visit with the residents and serve them herbal tea. The nursing home staff reports that after several months many of the residents are more cheerful and healthy. A skeptical sociologist commends the students for their good deeds but scoffs at the idea that herbal tea helped the residents. Identify the explanatory and response variables in this informal study. Then explain what lurking variables account for the observed association.

4.29 STRIDE RATE The data in Exercise 3.71 (page 187) give the average steps per second for a group of top female runners at each of several running speeds. There is a high positive correlation between steps per second and speed. Suppose that you had the full data, which record steps per second for each runner separately at each speed. If you

plotted each individual observation and computed the correlation, would you expect the correlation to be lower than, about the same as, or higher than the correlation for the published data? Why?

4.30 HOW TO SHORTEN A HOSPITAL STAY A study shows that there is a positive correlation between the size of a hospital (measured by its number of beds x) and the median number of days y that patients remain in the hospital. Does this mean that you can shorten a hospital stay by choosing a small hospital?

4.31 STOCK MARKET INDEXES The Standard & Poor's 500-stock index is an average of the price of 500 stocks. There is a moderately strong correlation (roughly $r = 0.6$) between how much this index changes in January and how much it changes during the entire year. If we looked instead at data on all 500 individual stocks, we would find a quite different correlation. Would the correlation be higher or lower? Why?

4.32 GOLF SCORES Here are the golf scores of 11 members of a women's golf team in two rounds of tournament play:

Player	1	2	3	4	5	6	7	8	9	10	11
Round 1	89	90	87	95	86	81	105	83	88	91	79
Round 2	94	85	89	89	81	76	89	87	91	88	80

(a) Plot the data with the Round 1 scores on the x axis and the Round 2 scores on the y axis. There is a generally linear pattern except for one potentially influential observation. Circle this observation on your graph.

(b) Here are the equations of two least-squares lines. One of them is calculated from all 11 data points and the other omits the influential observation.

$$\hat{y} = 20.49 + 0.754x$$
$$\hat{y} = 50.01 + 0.410x$$

Draw both lines on your scatterplot. Which line omits the influential observation? How do you know this?

The question of causation

In many studies of the relationship between two variables, the goal is to establish that changes in the explanatory variable *cause* changes in the response variable. Even when a strong association is present, the conclusion that this association is due to a causal link between the variables is often elusive. What ties between two variables (and others lurking in the background) can explain an observed association? What constitutes good evidence for causation? We begin our consideration of these questions with a set of examples. In each case, there is a clear association between an explanatory variable x and a response variable y. Moreover, the association is positive whenever the direction makes sense.

EXAMPLE 4.13 ASSOCIATIONS

The following are some examples of observed associations between x and y:

1. x = mother's body mass index
   y = daughter's body mass index
2. x = amount of the artificial sweetener saccharin in a rat's diet
   y = count of tumors in the rat's bladder
3. x = a high school senior's SAT score
   y = the student's first-year college grade point average
4. x = monthly flow of money into stock mutual funds
   y = monthly rate of return for the stock market
5. x = whether a person regularly attends religious services
   y = how long the person lives
6. x = the number of years of education a worker has
   y = the worker's income

Explaining association: causation

Figure 4.22 shows in outline form how a variety of underlying links between variables can explain association. The dashed line represents an observed association between the variables x and y. Some associations are explained by a direct cause-and-effect link between these variables. The first diagram in Figure 4.22 shows "x causes y" by a solid arrow running from x to y.

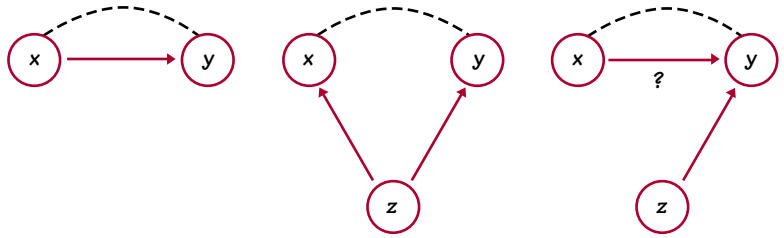


FIGURE 4.22 Variables *x* and *y* show a strong association (*dashed line*). **This association may be the result of any of several causal relationships** (*solid arrow*). (a) Causation: Changes in x cause changes in y. (b) Common response: Changes in both x and y are caused by changes in a lurking variable z. (c) Confounding: The effect (if any) of x on y is confounded with the effect of a lurking variable z.

EXAMPLE 4.14 CAUSATION?

Items 1 and 2 in Example 4.13 are examples of direct causation. Thinking about these examples, however, shows that "causation" is not a simple idea.

1. A study of Mexican American girls aged 9 to 12 years recorded body mass index (BMI), a measure of weight relative to height, for both the girls and their mothers. People with high BMI are overweight or obese. The study also measured hours of television, minutes of physical activity, and intake of several kinds of food. The strongest correlation ($r = 0.506$) was between the BMI of daughters and the BMI of their mothers.[19]

Body type is in part determined by heredity. Daughters inherit half their genes from their mothers. There is therefore a direct causal link between the BMI of mothers and daughters. Yet the mothers' BMIs explain only 25.6% (that's r^2 again) of the variation among the daughters' BMIs. Other factors, such as diet and exercise, also influence BMI. **Even when direct causation is present, it is rarely a complete explanation of an association between two variables.**

2. The best evidence for causation comes from experiments that actually change x while holding all other factors fixed. If y changes, we have good reason to think that x caused the change in y. Experiments show conclusively that large amounts of saccharin in the diet cause bladder tumors in rats. Should we avoid saccharin as a replacement for sugar in food? Rats are not people. Although we can't experiment with people, studies of people who consume different amounts of saccharin show little association between saccharin and bladder tumors.[20] **Even well-established causal relations may not generalize to other settings.**

Explaining association: common response

"Beware the lurking variable" is good advice when thinking about an association between two variables. The second diagram in Figure 4.22 illustrates ***common response.*** The observed association between the variables x and y is explained by a lurking variable z. Both x and y change in response to changes in z. This common response creates an association even though there may be no direct causal link between x and y.

common response

EXAMPLE 4.15 COMMON RESPONSE

The third and fourth items in Example 4.13 illustrate how common response can create an association.

3. Students who are smart and who have learned a lot tend to have both high SAT scores and high college grades. The positive correlation is explained by this common response to students' ability and knowledge.

4. There is a strong positive correlation between how much money individuals add to mutual funds each month and how well the stock market does the same month. Is the new money driving the market up? The correlation may be explained in part by common response to underlying investor sentiment: when optimism reigns, individuals send money to funds and large institutions also invest more. The institutions would drive up prices even if individuals did nothing. In addition, what causation there is may operate in the other direction: when the market is doing well, individuals rush to add money to their mutual funds.[21]

Explaining association: confounding

We noted in Example 4.14 that inheritance no doubt explains part of the association between the body mass indexes (BMIs) of daughters and their mothers. Can we use r or r^2 to say how much inheritance contributes to the daughters' BMIs? No. It may well be that mothers who are overweight also set an example of little exercise, poor eating habits, and lots of television. Their daughters pick up these habits to some extent, so the influence of heredity is mixed up with influences from the girls' environment. We call this mixing of influences *confounding*.

CONFOUNDING

Two variables are **confounded** when their effects on a response variable cannot be distinguished from each other. The confounded variables may be either explanatory variables or lurking variables.

When many variables interact with each other, confounding of several variables often prevents us from drawing conclusions about causation. The third diagram in Figure 4.22 illustrates confounding. Both the explanatory variable x and the lurking variable z may influence the response variable y. Because x is confounded with z, we cannot distinguish the influence of x from the influence of z. We cannot say how strong the direct effect of x on y is. In fact, it can be hard to say if x influences y at all.

EXAMPLE 4.16 CONFOUNDING

The last two associations in Example 4.13 (Items 5 and 6) are explained in part by confounding.

5. Many studies have found that people who are active in their religion live longer than nonreligious people. But people who attend church or mosque or synagogue also take better care of themselves than nonattenders. They are less likely to smoke, more likely to exercise, and less likely to be overweight. The effects of these good habits are confounded with the direct effects of attending religious services.

6. It is likely that more education is a cause of higher income—many highly paid professions require advanced education. However, confounding is also present. People who have high ability and come from prosperous homes are more likely to get many years of education than people who are less able or poorer. Of course, people who start out able and rich are more likely to have high earnings even without much education. We can't say how much of the higher income of well-educated people is actually caused by their education.

Many observed associations are at least partly explained by lurking variables. Both common response and confounding involve the influence of a

lurking variable (or variables) z on the response variable y. The distinction between these two types of relationships is less important than the common element, the influence of lurking variables. The most important lesson of these examples is one we have already emphasized: **even a very strong association between two variables is not by itself good evidence that there is a cause-and-effect link between the variables.**

Establishing causation

How can a direct causal link between x and y be established? The best method—indeed, the only fully compelling method—of establishing causation is to conduct a carefully designed experiment in which the effects of possible lurking variables are controlled. Much of Chapter 5 is devoted to the art of designing convincing experiments.

Many of the sharpest disputes in which statistics plays a role involve questions of causation that cannot be settled by experiment. Does gun control reduce violent crime? Does living near power lines cause cancer? Has increased free trade helped to increase the gap between the incomes of more educated and less educated American workers? All of these questions have become public issues. All concern associations among variables. And all have this in common: they try to pinpoint cause and effect in a setting involving complex relations among many interacting variables. Common response and confounding, along with the number of potential lurking variables, make observed associations misleading. Experiments are not possible for ethical or practical reasons. We can't assign some people to live near power lines or compare the same nation with and without free-trade agreements.

EXAMPLE 4.17 DO POWER LINES INCREASE THE RISK OF LEUKEMIA?

Electric currents generate magnetic fields. So living with electricity exposes people to magnetic fields. Living near power lines increases exposure to these fields. Really strong fields can disturb living cells in laboratory studies. What about the weaker fields we experience if we live near power lines?

It isn't ethical to do experiments that expose children to magnetic fields. It's hard to compare cancer rates among children who happen to live in more and less exposed locations, because leukemia is rare and locations vary in many ways other than magnetic fields. We must rely on studies that compare children who have leukemia with children who don't.

A careful study of the effect of magnetic fields on children took five years and cost $5 million. The researchers compared 638 children who had leukemia and 620 who did not. They went into the homes and actually measured the magnetic fields in the children's bedrooms, in other rooms, and at the front door. They recorded facts about nearby power lines for the family home and also for the mother's residence when she was pregnant. Result: no evidence of more than a chance connection between magnetic fields and childhood leukemia.[22]

"No evidence" that magnetic fields are connected with childhood leukemia doesn't prove that there is no risk. It says only that a careful study could not find any risk that stands out from the play of chance that distributes leukemia cases across the landscape. Critics continue to argue that the study failed to measure some lurking variables, or that the children studied don't fairly represent all children. Nonetheless, a carefully designed study comparing children with and without leukemia is a great advance over haphazard and sometimes emotional counting of cancer cases.

EXAMPLE 4.18 DOES SMOKING CAUSE LUNG CANCER?

Despite the difficulties, it is sometimes possible to build a strong case for causation in the absence of experiments. The evidence that smoking causes lung cancer is about as strong as nonexperimental evidence can be.

Doctors had long observed that most lung cancer patients were smokers. Comparison of smokers and similar nonsmokers showed a very strong association between smoking and death from lung cancer. Could the association be due to common response? Might there be, for example, a genetic factor that predisposes people both to nicotine addiction and to lung cancer? Smoking and lung cancer would then be positively associated even if smoking had no direct effect on the lungs. Or perhaps confounding is to blame. It might be that smokers live unhealthy lives in other ways (diet, alcohol, lack of exercise) and that some other habit confounded with smoking is a cause of lung cancer. How were these objections overcome?

Let's answer this question in general terms: What are the criteria for establishing causation when we cannot do an experiment?

- *The association is strong.* The association between smoking and lung cancer is very strong.
- *The association is consistent.* Many studies of different kinds of people in many countries link smoking to lung cancer. That reduces the chance that a lurking variable specific to one group or one study explains the association.
- *Higher doses are associated with stronger responses.* People who smoke more cigarettes per day or who smoke over a longer period get lung cancer more often. People who stop smoking reduce their risk.
- *The alleged cause precedes the effect in time.* Lung cancer develops after years of smoking. The number of men dying of lung cancer rose as smoking became more common, with a lag of about 30 years. Lung cancer kills more men than any other form of cancer. Lung cancer was rare among women until women began to smoke. Lung cancer in women rose along with smoking, again with a lag of about 30 years, and has now passed breast cancer as the leading cause of cancer death among women.
- *The alleged cause is plausible.* Experiments with animals show that tars from cigarette smoke do cause cancer.

Medical authorities do not hesitate to say that smoking causes lung cancer. The U.S. Surgeon General states that cigarette smoking is "the largest avoidable cause of death and disability in the United States."[23] The evidence for causation is overwhelming—but it is not as strong as the evidence provided by well-designed experiments.

EXERCISES

For Exercises 4.33 through 4.37, answer the question. State whether the relationship between the two variables involves causation, common response, or confounding. Identify possible lurking variable(s). Draw a diagram of the relationship in which each circle represents a variable. Write a brief description of the variable by each circle.

4.33 FIGHTING FIRES Someone says, "There is a strong positive correlation between the number of firefighters at a fire and the amount of damage the fire does. So sending lots of firefighters just causes more damage." Why is this reasoning wrong?

4.34 HOW'S YOUR SELF-ESTEEM? People who do well tend to feel good about themselves. Perhaps helping people feel good about themselves will help them do better in school and life. Raising self-esteem became for a time a goal in many schools. California even created a state commission to advance the cause. Can you think of explanations for the association between high self-esteem and good school performance other than "Self-esteem causes better work in school"?

4.35 SAT MATH AND VERBAL SCORES Table 1.15 (page 70) gives education data for the states. The correlation between the average SAT math scores and the average SAT verbal scores for the states is $r = 0.962$

(a) Find r^2 and explain in simple language what this number tells us.

(b) If you calculated the correlation between the SAT math and verbal scores of a large number of individual students, would you expect the correlation to be about 0.96 or quite different? Explain your answer.

4.36 BETTER READERS A study of elementary school children, ages 6 to 11, finds a high positive correlation between shoe size x and score y on a test of reading comprehension. What explains this correlation?

4.37 THE BENEFITS OF FOREIGN LANGUAGE STUDY Members of a high school language club believe that study of a foreign language improves a student's command of English. From school records, they obtain the scores on an English achievement test given to all seniors. The mean score of seniors who studied a foreign language for at least two years is much higher than the mean score of seniors who studied no foreign language. These data are not good evidence that language study strengthens English skills. Identify the explanatory and response variables in this study. Then explain what lurking variable prevents the conclusion that language study improves students' English scores.

SUMMARY

Correlation and regression must be **interpreted with caution.** Plot the data to be sure that the relationship is roughly linear and to detect outliers and influential observations. Remember that correlation and regression describe **only linear** relations.

Avoid **extrapolation,** which is the use of a regression line or curve for prediction for values of the explanatory variable outside the domain of the data from which the line was calculated.

Remember that **correlations based on averages** are usually too high when applied to individuals.

Lurking variables may explain the relationship between the explanatory and response variables. Correlation and regression can be misleading if you ignore important lurking variables.

The effect of lurking variables can operate through **common response** if changes in both the explanatory and response variables are caused by changes in lurking variables. **Confounding** of two variables (either explanatory or lurking variables) means that we cannot distinguish their effects on the response variable.

Most of all, be careful not to conclude that there is a cause-and-effect relationship between two variables just because they are strongly associated. The relationship could involve common response or confounding. **High correlation does not imply causation.** The best evidence that an association is due to causation comes from an **experiment** in which the explanatory variable is directly changed and other influences on the response are controlled.

In the absence of experimental evidence be cautious in accepting claims of causation. Good evidence of causation requires a strong association that appears consistently in many studies, a clear explanation for the alleged causal link, and careful examination of possible lurking variables.

SECTION 4.2 EXERCISES

For Exercises 4.38 through 4.45, carry out the instructions. Then state whether the relationship between the two variables involves causation, common response, or confounding. Then identify possible lurking variable(s). Draw a diagram of the relationship in which each circle represents a variable. By each circle, write a brief description of the variable.

4.38 DO ARTIFICIAL SWEETENERS CAUSE WEIGHT GAIN? People who use artificial sweeteners in place of sugar tend to be heavier than people who use sugar. Does this mean that artificial sweeteners cause weight gain? Give a more plausible explanation for this association.

4.39 DOES EXPOSURE TO INDUSTRIAL CHEMICALS CAUSE MISCARRIAGES? A study showed that women who work in the production of computer chips have abnormally high numbers of miscarriages. The union claimed that exposure to chemicals used in production causes the miscarriages. Another possible explanation is that these workers spend most of their time standing up.

4.40 IS MATH THE KEY TO SUCCESS IN COLLEGE? Here is the opening of a newspaper account of a College Board study of 15,941 high school graduates:

Minority students who take high school algebra and geometry succeed in college at almost the same rate as whites, a new study says.

The link between high school math and college graduation is "almost magical," says College Board President Donald Stewart, suggesting "math is the gatekeeper for success in college."

"These findings," he says, "justify serious consideration of a national policy to ensure that all students take algebra and geometry."[24]

What lurking variables might explain the association between taking several math courses in high school and success in college? Explain why requiring algebra and geometry may have little effect on who succeeds in college.

4.41 ARE GRADES AND TV WATCHING LINKED? Children who watch many hours of television get lower grades in school on the average than those who watch less TV. Explain clearly why this fact does not show that watching TV *causes* poor grades. In particular, suggest some other variables that may be confounded with heavy TV viewing and may contribute to poor grades.

4.42 MOZART FOR MINORS In 1998, the Kalamazoo (Michigan) Symphony advertised a "Mozart for Minors" program with this statement: "Question: Which students scored 51 points higher in verbal skills and 39 points higher in math? Answer: Students who had experience in music."[25] What do you think of the claim that "experience in music" causes higher test scores?

4.43 RAISING SAT SCORES A study finds that high school students who take the SAT, enroll in an SAT coaching course, and then take the SAT a second time raise their SAT mathematics scores from a mean of 521 to a mean of 561.[26] What factors other than "taking the course causes higher scores" might explain this improvement?

4.44 ECONOMISTS' EDUCATION AND INCOME There is a strong positive correlation between years of education and income for economists employed by business firms. (In particular, economists with doctorates earn more than economists with only a bachelor's degree.) There is also a strong positive correlation between years of education and income for economists employed by colleges and universities. But when all economists are considered, there is a *negative* correlation between education and income. The explanation for this is that business pays high salaries and employs mostly economists with bachelor's degrees, while colleges pay lower salaries and employ mostly economists with doctorates. Sketch a scatterplot with two groups of cases (business and academic) that illustrates how a strong positive correlation within each group and a negative overall correlation can occur together. (*Hint:* Begin by studying Figure 4.18 on page 227.)

4.45 TV AND OBESITY Over the last 20 years there has developed a positive association between sales of television sets and the number of obese adolescents in the United States. Do more TVs cause more children to put on weight, or are there other factors involved? List some of the possible lurking variables.

4.46 THE S&P 500 The Standard & Poor's 500-stock index is an average of the price of 500 stocks. There is a moderately strong correlation (roughly $r = 0.6$) between how much this index changes in January and how much it changes during the entire year.

If we looked instead at data on all 500 individual stocks, we would find a quite different correlation. Would the correlation be higher or lower? Why?

4.47 THE LINK BETWEEN HEALTH AND INCOME An article entitled "The Health and Wealth of Nations" says: "The positive correlation between health and income per capita is one of the best-known relations in international development. This correlation is commonly thought to reflect a causal link running from income to health. . . . Recently, however, another intriguing possibility has emerged: that the health-income correlation is partly explained by a causal link running the other way—from health to income."[27]

Explain how higher income in a nation can cause better health. Then explain how better health can cause higher income. There is no simple way to determine the direction of the link.

4.48 RETURNS FOR U.S. AND OVERSEAS STOCKS Exercise 3.56 (page 179) examined the relationship between returns on U.S. and overseas stocks. Return to the scatterplot and regression line for predicting overseas returns from U.S. returns.

(a) Circle the point that has the largest residual (either positive or negative). What year is this? Redo the regression without this point and add the new regression line to your plot. Was this observation very influential?

(b) Whenever we regress two variables that both change over time, we should plot the residuals against time as a check for time-related lurking variables. Make this plot for the stock returns data. Are there any suspicious patterns in the residuals?

4.49 HEART ATTACKS AND HOSPITALS If you need medical care, should you go to a hospital that handles many cases like yours? Figure 4.23 presents some data for heart attacks.

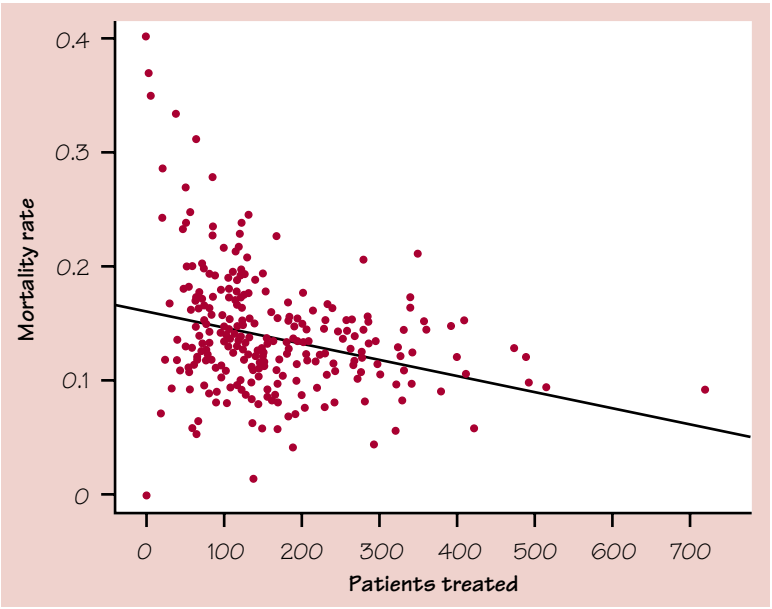


FIGURE 4.23 Mortality of heart attack patients and number of heart attack cases treated for a large group of hospitals.

The figure plots mortality rate (the proportion of patients who died) against the number of heart attack patients treated for a large number of hospitals in a recent year. The line on the plot is the least-squares regression line for predicting mortality from number of patients.

(a) Do the plot and regression generally support the thesis that mortality is lower at hospitals that treat more heart attacks? Is the relationship very strong?

(b) In what way is the pattern of the plot nonlinear? Does the nonlinearity strengthen or weaken the conclusion that heart attack patients should avoid hospitals that treat few heart attacks? Why?

4.3 RELATIONS IN CATEGORICAL DATA

To this point we have concentrated on relationships in which at least the response variable was quantitative. Now we will shift to describing relationships between two or more categorical variables. Some variables—such as sex, race, and occupation—are inherently categorical. Other categorical variables are created by grouping values of a quantitative variable into classes. Published data are often reported in grouped form to save space. To analyze categorical data, we use the *counts* or *percents* of individuals that fall into various categories.

EXAMPLE 4.19 EDUCATION AND AGE

Table 4.6 presents Census Bureau data on the years of school completed by Americans of different ages. Many people under 25 years of age have not completed their education, so they are left out of the table. Both variables, age and education, are grouped into categories. This is a ***two-way table*** because it describes two categorical variables. Education is the ***row variable*** because each row in the table describes people with one level of education. Age is the ***column variable*** because each column describes one age group. The entries in the table are the counts of persons in each age-by-education class. Although both age and education in this table are categorical variables, both have a natural order from least to most. The order of the rows and the columns in Table 4.6 reflects the order of the categories.

two-way table
row variable
column variable

TABLE 4.6 Years of school completed, by age, 2000 (thousands of persons)

	Age group			
Education	25 to 34	35 to 54	55+	Total
Did not complete high school	4,474	9,155	14,224	27,853
Completed high school	11,546	26,481	20,060	58,087
1 to 3 years of college	10,700	22,618	11,127	44,445
4 or more years of college	11,066	23,183	10,596	44,845
Total	37,786	81,435	56,008	175,230

Marginal distributions

How can we best grasp the information contained in Table 4.6 First, *look at the distribution of each variable separately.* The distribution of a categorical variable just says how often each outcome occurred. The "Total" column at the right of the table contains the totals for each of the rows. These row totals give the distribution of education level (the row variable) among all people over 25 years of age: 27,853,000 did not complete high school, 58,087,000 finished high school but did not attend college, and so on. In the same way, the "Total" row on the bottom gives the age distribution. If the row and column totals are missing, the first thing to do in studying a two-way table is to calculate them. The distributions of education alone and age alone are often called ***marginal distributions*** because they appear at the right and bottom margins of the two-way table.

marginal distributions

If you check the column totals in Table 4.6, you will notice a few discrepancies. For example, the sum of the entries in the "35 to 54" column is 81,437. The entry in the "Total" row for that column is 81,435. The explanation is ***roundoff error.*** The table entries are in the thousands of persons, and each is rounded to the nearest thousand. The Census Bureau obtained the "Total" entry by rounding the exact number of people aged 35 to 54 to the nearest thousand. The result was 81,435,000. Adding the column entries, each of which is already rounded, gives a slightly different result.

roundoff error

Percents are often more informative than counts. We can display the marginal distribution of education level in terms of percents by dividing each row total by the table total and converting to a percent.

EXAMPLE 4.20 MARGINAL DISTRIBUTION

The percent of people 25 years of age or older who have at least 4 years of college is

$$\frac{\text{total with four years of college}}{\text{table total}} = \frac{44,845}{175,230} = 0.256 = 25.6\%$$

Do three more such calculations to obtain the marginal distribution of education level in percents. Here it is.

Education:	Did not finish high school	Completed high school	1–3 years of college	≥ 4 years of college
Percent:	15.9	33.1	25.4	25.6

The total is 100% because everyone is in one of the four education categories.

Each marginal distribution from a two-way table is a distribution for a single categorical variable. As we saw in Chapter 1, we can use a bar graph or a pie chart to display such a distribution. Figure 4.24 is a bar graph of the distribu-

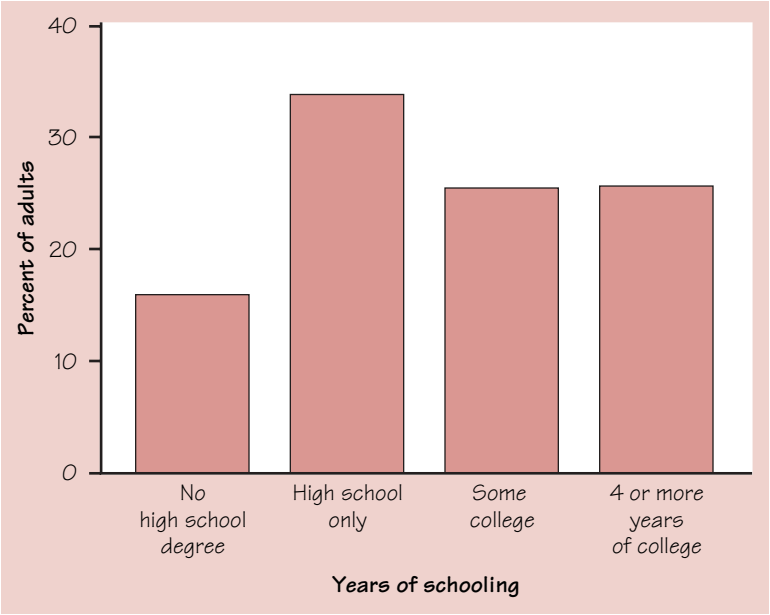


FIGURE 4.24 A bar graph of the distribution of years of schooling completed among people aged 25 years and over. This is one of the marginal distributions for Table 4.6.

tion of years of schooling. We see that people with at least some college education make up about half of the 25-or-older population.

In working with two-way tables, you must calculate lots of percents. Here's a tip to help decide what fraction gives the percent you want. Ask, "What group represents the total that I want a percent of?" The count for that group is the denominator of the fraction that leads to the percent. In Example 4.20, we wanted a percent "of people 25 or older years of age," so the count of people 25 or older (the table total) is the denominator.

Describing relationships

The marginal distributions of age and of education separately do not tell us how the two variables are related. That information is in the body of the table. How can we describe the relationship between age and years of school completed? No single graph (such as a scatterplot) portrays the form of the relationship between categorical variables, and no single numerical measure (such as the correlation) summarizes the strength of an association. *To describe relationships among categorical variables, calculate appropriate percents from the counts given.* We use percents because counts are often hard to compare. For example, 11,066,000 people age 25 to 34 have completed college, and only 10,596,000 people in the 55 and over age group have done so. But the older age group is larger, so we can't directly compare these counts.

EXAMPLE 4.21 HOW COMMON IS COLLEGE EDUCATION?

What percent of people aged 25 to 34 have completed 4 years of college? This is the count who are 25 to 34 and have 4 years of college as a percent of the age group total:

$$\frac{11{,}066}{37{,}786} = 0.293 = 29.3\%$$

"People aged 25 to 34" is the group we want a percent of, so the count for that group is the denominator. In the same way, the percent of people in the 55 and over age group who completed college is

$$\frac{10{,}596}{56{,}008} = 0.189 = 18.9\%$$

Here are the results for all three age groups:

Age group:	25 to 34	35 to 54	55+
Percent with 4 years of college:	29.3	28.5	18.9

These percents help us see how the education of Americans varies with age. Older people are less likely to have completed college.

Although graphs are not as useful for describing categorical variables as they are for quantitative variables, a graph still helps an audience to grasp the data quickly. The bar graph in Figure 4.25 presents the information in Example 4.20.

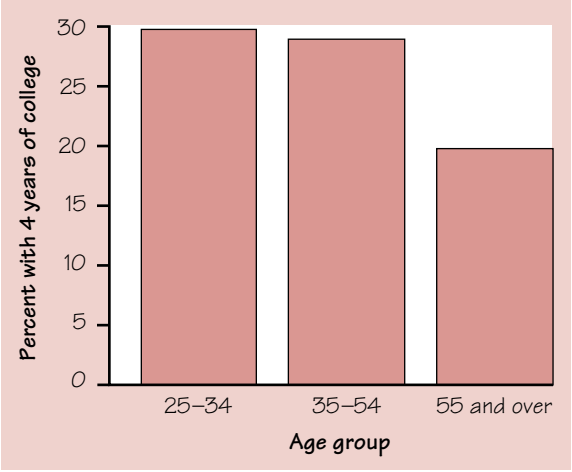


FIGURE 4.25 Bar graph comparing the percents of three age groups who have completed 4 or more years of college. The height of each bar is the percent of people in one age group who have completed at least 4 years of college.

Each bar represents one age group. The height of the bar is the percent of that age group with at least 4 years of college. Although bar graphs look a bit like histograms, their details and uses are different. A histogram shows the distribution of the values of a quantitative variable. A bar graph compares the sizes of different items. The horizontal axis of a bar graph need not have any measurement scale but may simply identify the items being compared. The items compared in Figure 4.25 are the three age groups. Because each bar in a bar graph describes a different item, we draw the bars with space between them.

EXERCISES

4.50 Sum the counts in the "55+" age column in Table 4.6 (page 241). Then explain why the sum is not the same as the entry for this column in the "Total" row.

4.51 Give the marginal distribution of age among people 25 years of age or older in percents, starting from the counts in Table 4.6 (page 241).

4.52 Using the counts in Table 4.6 (page 241), find the percent of people in each age group who did not complete high school. Draw a bar graph that compares these percents. State briefly what the data show.

4.53 SMOKING BY STUDENTS AND THEIR PARENTS Here are data from eight high schools on smoking among students and among their parents:[28]

	Neither parent smokes	One parent smokes	Both parents smoke
Student does not smoke	1168	1823	1380
Student smokes	188	416	400

(a) How many students do these data describe?

(b) What percent of these students smoke?

(c) Give the marginal distribution of parents' smoking behavior, both in counts and in percents.

4.54 PYTHON EGGS How is the hatching of water python eggs influenced by the temperature of the snake's nest? Researchers assigned newly laid eggs to one of three temperatures: hot, neutral, or cold. Hot duplicates the extra warmth provided by the mother python, and cold duplicates the absence of the mother. Here are the data on the number of eggs and the number that hatched:[29]

	Cold	Neutral	Hot
Number of eggs	27	56	104
Number hatched	16	38	75

(a) Make a two-way table of temperature by outcome (hatched or not).

(b) Calculate the percent of eggs in each group that hatched. The researchers anticipated that eggs would not hatch in cold water. Do the data support that anticipation?

4.55 IS HIGH BLOOD PRESSURE DANGEROUS? Medical researchers classified each of a group of men as "high" or "low" blood pressure, then watched them for 5 years. (Men with systolic blood pressure 140 mm Hg or higher were "high"; the others, "low.") The following two-way table gives the results of the study:[30]

	Died	Survived
Low blood pressure	21	2655
High blood pressure	55	3283

(a) How many men took part in the study? What percent of these men died during the 5 years of the study?

(b) The two categorical variables in the table are blood pressure (high or low) and outcome (died or survived). Which is the explanatory variable?

(c) Is high blood pressure associated with a higher death rate? Calculate and compare percents to answer this question.

Conditional distributions

Example 4.21 does not compare the complete distributions of years of schooling in the three age groups. It compares only the percents who finished college. Let's look at the complete picture.

EXAMPLE 4.22 CONDITIONAL DISTRIBUTION

Information about the 25 to 34 age group occupies the first column in Table 4.6. To find the complete distribution of education in this age group, look only at that column. Compute each count as a percent of the column total: 37,786. Here is the distribution:

Education:	Did not finish high school	Completed high school	1–3 years of college	≥ 4 years of college
Percent:	11.8	30.6	28.3	29.3

conditional distribution

These percents add to 100% because all 25- to 34-year-olds fall in one of the educational categories. The four percents together are the ***conditional distribution*** of education, given that a person is 25 to 34 years of age. We use the term "conditional" because the distribution refers only to people who satisfy the condition that they are 25 to 34 years old.

For comparison, here is the conditional distribution of years of school completed among people age 55 and over. To find these percents, look only at the "55+" column in Table 4.6. The column total is the denominator for each percent calculation.

Education:	Did not finish high school	Completed high school	1–3 years of college	≥ 4 years of college
Percent:	25.4	35.8	19.9	18.9

The percent who did not finish high school is much higher in the older age group, and the percents with some college and who finished college are much lower. Comparing the conditional distributions of education in different age groups describes the association between age and education. There are three different conditional distributions of education given age, one for each of the three age groups. All of these conditional distributions differ from the marginal distribution of education found in Example 4.20.

Statistical software can speed the task of finding each entry in a two-way table as a percent of its column total. Figure 4.26 displays the result. The software found the row and column totals from the table entries, so they may differ slightly from those in Table 4.6.

TABLE OF EDU BY AGE

EDU Frequency Col Pct	AGE 25-34	35-54	55 over	Total
NoHS	4474	9155	14224	27853
	11.84	11.24	25.40	
HSonly	11546	26481	20060	58087
	30.56	32.52	35.82	
SomeColl	10700	22618	11127	44445
	28.32	27.77	19.87	
Coll4yrs	11066	23183	10596	44845
	29.29	28.47	18.92	
Total	37786	81435	56008	175230

FIGURE 4.26 SAS output of the two-way table of education by age with the three conditional distributions of education, one for each age group. The percents in each column add to 100%.

Each cell in this table contains a count from Table 4.6 along with that count as a percent of the column total. The percents in each column form the conditional distribution of years of schooling for one age group.

The percents in each column add to 100% because everyone in the age group is accounted for. Comparing the conditional distributions reveals the nature of the association between age and education. The distributions of education in the two younger groups are quite similar, but higher education is less common in the 55 and over group.

Bar graphs can help make the association visible. We could make three side-by-side bar graphs, each resembling Figure 4.24 (page 243), to present the three conditional distributions. Figure 4.27 shows an alternative form of bar graph. Each set of three bars compares the percents in the three age groups who have reached a specific educational level.

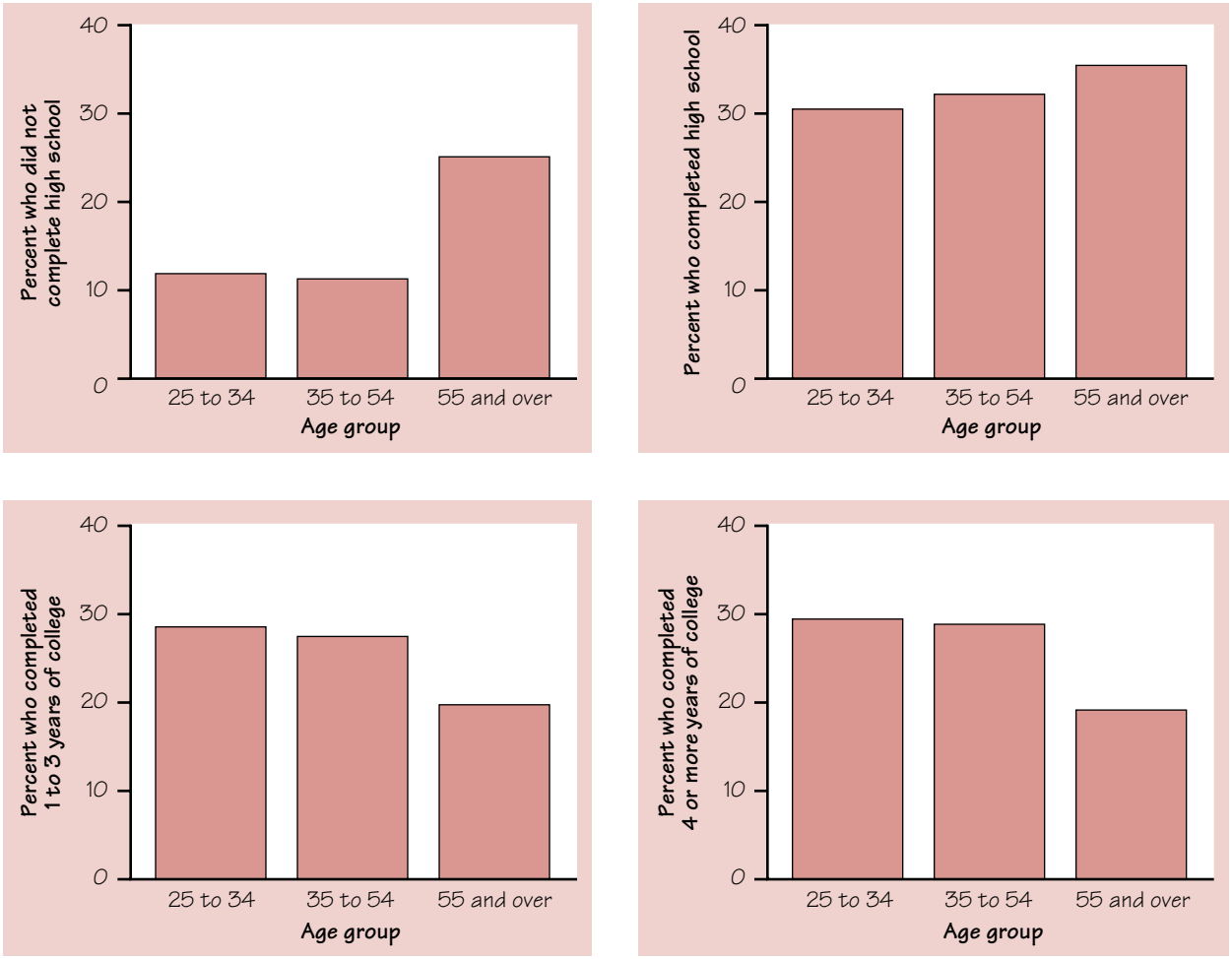


FIGURE 4.27 Bar graphs to compare the education levels of three age groups. Each graph compares the percents of three groups who fall in one of the four education levels.

We see at once that the "25 to 34" and "35 to 54" bars are similar for all four levels of education, and that the "55 and over" bars show that many more people in this group did not finish high school and that many fewer have any college.

No single graph (such as a scatterplot) portrays the form of the relationship between categorical variables. No single numerical measure (such as the correlation) summarizes the strength of the association. Bar graphs are flexible enough to be helpful, but you must think about what comparisons you want to display. For numerical measures, we rely on well-chosen percents. You must decide which percents you need. Here is a hint: compare the conditional distributions of the response variable (education) for the separate values of the explanatory variable (age). That's what we did in Figure 4.26.

In Example 4.22 we compared the education of different age groups. That is, we thought of age as the explanatory variable and education as the response variable. We might also be interested in the distribution of age among persons having a certain level of education. To do this, look only at one row in Table 4.6. Calculate each entry in that row as a percent of the row total, the total of that education group. The result is another conditional distribution, the conditional distribution of age given a certain level of education.

A two-way table contains a great deal of information in compact form. Making that information clear almost always requires finding percents. You must decide which percents you need. If you are studying trends in the training of the American workforce, comparing the distributions of education for different age groups reveals the more extensive education of younger people. If, on the other hand, you are planning a program to improve the skills of people who did not finish high school, the age distribution within this educational group is important information.

Simpson's paradox

As is the case with quantitative variables, the effects of lurking variables can change or even reverse relationships between two categorical variables. Here is a hypothetical example that demonstrates the surprises that can await the unsuspecting user of data.

EXAMPLE 4.23 PATIENT OUTCOMES IN HOSPITALS

To help consumers make informed decisions about health care, the government releases data about patient outcomes in hospitals. You want to compare Hospital A and Hospital B, which serve your community. Here is a two-way table of data on the survival of patients after surgery in these two hospitals. All patients undergoing surgery in a recent time period are included. "Survived" means that the patient lived at least 6 weeks following surgery.

	Hospital A	Hospital B
Died	63	16
Survived	2037	784
Total	2100	800

The evidence seems clear: Hospital A loses 3% (63/2100) of its surgery patients, and Hospital B loses only 2% (16/800). It seems that you should choose Hospital B if you need surgery.

Not all surgery cases are equally serious, however. Patients are classified as being in either "poor" or "good" condition before surgery. Here are the data broken down by patient condition. Check that the entries in the original two-way table are just the sums of the "poor" and "good" entries in this pair of tables.

	Good Condition			Poor Condition	
	Hospital A	Hospital B		Hospital A	Hospital B
Died	6	8	Died	57	8
Survived	594	592	Survived	1443	192
Total	600	600	Total	1500	200

Hospital A beats Hospital B for patients in good condition: only 1% (6/600) died in Hospital A, compared with 1.3% (8/600) in Hospital B. And Hospital A wins again for patients in poor condition, losing 3.8% (57/1500) to Hospital B's 4% (8/200). So Hospital A is safer for both patients in good condition and patients in poor condition. If you are facing surgery, you should choose Hospital A.

The patient's condition is a lurking variable when we compare the death rates at the two hospitals. When we ignore the lurking variable, Hospital B seems safer, even though Hospital A does better for both classes of patients. How can A do better in each group, yet do worse overall? Look at the data. Hospital A is a medical center that attracts seriously ill patients from a wide region. It had 1500 patients in poor condition. Hospital B had only 200 such cases. Because patients in poor condition are more likely to die, Hospital A has a higher death rate despite its superior performance for each class of patients. The original two-way table, which did not take account of the condition of the patients, was misleading. Example 4.23 illustrates *Simpson's paradox*.

SIMPSON'S PARADOX

Simpson's paradox refers to the reversal of the direction of a comparison or an association when data from several groups are combined to form a single group.

The lurking variables in Simpson's paradox are categorical. That is, they break the individuals into groups, as when surgery patients are classified as "good condition" or "poor condition." Simpson's paradox is just an extreme form of the fact that observed associations can be misleading when there are lurking variables.

EXERCISES

4.56 Verify that the results for the conditional distribution of education level among people aged 55 and over given in Example 4.22 (page 246) are correct.

4.57 Example 4.22 (page 246) gives the conditional distributions of education level among 25- to 34-year-olds and among people 55 and over. Find the conditional distribution of education level among 35- to 54-year-olds in percents. Is this distribution more like the distribution for 25- to 34-year-olds or the distribution for people 55 and over?

4.58 Find the conditional distribution of age among people with at least 4 years of college using the data from Example 4.22 (page 246).

4.59 MAJORS FOR MEN AND WOMEN IN BUSINESS A study of the career plans of young women and men sent questionnaires to all 722 members of the senior class in the College of Business Administration at the University of Illinois. One question asked which major within the business program the student had chosen. Here are the data from the students who responded:[31]

	Female	Male
Accounting	68	56
Administration	91	40
Economics	5	6
Finance	61	59

(a) Find the two conditional distributions of major, one for women and one for men. Based on your calculations, describe the differences between women and men with a graph and in words.

(b) What percent of the students did not respond to the questionnaire? The nonresponse weakens conclusions drawn from these data.

4.60 COLLEGE ADMISSIONS PARADOX Upper Wabash Tech has two professional schools, business and law. Here are two-way tables of applicants to both schools, categorized by gender and admission decision. (Although these data are made up, similar situations occur in reality.)[32]

	Business			Law	
	Admit	Deny		Admit	Deny
Male	480	120	Male	10	90
Female	180	20	Female	100	200

(a) Make a two-way table of gender by admission decision for the two professional schools together by summing entries in this table.

(b) From the two-way table, calculate the percent of male applicants who are admitted and the percent of female applicants who are admitted. Wabash admits a higher percent of male applicants.

(c) Now compute separately the percents of male and female applicants admitted by the business school and by the law school. Each school admits a higher percent of female applicants.

(d) This is Simpson's paradox: both schools admit a higher percent of the women who apply, but overall Wabash admits a lower percent of female applicants than of male applicants. Explain carefully, as if speaking to a skeptical reporter, how it can happen that Wabash appears to favor males when each school individually favors females.

4.61 RACE AND THE DEATH PENALTY Whether a convicted murderer gets the death penalty seems to be influenced by the race of the victim. Here are data on 326 cases in which the defendant was convicted of murder:[33]

	White defendant		Black defendant	
	White victim	Black victim	White victim	Black victim
Death	19	0	11	6
Not	132	9	52	97

(a) Use these data to make a two-way table of defendant's race (white or black) versus death penalty (yes or no).

(b) Show that Simpson's paradox holds: a higher percent of white defendants are sentenced to death overall, but for both black and white victims a higher percent of black defendants are sentenced to death.

(c) Use the data to explain why the paradox holds in language that a judge could understand.

SUMMARY

A **two-way table** of counts organizes data about two categorical variables. Values of the **row variable** label the rows that run across the table, and values of the **column variable** label the columns that run down the table. Two-way tables are often used to summarize large amounts of data by grouping outcomes into categories.

The **row totals** and **column totals** in a two-way table give the **marginal distributions** of the two individual variables. It is clearer to present these distributions as percents of the table total. Marginal distributions tell us nothing about the relationship between the variables.

To find the **conditional distribution** of the row variable for one specific value of the column variable, look only at that one column in the table. Find each entry in the column as a percent of the column total.

There is a conditional distribution of the row variable for each column in the table. Comparing these conditional distributions is one way to describe the association between the row and the column variables. It is particularly useful when the column variable is the explanatory variable.

Bar graphs are a flexible means of presenting categorical data. There is no single best way to describe an association between two categorical variables.

A comparison between two variables that holds for each individual value of a third variable can be changed or even reversed when the data for all values of the third variable are combined. This is **Simpson's paradox.** Simpson's paradox is an example of the effect of lurking variables on an observed association.

SECTION 4.3 EXERCISES

COLLEGE UNDERGRADUATES Exercises 4.62 to 4.66 are based on Table 4.7. This two-way table reports data on all undergraduate students enrolled in U.S. colleges and universities in the fall of 1995 whose age was known.

TABLE 4.7 Undergraduate college enrollment, fall 1995 (thousands of students)

Age	2-year full-time	2-year part-time	4-year full-time	4-year part-time
under 18	41	125	75	45
18 to 24	1378	1198	4607	588
25 to 39	428	1427	1212	1321
40 and up	119	723	225	605
Total	1966	3472	6119	2559

Source: Digest of Education Statistics 1997, accessed on the National Center for Education Statistics Web site, http://www.ed.gov/NCES.

4.62

(a) How many undergraduate students were enrolled in colleges and universities?

(b) What percent of all undergraduate students were 18 to 24 years old in the fall of the academic year?

(c) Find the percent of the undergraduates enrolled in each of the four types of program who were 18 to 24 years old. Make a bar graph to compare these percents.

(d) The 18 to 24 group is the traditional age group for college students. Briefly summarize what you have learned from the data about the extent to which this group predominates in different kinds of college programs.

4.63

(a) An association of two-year colleges asks: "What percent of students enrolled part-time at 2-year colleges are 25 to 39 years old?"

(b) A bank that makes education loans to adults asks: "What percent of all 25- to 39-year-old students are enrolled part-time at 2-year colleges?"

4.64

(a) Find the marginal distribution of age among all undergraduate students, first in counts and then in percents. Make a bar graph of the distribution in percents.

(b) Find the conditional distribution of age (in percents) among students enrolled part-time in 2-year colleges and make a bar graph of this distribution.

(c) Briefly describe the most important differences between the two age distributions.

(d) The sum of the entries in the "2-year part-time" column is not the same as the total given for that column. Why is this?

4.65 Call students aged 40 and up "older students." Compare the presence of older students in the four types of program with numbers, a graph, and a brief summary of your findings.

4.66 With a little thought, you can extract from Table 4.7 information other than marginal and conditional distributions. The traditional college age group is ages 18 to 24 years.

(a) What percent of all undergraduates fall in this age group?

(b) What percent of students at 2-year colleges fall in this age group?

(c) What percent of part-time students fall in this group?

4.67 FIREARM DEATHS Firearms are second to motor vehicles as a cause of nondisease deaths in the United States. Here are counts from a study of all firearm-related deaths in Milwaukee, Wisconsin, between 1990 and 1994.[34] We want to compare the types of firearms used in homicides and in suicides. We suspect that long guns (shotguns and rifles) will more often be used in suicides because many people keep them at home for hunting. Make a careful comparison of homicides and suicides, with a bar graph. What do you find about long guns versus handguns?

	Handgun	Shotgun	Rifle	Unknown	Total
Homicides	468	28	15	13	524
Suicides	124	22	24	5	175

4.68 HELPING COCAINE ADDICTS Cocaine addiction is hard to break. Addicts need cocaine to feel any pleasure, so perhaps giving them an antidepressant drug will help. A 3-year study with 72 chronic cocaine users compared an antidepressant drug called desipramine with lithium and a placebo. (Lithium is a standard drug to treat cocaine addiction. A placebo is a dummy drug, used so that the effect of being in the study but not taking any drug can be seen.) One-third of the subjects, chosen at random, received each drug. Here are the results:[35]

	Desipramine	Lithium	Placebo
Relapse	10	18	20
No relapse	14	6	4
Total	24	24	24

(a) Compare the effectiveness of the three treatments in preventing relapse. Use percents and draw a bar graph.

(b) Do you think that this study gives good evidence that desipramine actually *causes* a reduction in relapses?

4.69 SEAT BELTS AND CHILDREN Do child restraints and seat belts prevent injuries to young passengers in automobile accidents? Here are data on the 26,971 passengers under the age of 15 in accidents reported in North Carolina during two years before the law required restraints:[36]

	Restrained	Unrestrained
Injured	197	3,844
Uninjured	1,749	21,181

(a) What percent of these young passengers were restrained?

(b) Do the data provide evidence that young passengers are less likely to be injured in an accident if they wear restraints? Calculate and compare percents to answer this question.

4.70 BASEBALL PARADOX Most baseball hitters perform differently against right-handed and left-handed pitching. Consider two players, Joe and Moe, both of whom bat right-handed. The table below records their performance against right-handed and left-handed pitchers.

Player	Pitcher	Hits	At bats
Joe	Right	40	100
	Left	80	400
Moe	Right	120	400
	Left	10	100

(a) Make a two-way table of player (Joe or Moe) versus outcome (hit or no hit) by summing over both kinds of pitcher.

(b) Find the overall batting average (hits divided by total times at bat) for each player. Who has the higher batting average?

(c) Make a separate two-way table of player versus outcome for each kind of pitcher. From these tables, find the batting averages of Joe and Moe against right-handed pitching. Who does better? Do the same for left-handed pitching. Who does better?

(d) The manager doesn't believe that one player can hit better against both left-handers and right-handers yet have a lower overall batting average. Explain in simple language why this happens to Joe and Moe.

4.71 OBESITY AND HEALTH Recent studies have shown that earlier reports underestimated the health risks associated with being overweight. The error was due to overlooking lurking variables. In particular, smoking tends both to reduce weight and to lead to earlier death. Illustrate Simpson's paradox by a simplified version of this situation. That is, make up tables of overweight (yes or no) by early death (yes or no) by smoker (yes or no) such that

- Overweight smokers and overweight nonsmokers both tend to die earlier than those not overweight.
- But when smokers and nonsmokers are combined into a two-way table of overweight by early death, persons who are not overweight tend to die earlier.

CHAPTER REVIEW

In Chapter 3, we learned how to analyze two-variable data that show a linear pattern. We learned about positive and negative associations and how to measure the strength of association between two variables. We also developed a procedure for constructing a model (the least-squares regression line) that captures the trend of the data. This LSRL is useful for prediction purposes. A recurring theme is that data analysis begins with graphs and then adds numerical summaries of specific aspects of the data.

In this chapter we learned how to construct mathematical models for data that fit a curve, such as an exponential function or a power function. We also learned that although correlation and regression are powerful tools for understanding two-variable data when both variables are quantitative, both correlation and regression have their limitations. In particular, we are cautioned that a strong observed association between two variables may exist without a cause-and-effect link between them. If both variables are categorical, there is no satisfactory graph for displaying the data, although bar graphs can be helpful. We describe the relationship by comparing percents.

Here is a review list of the most important skills you should have gained from studying this chapter.

A. MODELING NONLINEAR DATA

1. Recognize that when a variable is multiplied by a fixed number greater than 1 in each equal time period, exponential growth results; when the ratio is a positive number less than 1, it's called exponential decay.

2. Recognize that when one variable is proportional to a power of a second variable, the result is a power function.

3. In the case of both exponential growth and power function, perform a logarithmic transformation and obtain points that lie in a linear pattern. Then use least-squares regression on the transformed points. An inverse transformation then produces a curve that is a model for the original points.

4. Know that deviations from the overall pattern are most easily examined by fitting a line to the transformed points and plotting the residuals from this line against the explanatory variable (or fitted values).

B. INTERPRETING CORRELATION AND REGRESSION

1. Understand that both r and the least-squares regression line can be strongly influenced by a few extreme observations.

2. Recognize possible lurking variables that may explain the observed association between two variables x and y.

3. Understand that even a strong correlation does not mean that there is a cause-and-effect relationship between x and y.

C. RELATIONS IN CATEGORICAL DATA

1. From a two-way table of counts, find the marginal distributions of both variables by obtaining the row sums and column sums.

2. Express any distribution in percents by dividing the category counts by their total.

3. Describe the relationship between two categorical variables by computing and comparing percents. Often this involves comparing the conditional distributions of one variable for the different categories of the other variable.

4. Recognize Simpson's paradox and be able to explain it.

CHAPTER 4 REVIEW EXERCISES

4.72 LIGHT INTENSITY In physics class, the intensity of a 100-watt light bulb was measured by a sensing device at various distances from the light source, and the following data were collected. Note that a *candela* (cd) is an international unit of luminous intensity.

Distance (meters)	Intensity (candelas)
1.0	0.2965
1.1	0.2522
1.2	0.2055
1.3	0.1746
1.4	0.1534
1.5	0.1352
1.6	0.1145
1.7	0.1024
1.8	0.0923
1.9	0.0832
2.0	0.0734

(a) Plot the data. Based on the pattern of points, propose a model form for the data. Then use a transformation followed by linear regression and then an inverse transformation to construct a model.

(b) Report the equation, and plot the original data with the model on the same axes.

(c) Describe the relationship between the intensity and the distance from the light source.

(d) Consult the physics textbooks used in your school and find the formula for the intensity of light as a function of distance from the light source. How do your experimental results compare with the theoretical formula?

4.73 PENDULUM An experiment was conducted with a pendulum of variable length. The *period*, or length of time to complete one complete oscillation, was recorded for several lengths. Here are the data:

Length (feet):	1	2	3	4	5	6	7
Period (seconds):	1.10	1.56	1.92	2.20	2.50	2.71	2.93

(a) Make a plot of period against length. Describe the pattern that you see.

(b) Propose a model form. Then use a transformation to construct a model for the data. Report the equation, and plot the original data with the model on the same axes.

(c) Describe the relationship between the length of a pendulum and its period.

4.74 EXACT EXPONENTIAL GROWTH, I A clever courtier, offered a reward by an ancient king of Persia, asked for a grain of rice on the first square of a chess board, 2 grains on the second square, then 4, 8, 16, and so on.

(a) Make a table of the number of grains on each of the first 10 squares of the board.

(b) Plot the number of grains on each square against the number of the square for squares 1 to 10, and connect the points with a smooth curve. This is an exponential curve.

(c) How many grains of rice should the king deliver for the 64th (and final) square?

(d) Take the logarithm of each of your numbers of grains from **(a)**. Plot these logarithms against the number of squares from 1 to 10. You should get a straight line.

(e) From your graph in **(d)** find the approximate values of the slope b and the intercept a for the line. Use the equation $y = a + bx$ to predict the logarithm of the amount for the 64th square. Check your result by taking the logarithm of the amount you found in **(c)**.

4.75 800-METER RUN Return to the 800-meter world record times for men and women of Exercise 3.75 (page 188). Suppose you are uncomfortable with the linear model for the declinr in winning times that will eventually intersect the horizontal axis.

(a) Construct exponential and power regression models for the *men's* record times. Which do you consider to be a better model?

(b) Based on your answer to **(a)**, construct a similar model for the *women's* record times.

(c) Will either of these curves eventually reach zero? Will the curves intersect each other? If so, in what year will the curves intersect?

(d) Is this a satisfactory model, or is there a better model tor these data?

4.76 SOCIAL INSURANCE Federal expenditures on social insurance (chiefly social security and Medicare) increased rapidly after 1960. Here are the amounts spent, in millions of dollars:

Year:	1960	1965	1970	1975	1980	1985	1990
Spending:	14,307	21,807	45,246	99,715	191,162	310,175	422,257

(a) Plot social insurance expenditures against time. Does the pattern appear closer to linear growth or to exponential growth?

(b) Take the logarithm of the amounts spent. Plot these logarithms against time. Do you think that the exponential growth model fits well?

(c) After entering the data into the Minitab statistical system, with year as C1 and expenditures as C2, we obtain the least-squares line for the logarithms as follows:

```
MTB> LET C3 = LOGT(C2)
MTB> REGRESS C3 ON 1, C1

The regression equation is
C3 = -98.63833 + 0.05244 C1
```

That is, the least-squares line is

$$\log y = -98.63833 + (0.05244 \times \text{year})$$

Draw this line on your graph from **(b)**.

(d) Use this line to predict the logarithm of social insurance outlays for 1988. Then compute

$$y = 10^{\log y}$$

to predict the amount y spent in 1988.

(e) The actual amount (in millions) spent in 1988 was $358,412. Take the logarithm of this amount and add the 1988 point to your graph in **(b)**. Does it fall close to the line? When President Reagan took office in 1981, he advocated a policy of slowing growth in spending on social progams. Did the trend of exponential growth in spending for social insurance change in a major way during the Reagan years, 1981 to 1988?

4.77 KILLING BACTERIA Expose marine bacteria to X-rays for time periods from 1 to 15 minutes. Here are the number of surviving bacteria (in hundreds) on a culture plate after each exposure time:[37]

Time t	Count y	Time t	Count y
1	355	9	56
2	211	10	38
3	197	11	36
4	166	12	32
5	142	13	21
6	106	14	19
7	104	15	15
8	60		

Theory suggests an exponential growth or decay model. Do the data appear to conform to this theory?

4.78 BANK CARDS Electronic fund transfers, from bank automatic teller machines and the use of debit cards by consumers, have grown rapidly in the United States. Here are data on the number of such transfers (in millions).[38]

Year	EFT	Year	EFT	Year	EFT
1985	3,579	1991	6,642	1996	11,780
1987	4,108	1992	7,537	1997	12,580
1988	4,581	1993	8,135	1998	13,160
1989	5,274	1994	9,078	1999	13,316
1990	5,942	1995	10,464		

Write a clear account of the pattern of growth of electronic transfers over time, supporting your description with plots and calculations as needed. Has the pattern changed in the most recent years?

4.79 ICE CREAM AND FLU There is a negative correlation between the number of flu cases reported each week throughout the year and the amount of ice cream sold in that particular week. It's unlikely that ice cream prevents flu. What is a more plausible explanation for this observed correlation?

4.80 VOTING FOR PRESIDENT The following table gives the U.S. resident population of voting age and the votes cast for president, both in thousands, for presidential elections between 1960 and 2000:

Year	Population	Votes	Year	Population	Votes
1960	109,672	68,838	1984	173,995	92,653
1964	114,090	70,645	1988	181,956	91,595
1968	120,285	73,212	1992	189,524	104,425
1972	140,777	77,719	1996	196,511	96,456
1976	152,308	81,556	2000	209,128	105,363
1980	163,945	86,515			

(a) For each year compute the percent of people who voted. Make a time plot of the percent who voted. Describe the change over time in participation in presidential elections.

(b) Before proposing political explanations for this change, we should examine possible lurking variables. The minimum voting age in presidential elections dropped from 21 to 18 years in 1970. Use this fact to propose a partial explanation for the trend you saw in (a).

4.81 WOMEN AND MARITAL STATUS The following two-way table describes the age and marital status of American women in 2000. The table entries are in thousands of women.

	Marital status				
Age	Single	Married	Widowed	Divorced	Total
15–24	16,121	2,694	21	203	19,040
25–39	7,409	19,925	212	2,965	30,510
40–64	3,553	29,687	2,338	6,797	42,373
≥65	680	8,223	8,490	1,344	18,735
Total					110,660

(a) Find the sum of the entries in the 15–24 row. Why does this sum differ from the "Total" entry for that row?

(b) Give the marginal distribution of marital status for all adult women (use percents). Draw a bar graph to display this distribution.

(c) Compare the conditional distributions of marital status for women aged 15 to 24 and women aged 40 to 64. Briefly describe the most important differences between the two groups of women, and back up your description with percents.

(d) You are planning a magazine aimed at single women who have never been married. (That's what "single" means in government data.) Find the conditional distribution of ages among single women.

4.82 WOMEN SCIENTISTS A study by the National Science Foundation[39] found that the median salary of newly graduated female engineers and scientists was only 73% of the median salary for males. When the new graduates were broken down by field, however, the picture changed. Women's median salaries as a percent of the male median in the 16 fields studied were

94%	96%	98%	95%	85%	85%	84%	100%
103%	100%	107%	93%	104%	93%	106%	100%

How can women do nearly as well as men in every field yet fall far behind men when we look at all young engineers and scientists?

4.83 SMOKING AND STAYING ALIVE In the mid-1970s, a medical study contacted randomly chosen people in a district in England. Here are data on the 1314 women contacted who were either current smokers or who had never smoked. The table classifies these women by their smoking status and age at the time of the survey and whether they were still alive 20 years later.[40]

	Age 18 to 44	
	Smoker	Not
Dead	19	13
Alive	269	327

	Age 45 to 64	
	Smoker	Not
Dead	78	52
Alive	167	147

	Age 65+	
	Smoker	Not
Dead	42	165
Alive	7	28

(a) Make a two-way table of smoking (yes or no) by dead or alive. What percent of the smokers stayed alive for 20 years? What percent of the nonsmokers survived? It seems surprising that a higher percent of smokers stayed alive.

(b) The age of the women at the time of the study is a lurking variable. Show that within each of the three age groups in the data, a higher percent of nonsmokers remained alive 20 years later. This is another example of Simpson's paradox.

(c) The study authors give this explanation: "Few of the older women (over 65 at the original survey) were smokers, but many of them had died by the time of follow-up." Compare the percent of smokers in the three age groups to verify the explanation.

NOTES AND DATA SOURCES

1. This activity was described in Elizabeth B. Applebaum, "A simulation to model exponential growth," *Mathematics Teacher*, 93, No.7 (October 2000), pp. 614–615.
2. Data from G. A. Sacher and E. F. Staffelt, "Relation of gestation time to brain weight for placental mammals: implications for the theory of vertebrate growth," *American Naturalist*, 108 (1974), pp. 593–613. We found these data in F. L. Ramsey and D. W. Schafer, *The Statistical Sleuth: A Course in Methods of Data Analysis*, Duxbury, 1997.
3. There are several mathematical ways to show that log t fits into the power family at $p = 0$. Here's one. For powers $p \neq 0$, the indefinite integral $\int t^{p-1}\,dt$ is a multiple of t^p. When $p = 0$, $\int t^{-1}dt$ is log t.
4. Data from the World Bank's *1999 World Development Indicators*. Life expectancy is estimated for 1997, and GDP per capita (purchasing-power parity basis) is estimated for 1998.
5. The power law connecting heart rate with body weight was found online at "The Worldwide Anaesthetist," www.anaesthetist.com. Anesthesiologists are interested in power laws because they must judge how drug doses should increase in bigger patients.
6. Data from *Statistical Abstract of the United States, 2000*. Data for Alaska and Hawaii were included for the first time in 1950.
7. Gypsy moth data provided by Chuck Schwalbe, U.S. Department of Agriculture.
8. From Intel Web site, www.intel.com/research/silicon/mooreslaw.htm.
9. From Joel Best, *Damned Lies and Statistics: Untangling Numbers from the Media, Politicians, and Activists*, University of California Press, Berkeley and Los Angeles, 2001.
10. Fish data from Gordon L. Swartzman and Stephen P. Kaluzny, *Ecological Simulation Primer*, Macmillan, New York, 1987, p. 98.
11. Data originally from A. J. Clark, *Comparative Physiology of the Heart*, Macmillan, New York, 1927, p. 84. Obtained from Frank R. Giordano and Maurice D. Weir, *A First Course in Mathematical Modeling*, Brooks/Cole, Belmont, Calif., 1985, p. 56.
12. Data on a sample of 12 of 56 perch in a data set contributed to the *Journal of Statistics Education* data archive (www.amstat.org/publications/jse/) by Juha Puranen of the University of Helsinki.
13. Similar experiments are described by A. P. J. Trinci, "A kinetic study of the growth of *Aspergillus nidulans* and other fungi," *Journal of General Microbiology*, 57 (1969), pp. 11–24. These data were provided by Thomas Kuczek, Purdue University.
14. Jérôme Chave, Bernard Riéra, and Marc-A. Dubois, "Estimation of biomass in a neotropical forest of French Guiana: spatial and temporal variability," *Journal of Tropical Ecology*, 17 (2001), pp. 79–96.
15. Data from Stillman Drake, *Galileo at Work*, University of Chicago Press, 1978. We found these data in D. A. Dickey and J. T. Arnold, "Teaching statistics with data of historic significance," *Journal of Statistics Education*, 3 (1995), www.amstat.org/publications/jse/.
16. The quotation is from Dr. Daniel Mark of Duke University, in "Age, not bias, may explain differences in treatment," *New York Times*, April 26, 1994.

17. This example is drawn from M. Goldstein, "Preliminary inspection of multivariate data," *The American Statistician*, 36(1982), pp. 358–362.
18. Data provided by Peter Cook, Department of Mathematics, Purdue University.
19. Laura L. Calderon *et al.*, "Risk factors for obesity in Mexican-American girls: dietary factors, anthropometric factors, physical activity, and hours of television viewing," *Journal of the American Dietetic Association*, 96 (1996), pp. 1177–1179.
20. Saccharin appears on the National Institute of Health's list of suspected carcinogens but remains in wide use. In October 1997, an expert panel recommended by a 4 to 3 vote to keep it on the list, despite recommendations from other scientific groups that it be removed.
21. A detailed study of this correlation appears in E. M. Remolona, P. Kleinman, and D. Gruenstein, "Market returns and mutual fund flows," *Federal Reserve Bank of New York Economic Policy Review*, 3, No. 2(1997), pp. 33–52.
22. M. S. Linet *et al.*, "Residential exposure to magnetic fields and acute lymphoblastic leukemia in children," *New England Journal of Medicine*, 337 (1997), pp.1–7.
23. *The Health Consequences of Smoking: 1983*, U.S. Health Service, 1983.
24. From a Gannett News Service article appearing in the Lafayette (Indiana) *Journal and Courier*, April 23, 1994.
25. Contributed by Marigene Arnold of Kalamazoo College.
26. D. E. Powers and D. A. Rock, *Effects of Coaching on SAT I: Reasoning Test Scores*, Educational Testing Service Research Report 98-6, College Entrance Examination Board, 1998.
27. David E. Bloom and David Canning, "The health and wealth of nations," *Science*, 287 (2000), pp. 1207–1208.
28. From S. V. Zagona (ed.), *Studies and Issues in Smoking Behavior*, University of Arizona Press, Tucson, 1967, pp. 157–180.
29. R. Shine, T. R. L. Madsen, M. J. Elphick, and P. S. Harlow, "The influence of nest temperature and maternal brooding on hatchling phenotypes in water python," *Ecology*, 78 (1997), pp. 1713–1721.
30. From J. Stamler, "The mass treatment of hypertensive disease: defining the problem," *Mild Hypertension: To Treat or Not to Treat*, New York Academy of Sciences, 1978, pp. 333–358.
31. From F. D. Blau and M. A. Ferber, "Career plans and expectations of young women and men," *Journal of Human Resources*, 26 (1991), pp. 581–607.
32. See P. J. Bickel and J. W. O'Connell, "Is there a sex bias in graduate admissions?" *Science*, 187 (1975), pp. 398–404.
33. From M. Radelet, "Racial characteristics and imposition of the death penalty," *American Sociological Review*, 46 (1981), pp. 918–927.
34. S. W. Hargarten *et al.*, "Characteristics of firearms involved in fatalities," *Journal of the American Medical Association*, 275 (1996), pp. 42–45.
35. From D. M. Barnes, "Breaking the cycle of addiction," *Science*, 241 (1988), pp. 1029–1030.
36. Adapted from data of Williams and Zador in *Accident Analysis and Prevention*, 9 (1977), pp. 69–76.
37. S. Chatterjee and B. Price, *Regression Analysis by Example*, Wiley, New York, 1977.
38. From several editions of the *Statistical Abstract of the United States*.

39. National Science Board, *Science and Engineering Indicators, 1991*, U.S. Government Printing Office, Washington, D.C., 1991. The detailed data appear in Appendix Table 3-5, p. 274.

40. Condensed from D. R. Appleton, J. M. French, and M. P. J. Vanderpump, "Ignoring a covariate: an example of Simpson's paradox," *The American Statistician*, 50 (1996), pp. 340–341.