CHAPTER

13 Applications of Statistics

This untitled painting (1990) by Japanese-born artist Naoki Okamoto creates a “crowd” of faces. In this chapter you’ll explore how numerical data about a population can be represented with a few summary numbers. You’ll also explore whether data about a small but random group of people can lead to valid generalizations about a whole population.

OBJECTIVES

In this chapter you will

- **learn methods for making predictions about a population based on a random sample**
- **discover the association between the statistics of a sample and the parameters of an entire population**
- **study population distributions, including normal distributions**
- **fit functions to data and make predictions using least squares lines and other regression equations**

© 2004 Key Curriculum Press

LESSON

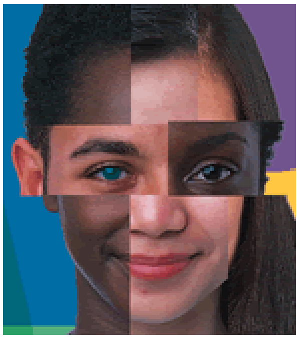
13.1

Probability Distributions

Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.

H. G. WELLS

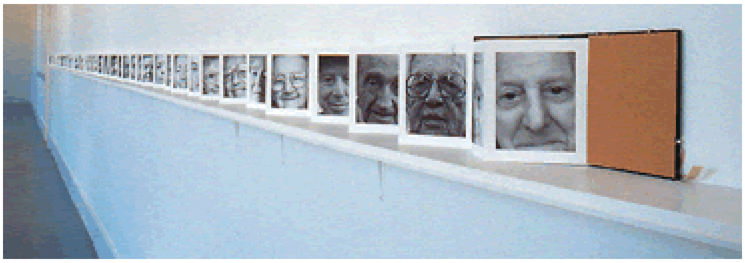
"I did a survey and 22 percent of the kids in this school have blue eyes, so 22 percent of the people in this neighborhood must have blue eyes," claims Sean. "Exactly 22 percent?" asks Yiscah.



One of the main uses of statistics is to find out about large collections, such as the residents of a city, by looking at smaller collections, such as the students in the school. The larger collection is called a **population,** and the smaller collection is a **sample** of that population.

You've seen in earlier chapters that statistics are numbers, such as mean or standard deviation, that describe a sample. The corresponding numbers describing the entire population are called **parameters.** The larger the sample, the closer its statistics will be to the parameters.

When you examined probabilities in Chapter 12, you used discrete random variables. The data had integer values, such as 5 heads, 3 tails, or 454 students. However, sometimes data can take on any real value within an interval. This is represented by a **continuous random variable.** For example, the ages of all the students in your class are continuous. You might say that everyone is 15 or 16 years old; but actually no one is exactly 16, because a person is exactly 16 at only one instant. Your age is a continuous variable, and there are infinitely many of these ages.



This 14-meter installation, shown here in two views, is called *100 edition of 12* (1995). To create this piece, American artist Paul Ramirez Jonas (b 1965) chronologically arranged 100 photos of people aged 0-99.

© 2004 Key Curriculum Press

Investigation

Pencil Lengths

You will need

- centimeter rulers
- a few pencils
- graph paper

In this investigation you'll explore the difference between discrete and continuous random variables.

Begin by collecting all the pencils that your group has.

Step 1 Measure your pencils accurate to a tenth of a centimeter. Before you share data with other groups, predict the shape of a histogram of the class data.

Step 2 Share all measurements so that the class has one set of data. On graph paper, draw a histogram with bins representing 1 cm increments in pencil length.

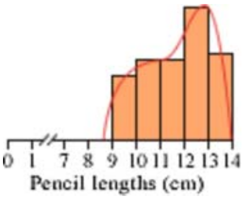
Step 3 Divide the number of pencils in each bin by the total number of pencils. Make a new histogram, using these quotients as the values on the y-axis.

Step 4 Check that the area of your second histogram is 1. Why must this be true?

Step 5 Imagine that you collect more and more pencils and draw a histogram using the method described in Step 3. Sketch what this histogram of infinitely many pencil lengths would look like. Give reasons for your answer.

Step 6 Imagine doing a very complete and precise survey of all the pencils in the world. Assume that their distribution is about the same as the distribution of pencils in your sample. Also assume that you use infinitely many very narrow bins. What will this histogram look like?

To approximate this plot, sketch over the top of your histogram with a smooth curve, as shown at right. Make the area between the curve and the horizontal axis about the same as the area of the histogram. Make sure that the extra area enclosed by the curve above the histogram is about the same as the area cut off the corners of the bins as you smooth out the shape.



Step 7 Let x represent pencil length. Use your histogram from Step 6 to estimate the areas of various regions between the curve and the x-axis that satisfy these conditions:

a. $x < 10$

b. $11 < x < 12$

c. $x > 12.5$

d. $x = 11$

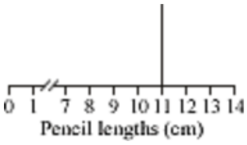
The histogram you made in Step 3 of the investigation, giving the proportions of pencils in the bins, is a **relative frequency histogram.** It shows what fraction of the time the value of a discrete random variable falls within each bin. The continuous curve you drew in Step 6 approximates a continuous random variable for the infinite set of measurements. This graph represents a function called a **probability distribution.**

© 2004 Key Curriculum Press

The areas you found in Step 7 of the investigation give probabilities that a randomly chosen pencil length will satisfy a condition. As with geometric probabilities in Lesson 12.1, these probabilities are given by areas. If x represents the continuous random variable giving the pencil lengths in centimeters, then you can write these areas as

$P(x < 10 \text{ cm})$, $P(11 \text{ cm} < x < 12 \text{ cm})$, $P(x > 12.5 \text{ cm})$, and $P(x = 11 \text{ cm})$

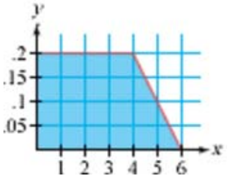
In a continuous probability distribution, the probability of any single outcome, such as the probability that x is exactly 11 centimeters, is the area of a line segment, which is 0. It is possible for a pencil to be exactly 11 cm long, but the probability of choosing any one value from infinitely many values is theoretically 0. As you learned in geometry, a single point or line has no area.



In the following example, you'll see how areas represent probabilities for a continuous random variable.

EXAMPLE A

A random-number generator selects a number between 0 and 6 according to the probability distribution at right. Because the random number can be any value of x with $0 \le x \le 6$, the graph is a continuous graph. Find the probability that a selected number is



a. Less than 2.

b. Between 2.5 and 3.5.

c. More than 4.

Solution

First, note that the region shaded for the entire distribution has area 1. To find the probability of a particular set of outcomes, find the area of the region that corresponds to it.

a. The region between 0 and 2 is a rectangle with width 2 and height .2. Its area is $2 \cdot .2$, or .4. So, the probability is .4 that a randomly selected number is between 0 and 2.

b. The region between 2.5 and 3.5 is a rectangle with width 1 and height .2. The area of this rectangle is $1 \cdot .2$, or .2. So, the probability is .2 that a number is between 2.5 and 3.5.

c. The region between 4 and 6 is a triangle with width 2 and height .2. The area of the triangle is $0.5 \cdot 2 \cdot .2 = .2$, so the probability is .2.

In Chapter 2 you learned three measures of center to describe a data set-mean, median, and mode. With a probability distribution you don't have a finite set of data. So these statistics must be defined and calculated in a slightly different way.

© 2004 Key Curriculum Press

Measures of Center for Probability Distributions

Mode

The value or values of x at which the graph reaches its maximum value.

Median

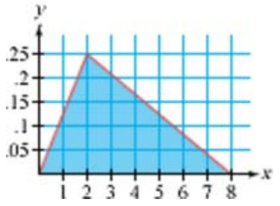
The number d such that the line $x = d$ divides the area into two parts with equal areas.

Mean

The sum of each value of x times its probability. Also, the x-coordinate of the centroid, or balance point, of the region.

EXAMPLE B

A large number of people were asked to complete a puzzle. The time it took each person was recorded. The data are shown in the probability distribution graph at right, with times ranging between 0 and 8 seconds.



a. Find the mode.

b. Find the median.

c. Find the mean.

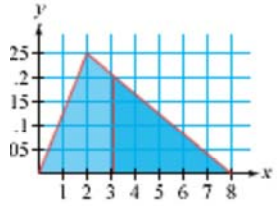
▸ Solution

Note that the shaded region has area 1.

a. The mode is the x-coordinate of the highest point, 2 seconds.

b. Find the vertical line that divides the triangle into two regions each having area .5. Some trial and error shows the median is about 3 s. The smaller triangle has base 5 and height about .2.

$A \approx 0.5 \cdot (5)\,(.2) \approx .5$



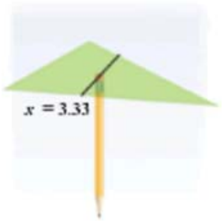
So about half the area of the larger triangle falls after 3. To calculate the median exactly, you can use equations of the lines that form boundaries of the region. The equation of the line through (2, .25) and (8, 0) is $y = -\frac{1}{24}(x - 8)$. Use this equation to find the value of the median, d, so that the area of the triangle to the right of the median is .5.

$A = 0.5bh = 0.5(8 - d)\left(\frac{-1}{24}(d - 8)\right)$	The area of a triangle is half the product of the base $(8 - d)$ and height. The height is the y-value at $x = d$.
$0.5 = 0.5(8 - d)\left(\frac{-1}{24}(d - 8)\right)$	Solve for d when the area is 0.5.
$1 = \frac{-1}{24}(-d^2 - 16d - 64)$	Divide both sides by 0.5 and multiply the binomials.
$-24 = -d^2 - 16d - 64$	Multiply both sides by –24.
$d^2 - 16d + 40 = 0$	Write the quadratic equation in general form.
$d \approx 3.101, 12.899$	Use the quadratic formula to solve for d.

© 2004 Key Curriculum Press

The second value, 12.899, is outside of the domain, so the median of this distribution is about 3.101.

c. By cutting out the triangle and balancing it on the eraser end of a pencil, you can get a pretty good estimate that the mean is 3.33 in.



Because the distribution forms a triangle, you can find its centroid using geometric construction or algebra. You might find the intersection of two medians of the triangle. Or you might recall from geometry that the coordinates of the centroid are the means of the coordinates of the vertices. The *x*-coordinate of the centroid, then, is the mean of the *x*-coordinates:

$$\frac{0+2+8}{3} = \frac{10}{3} = 3\frac{1}{3}$$

Calculus can be used to find the measures of center of some more-complicated regions. But for most probability distributions, finding the exact values of these parameters is almost impossible. They must be estimated using methods such as the balancing approach.

EXERCISES

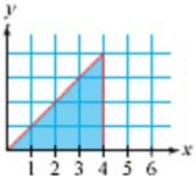
You will need

Geometry software for Exercise **17**

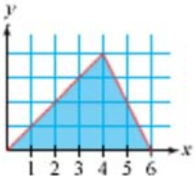
Practice Your Skills

Answer Exercises 1-4 on page 729 for each probability distribution below.

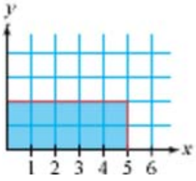
Distribution A



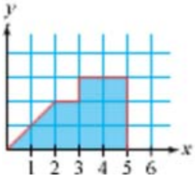
Distribution B



Distribution C



Distribution D



© 2004 Key Curriculum Press

1. Find the height of one grid box on the y-scale so that the area is 1.

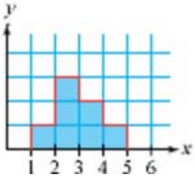
2. Find the probability that a randomly chosen value will be less than 3.

3. Estimate the median.

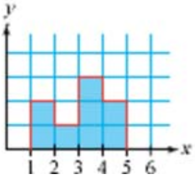
4. Estimate the mean.

5. Draw a probability distribution for each histogram below. Try to keep the area under the curve the same as the area under the histogram.

a.



b.



Reason and Apply

6. Suppose each person in your class selects a set of four numbers from 1 to 8 (repeats are allowed) and that each person calculates the mean of his or her own set.

   a. Sketch a possible histogram of these mean values. Explain the reasoning behind your histogram.
   b. Based on your histogram, estimate the mode and median of your distribution.

7. Classify each statement as true or false and if false, explain why.

   a. The y-value of the mode in a probability distribution can never be more than 1.
   b. It is impossible to tell how many data values were used to create a probability distribution.
   c. The mean, median, and mode of a continuous distribution can never all be the same value.

8. Imagine finding many random numbers from 0 to 1 and substituting them into each expression below. Sketch what you think the relative frequency histograms or probability distributions of the results would look like, and explain your reasoning.

   a. $(random\ number)^2$ **b.** $(random\ number)^4$ **c.** $\sqrt{random\ number}$

9. ***Technology*** Use statistics software or lists on your graphing calculator to investigate Exercise 8.

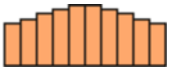
10. Sketch a relative frequency histogram to fit each set of conditions. You may want to sketch these using the squares on graph paper to be certain you have a total area of 1. Label each axis with an appropriate scale.

    a. The data values are continuous from 0 to 10. The mean and median are the same value.
    b. The data values are continuous from 10 to 15. The mean is larger than the median.

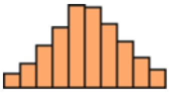
© 2004 Key Curriculum Press

11. Describe a data set that might produce each of these continuous distribution graphs. Indicate the range of values on the *x*-axis.

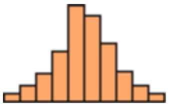
a.



b.



c.



12. This table lists the ages of the presidents and vice presidents of the United States when they first took office.

President	Age
Washington	57
J. Adams	61
Jefferson	57
Madison	57
Monroe	58
J. Q. Adams	57
Jackson	61
Van Buren	54
W. Harrison	68
Tyler	51
Polk	49
Taylor	64
Fillmore	50
Pierce	48
Buchanan	65
Lincoln	52
A. Johnson	56
Grant	46
Hayes	54
Garfield	49
Arthur	50
Cleveland	47
B. Harrison	55
McKinley	54
T. Roosevelt	42
Taft	51
Wilson	56
Harding	55
Coolidge	51
Hoover	54
F. D. Roosevelt	51
Truman	60
Eisenhower	62
Kennedy	43
L. B. Johnson	55
Nixon	56
Ford	61
Carter	52
Reagan	69
G. H. W. Bush	64
W. Clinton	46
G. W. Bush	54

Vice president	Age
J. Adams	53
Jefferson	53
Burr	45
G. H. Clinton	65
Gerry	68
Tompkins	42
Calhoun	42
Van Buren	50
R. M. Johnson	56
Tyler	50
Dallas	52
Fillmore	49
King	66
Breckinridge	36
Hamlin	51
A. Johnson	56
Colfax	45
Wilson	61
Wheeler	57
Arthur	51
Hendricks	65
Morton	64
Stevenson	57
Hobart	52
T. Roosevelt	42
Fairbanks	52
Sherman	53
Marshall	58
Coolidge	48
Dawes	59
Curtis	69
Garner	64
Wallace	52
Truman	60
Barkley	71
Nixon	40
L. B. Johnson	52
Humphrey	53
Agnew	50
Ford	60
Rockefeller	66
Mondale	49
G. H. W. Bush	56
Quayle	41
Gore	44
Cheney	59

(The World Almanac and Book of Facts 2003)

a. Enter the two separate lists of data into your calculator and calculate the mean, $\bar{x}$, the standard deviation, *s*, the median, and the *IQR* for each list. Compare the data sets based on these statistics.

b. Graph a histogram for each data set. Use the same range and scale for each graph. Describe how the histograms reflect the statistics of each data set.

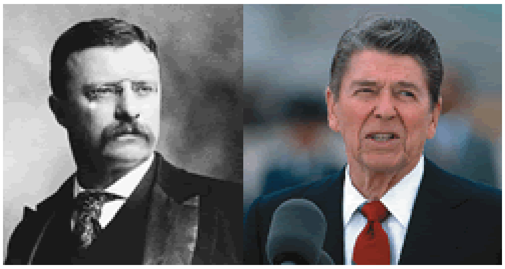
c. Calculate $\frac{x_i - \bar{x}}{s}$ for each entry, and create two new lists to convert the ages in each list to a standardized scale.

© 2004 Key Curriculum Press

d. What is the range of values in each of the new distributions? Explain what the new distributions represent.

e. Graph a histogram for each of these standardized distributions. Use domain $-3.5 \leq x \leq 3.5$.

f. Compare and describe the graphs.



The U.S. Constitution specifies that presidents must be natural-born citizens, have lived in the United States for at least 14 years, and be at least 35 years old. The 26th U.S. president, Theodore Roosevelt (1858-1919), was the youngest president, whereas the 40th president, Ronald Reagan (b 1911), was the oldest. Roosevelt became president when William McKinley (1843-1901) was assassinated. John F. Kennedy was the youngest president to be elected.

13. APPLICATION In order to provide better service, a customer service call center investigated the hang-up rates of people who called in for help on a recent evening. These data were collected.

Hang-Up Rates

Duration of call before hanging up (min)	0-3	3-6	6-9	9-12	12-15	15-18
Number of customers	1	3	4	6	13	9

Duration of call before hanging up (min)	18-21	21-24	24-27	27-30	30-33
Number of customers	8	10	6	4	1

a. Make a table showing the probability distribution of the random variable *x*, where *x* represents the duration of the call in three-minute intervals.

b. Construct a relative frequency histogram for the probability distribution.

c. What is the median length of time a customer waited before hanging up?

d. Draw a smooth curve on your histogram, so that the area under the curve is approximately the same as the area of the histogram.

Consumer CONNECTION

Customers often get frustrated when put on hold, and lost calls can mean lost business. Companies perform queue-time case studies to see how long a customer will wait on hold before abandoning a phone call. They can reschedule their staff to be available during high calling periods.

© 2004 Key Curriculum Press

Review

14. Consider the number of heads, x, when 15 fair coins are all tossed at once.

a. Use binomial expansion to find the probability distribution for $P(x)$. Then calculate the theoretical results for 500 trials of this experiment. Copy and complete the table below. Round off the frequencies in row 3 to whole-number values.

Heads (x)	0	1	2	...	14	15	Total
$P(x)$	(0.5^{15})						1
Frequency	$500\ (0.5^{15}) = \underline{?}$			...			500

b. Create a histogram showing the probability distribution for $P(x)$.

c. Find the mean and standard deviation of the number of heads.

d. How many of the 500 trials are within one standard deviation of the mean?

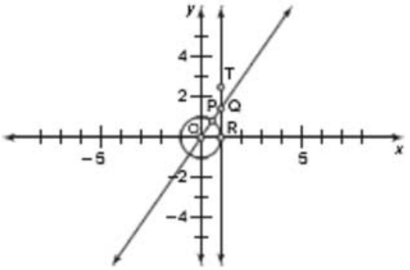
e. What percentage of the data is within one standard deviation of the mean?

f. What percentage of the data is within two standard deviations of the mean?

g. What percentage of the data is within three standard deviations of the mean?

15. Suppose you roll a pair of standard six-sided dice five times. What is the probability of rolling a sum of 8 at least three times?

16. The line RT is tangent to a unit circle, and $\overleftrightarrow{PO}$ intersects $\overleftrightarrow{RT}$ at Q. Point P makes one rotation around the unit circle every 20 seconds.



a. Find an equation to model the distance QR over a 1-minute period, starting when point P overlaps point R.

b. Graph the equation you found in 16a on your calculator.

17. ***Technology*** Use geometry software to construct the sketch in Exercise 16. Measure the distance QR, and animate point P. Describe the range of values of QR as P moves along the circle. Explain how these values relate to your answer to Exercise 16.

18. A spinner is divided into ten equal sectors numbered 1 through 10 in random order. If you get an even number, you add that number to your score. If you get an odd number, you subtract that number from your score. The game is over when your score reaches either + 50 or – 50. How many spins do you expect a typical game to last?

19. How is the area of a rectangle affected when its length is doubled and its width is halved? How is the area of a triangle affected when its base is doubled and its height is halved? How are the area of a rectangle and triangle affected when their horizontal dimensions are multiplied by 3 and their vertical dimensions are multiplied by one-third? Do you think this relationship is true for any two-dimensional figure?

Project

SIMPSON'S PARADOX

This table shows the number of male and female applicants, and the percentages admitted, to the six largest graduate school majors at the University of California, Berkeley, in Fall 1973.

	Men		Women		Total	
	Number of applicants	Percentage admitted	Number of applicants	Percentage admitted	Number of applicants	Percentage admitted
A	825	62	108	82	933	64
B	560	63	25	68	585	67
C	325	37	593	34	918	35
D	417	33	375	35	792	34
E	191	28	393	24	584	25
F	373	6	341	7	714	6

(D. Freedman, R. Pisani, R. Purves, and A. Adhikari (1991), *Statistics*, 2d ed., New York: Norton, p. 17)

Compare the percentages for men, women, and total admitted. Does it seem as though there is a bias in favor of men or women in admissions? Why or why not?

Now calculate the total number of men and women admitted to these six majors. (You'll need to use the data given for number of applicants and admission rate for each major.) Then calculate the overall percentages of men and of women admitted. Does it appear that there is a bias in favor of men or women? What happened?

Your project should include

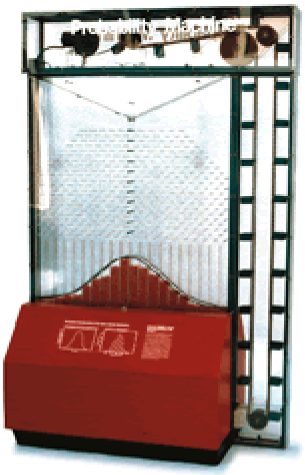
- Answers to the questions above.
- Any additional research you do on Simpson's paradox, including examples of other problems you find or one you make up yourself.

© 2004 Key Curriculum Press

LESSON 13.2

Normal Distributions

In Chapter 12, you studied the binomial distribution, $(p + q)^n$, for discrete random variables. The number of trials is represented by n, and p and q represent the only two possible outcomes of each event. In this lesson you will discover some properties of this probability distribution.



This 10-foot-tall machine drops balls through a grid of pins. The balls land in a bell-shaped curve-a visual representation of their probability distribution.

Investigation
The Bell

Consider the number of heads, x, when 15 fair coins are all tossed at once. The probability distribution, $P(x)$, is a binomial distribution, because there are exactly two possible outcomes for each toss-heads or tails.

Step 1 The binomial probability distribution for this theoretical experiment is $P(x) = {}_{15}C_x p^x(1 - p)^{15-x}$, where p is the probability of a head for each coin toss. Create a calculator table of this function with table entries at integer values of x from 0 to 15. What value should you use for p? What x-value gives the maximum for $P(x)$?

Step 2 Create two lists, L1 and L2. The entries in list L1 should contain all the possible values of x. Enter the corresponding values of $P(x)$ in list L2. You can do this quickly by defining L2 as the expression Y1 (L1). Complete the table below. What is the sum of values in list L2? Why does this answer make sense?

Heads (x)	0	1	2	...	15
***P* (x)**					

Step 3 Create a relative frequency histogram showing the distribution of heads. Use list L2 as the frequency. Describe the shape and range of this histogram. What is the maximum value?

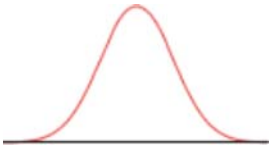
© 2004 Key Curriculum Press

Step 4 Graph $P(x)$ using a window with friendly domain, such as [0, 18.8, 1, –0.01, 0.25, 0.1]. You may choose to turn off the axes to see all the points. For what values of x is the function defined? Write a short description of this graph. Include the shape and your estimates of the mode, median, and mean of the distribution. How does this graph differ from the histogram in Step 3?

Step 5 Graph $P(x) = {}_{45}C_x p^x(1-p)^{45-x}$ using a window with friendly domain, such as [0, 47, 1, –0.01, 0.25, 0.1]. What theoretical experiment does this equation describe? Again, write a short description of the graph that includes the shape, the domain and range, and your estimate of the mode, median, and mean of the distribution. Compare this graph with the one in Step 4.

Step 6 Enter the defined values of x and $P(x)$ in lists L1 and L2. Then find the mean and standard deviation of the distribution. [▶ See Calculator Note 2B to learn how to find statistics of frequency tables. ◀]

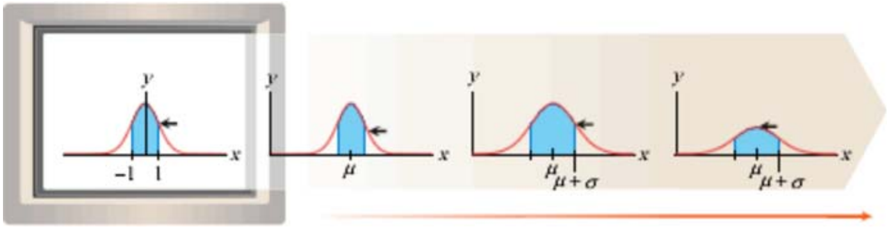
Step 7 If the number of coins increased, how would the answers to Steps 5 and 6 change? Write your predictions and then verify them.



As n grows increasingly large, the binomial distribution $(p + q)^n$ looks more and more like the continuous bell-shaped curve at right. Distributions for large populations often have this shape. Heights, clothing sizes, and test scores are a few examples. In fact, the bell-shaped curve is so common that it is called a **normal curve,** and a bell-shaped distribution is called a **normal distribution.**

Normal curves can describe the distribution of a sample or an entire population. You use $\bar{x}$ and s to represent the mean and standard deviation of a sample, but you use μ and σ (pronounced "mew" and "sigma") to represent the mean and standard deviation of an entire population.

In this lesson you'll see some properties of normal distributions. The general equation for a normal distribution curve is in the form $y = ab^{-x^2}$. If you graph a function like $y = 3^{-x^2}$, you'll get a bell-shaped curve that is symmetric about the vertical axis. To describe a particular distribution of data, you translate the curve horizontally to be centered at the mean of the data, and you stretch it horizontally to match the standard deviation of the data. Then you shrink it vertically so that the area is 1. These steps are shown graphically below. You'll want to begin with a parent function that has standard deviation 1.



History CONNECTION

German mathematician Carl Gauss (1777-1855) developed the idea of a normal distribution to describe the variation in the measurements made by surveyors and astronomers as they remeasured the same distances.

The parent function of a probability distribution has standard deviation 1, and is called the **standard normal distribution.** To meet the conditions for the standard normal distribution, statisticians have used advanced mathematics to determine the values of *a* and *b* in the equation $y = ab^{-x^2}$. The value of *a* is related to the number π.

$$a = \sqrt{\frac{1}{2\pi}} \approx 0.399$$

The value of *b* is related to another common mathematical constant, the transcendental number *e*.

$$b = \sqrt{e} \approx 1.649$$

Calculators allow you to work with these numbers fairly easily.

EXAMPLE A

The general equation for a normal curve is in the form $y = ab^{-x^2}$.

a. Write the equation for a standard normal curve, using $a = \sqrt{\frac{1}{2\pi}}$ and $b = \sqrt{e}$. Find a good graphing window for this equation and describe the graph. [► To learn how to enter the value of *e*, see Calculator Note 13A. ◄]

b. Write the equation for a normal distribution with mean μ and standard deviation σ.

c. Write the equation for the normal curve that fits the binomial distribution $P(x) = {}_{90}C_x p^x(1-p)^{90-x}$, where $p = .5$.

► Solution

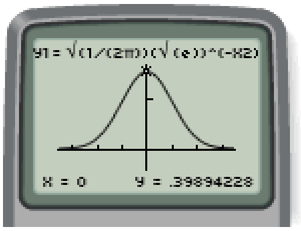
Substitute the values given for the constants *a* and *b* into the general equation for a normal curve.

a. The equation for a standard normal curve is

$$y = \sqrt{\frac{1}{2\pi}}\left(\sqrt{e}\right)^{-x^2}$$

Note that $\sqrt{\frac{1}{2\pi}}$ is the same as $\frac{1}{\sqrt{2\pi}}$.

A good window for this graph is [–3.5, 3.5, 1, –0.1, 0.5, 0.25]. The graph is bell-shaped and symmetric about $x = 0$. So the mean, median, and mode are all 0. Almost all of the data are in the interval $-3 \le x \le 3$.



[–3.5, 3.5, 1, –0.1, 0.5, 0.25]

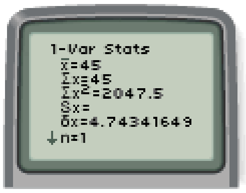
© 2004 Key Curriculum Press

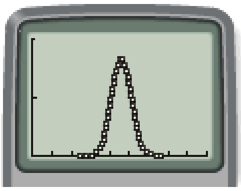
b. Translate the curve horizontally to shift the mean from 0 to μ. And stretch the curve horizontally to change the standard deviation from 1 to σ. The area under a probability distribution must be 1. A horizontal stretch will increase the area, so it must be accompanied by a vertical shrink.

$y = \sqrt{\frac{1}{2\pi}}\left(\sqrt{e}\right)^{-x^2}$	Start with the parent function.
$y = \sqrt{\frac{1}{2\pi}}\left(\sqrt{e}\right)^{-(x-\mu)^2}$	Substitute $(x - \mu)$ for x to translate the mean horizontally to μ.
$y = \frac{1}{\sqrt{2\pi}}\left(\sqrt{e}\right)^{-((x-\mu)/\sigma)^2}$	Divide $(x - \mu)$ by the horizontal scale factor, σ, so the curve reflects the correct standard deviation.
$y = \frac{1}{\sigma\sqrt{2\pi}}\left(\sqrt{e}\right)^{-((x-\mu)/\sigma)^2}$	Divide the right side of the equation by the vertical scale factor, σ, to keep the area under the curve equal to 1.

c. To write the equation for the normal curve that fits this distribution, you must first find the values of μ and σ. You can do this by entering values of x and $P(x)$ in lists L1 and L2, then using your calculator as shown below. For this distribution, $\mu = 45$ and $\sigma \approx 4.7434$, so the equation is

$$y = \frac{1}{4.7434\sqrt{2\pi}}\left(\sqrt{e}\right)^{-((x-45)/4.7434)^2} \approx (0.084)(1.649)^{-((x-45)/4.7434)^2}$$


```
1-Var Stats
 x̄=45
 Σx=45
 Σx²=2047.5
 Sx=
 σx=4.74341649
↓n=1
```


[0, 90, 10, 0, 0.1, 0.05]

You can see that the graph of this normal curve fits the discrete points of the binomial distribution graph.

The Normal Distribution

The equation for a **normal distribution** with mean μ and standard deviation σ is

$$y = \frac{1}{\sigma\sqrt{2\pi}}\left(\sqrt{e}\right)^{-((x-\mu)/\sigma)^2}$$

You can view the graph of a normal distribution well in the window

$$\mu - 3\sigma \le x \le \mu + 3\sigma \quad \text{and} \quad 0 \le y \le \frac{0.4}{\sigma}$$

This window will show 3 standard deviations above and below the mean, and the minimum and maximum frequencies of the distribution.

In the equation for a normal distribution, the data values are represented by x and their relative frequencies by y. The area under a section of the curve gives the probability that a data value will fall in that interval.

© 2004 Key Curriculum Press

Most graphing calculators provide the normal distribution equation as a built-in function, and you have to provide only the mean and standard deviation. In this chapter we will use the notation $n(x)$ to indicate a standard normal distribution function with mean 0 and standard deviation 1. Using this notation, a nonstandard normal distribution is written

n(x, mean, standard deviation)

For example, a normal distribution with mean 3.1 and standard deviation 0.14 is written $n(x, 3.1, 0.14)$. [► To learn how to graph a normal distribution on your calculator, see **Calculator Note 13B.** ◄] The area under a portion of a normal curve is written

N(lower, upper, mean, standard deviation)

This notation indicates the lower and upper endpoints of the interval, and the mean and standard deviation of the distribution. [► To learn how to find this value on your calculator, see **Calculator Note 13C.** ◄] This area determines the probability that a value in a normal distribution will fall within a particular range.

EXAMPLE B

A group of students weighs 500 U.S. pennies. They find that the pennies have normally distributed weights with a mean of 3.1 g and a standard deviation of 0.14 g.

a. Use your calculator to create a graph of this normal curve.

b. What is the probability that a randomly selected penny will weigh between 3.2 and 3.4 g?

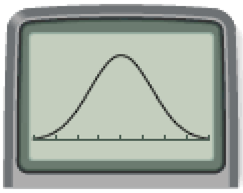
c. What is the probability that a randomly selected penny will weigh more than 3.3 g?

d. What is the probability that the weight of a penny will be within one standard deviation of the mean? Two standard deviations of the mean? Three standard deviations of the mean?

► Solution

The mean is 3.1, and the standard deviation is 0.14 g.

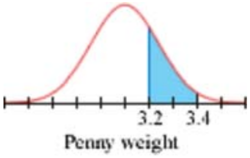
a. You can graph the probability curve using $n(x, 3.1, 0.14)$.



[2.7, 3.5, 0.1, – 0.5, 3, 0]

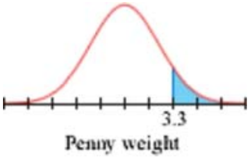
b. The probability that a randomly selected penny will weigh between 3.2 and 3.4 g is equal to the area under the normal curve between 3.2 and 3.4 g.

You can find this area on your calculator using $N(3.2, 3.4, 3.1, 0.14)$. The area is about .22, so there is a 22% chance that a randomly selected penny will have a mass between 3.2 and 3.4 g.



© 2004 Key Curriculum Press

c. You want to find the area under the curve to the right of 3.3 g. However, this interval has no upper limit. (Theoretically, there is no upper limit for the mass. Although a penny with a mass over 5 or 6 g would be extremely unlikely, it is not impossible.) How high should you set the upper bound? Whether you use 100 or 1000, you find the same answer, accurate to eight digits:



$N(3.3, 100, 3.1, 0.14) = .0765637714$

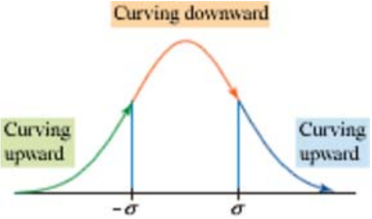
$N(3.3, 1000, 3.1, 0.14) = .0765637714$

So, you can use any fairly large number for the upper limit.

The probability that a penny will weigh more than 3.3g is approximately .07.

d. The probability that the mass will be within one standard deviation of the mean is $N(3.1 - 0.14, 3.1 + 0.14, 3.1, 0.14)$, or approximately .683. The probability that the mass will be within two standard deviations of the mean is $N(3.1 - 0.28, 3.1 + 0.28, 3.1, 0.14)$, or approximately .955. The probability that the mass will be within three standard deviations of the mean is $N(3.1 - 0.42, 3.1 + 0.42, 3.1, 0.14)$, or approximately .997.

Look at the curvature of a normal curve. At the points that are exactly one standard deviation from the mean, the curve changes between curving downward (the part of the curve with decreasing slope) and curving upward (the parts of the curve with increasing slope). These points are called **inflection points.** You can estimate the standard deviation of any normal distribution by locating the inflection points of its graph.



History CONNECTION

English nurse Florence Nightingale (1820-1910) contributed to the field of applied statistics by collecting and analyzing data during the Crimean War (1853-1856). While stationed in Turkey, she systematized data collection and record keeping at military hospitals and created a new type of graph, the polar-area diagram. Nightingale used statistics to show that improved sanitation in hospitals resulted in fewer deaths. For more information about Nightingale's contributions, see the web links at www.keymath.com/DAA.

Florence Nightingale

© 2004 Key Curriculum Press

EXERCISES

Practice Your Skills

1. The standard normal distribution equation, $y = ab^{-x^2}$, where $a = \sqrt{\frac{1}{2\pi}}$ and $b = \sqrt{e}$, is equivalent to the calculator's built-in function $n(x, 0, 1)$.

a. Use a table or graph to verify that these functions are equivalent.

b. Evaluate each function at $x = 1$.

2. From each equation, estimate the mean and standard deviation.

a. $y = \frac{1}{5\sqrt{2\pi}}\left(\sqrt{e}\right)^{-((x-47)/5)^2}$

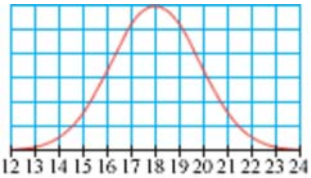
b. $y = \frac{0.4}{23}0.60653^{((x-250)/23)^2}$

c. $y = 1.29e^{-((x-5.5)^2/0.1922)}$

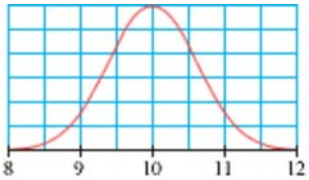
d. $y = 0.054(0.99091)^{(x-83)^2}$

3. From each graph, estimate the mean and standard deviation.

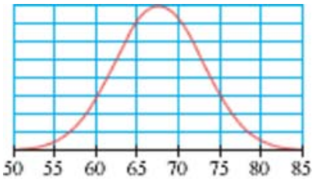
a.



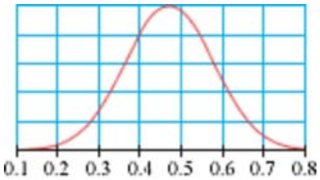
b.



c.



d.

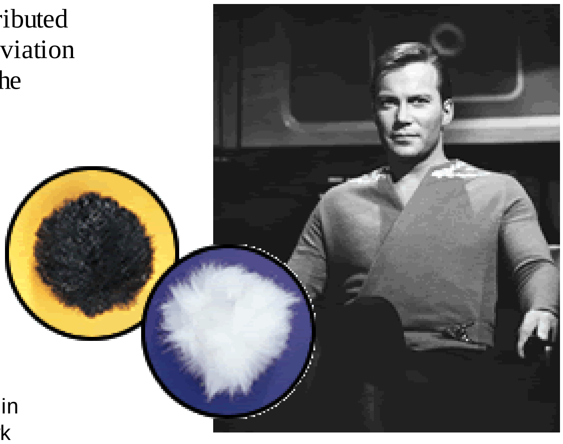


Keymath.com More Practice Your Skills

4. Estimate the equation of each graph in Exercise 3.

Reason and Apply

5. The life spans of wild tribbles are normally distributed with a mean value of 1.8 years and a standard deviation of 0.8 year. Sketch the normal curve, and shade the portion of the graph showing tribble life spans of 1.0 to 1.8 years.



William Shatner (b 1931) played Captain James T. Kirk in the television series *Star Trek* (1966-1969). Captain Kirk and two fictional alien creatures called tribbles are shown here.

© 2004 Key Curriculum Press

6. Assume that the mean height of an adult male gorilla is 5 ft 8 in., with a standard deviation of 7.2 in.

a. Sketch the graph of the normal distribution of gorilla heights.
b. Sketch the graph if, instead, the standard deviation were 4.3 in.
c. Shade the portion of each graph representing heights greater than 6 ft. Compare your sketches and explain your reasoning.

7. Frosted Sugar Squishies are packaged in boxes labeled "Net weight: 16 oz." The filling machine is set to put 16.8 oz in the box, with a standard deviation of 0.7 oz.

a. Sketch a graph of the normal distribution of package weights. Shade the portion of the graph representing boxes that are below the advertised weight.
b. What percentage of boxes does the shading represent? Is this acceptable? Why?

8. Makers of Sweet Sips 100% fruit drink have found that their filling machine will fill a bottle with a standard deviation of 0.75 oz. The control on the machine will change the mean value but will not affect the standard deviation.

a. Where should they set the mean so that 90% of the bottles have at least 12 oz of fruit drink in them?
b. If a fruit drink bottle can hold 13.5 oz before overflowing, what percentage of the bottles will overflow at the setting suggested in 8a?

9. The pH scale measures the acidity or alkalinity of a solution. Water samples from different locations and depths of a lake usually have normally distributed pH values. The mean of those pH values, plus or minus one standard deviation, is defined to be the pH range of the lake.

Lake Fishbegon has a pH range of 5.8 to 7.2. Sketch the normal curve, and shade those portions that are outside the pH range of the lake.

10. Data collected from 493 women are summarized in the table.

Height (cm)	148-50	150-52	152-54	154-56	156-58	158-60	160-62	162-64	164-66
Frequency	2	5	9	15	27	40	52	63	66

Height (cm)	166-68	168-70	170-72	172-74	174-76	176-78	178-80	180-82	182-84
Frequency	64	53	39	28	15	8	4	1	2

a. Find the mean and standard deviation of the heights, and sketch a histogram of the data.
b. Write an equation based on the model $y = ab^{-x^2}$ that approximates the histogram.
c. Find the equation for a normal curve using the height data.

11. The data at right show the pulse rates of 50 people.

a. Find the mean and standard deviation of the data and sketch a histogram.
b. Sketch a distribution that approximates the histogram.
c. Find the equation of a normal curve using the pulse-rate data.
d. Are these pulse rates normally distributed? Why or why not?

66	75	83	73	87	94	79	93	87	64
80	72	84	82	80	73	74	80	83	68
86	70	73	62	77	90	82	85	84	80
80	79	81	82	76	95	76	82	79	91
82	66	78	73	72	77	71	79	82	88

© 2004 Key Curriculum Press

12. Ridge counts in fingerprints are approximately normally distributed with a mean of about 150 and a standard deviation of about 50. Find the probability that a randomly chosen individual has a ridge count

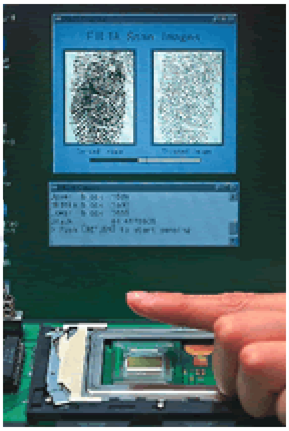
a. between 100 and 200

b. of more than 200

c. of less than 100

Science CONNECTION

Dactyloscopy is the comparison of fingerprints for identification. Francis Galton (1822-1911), an English anthropologist, demonstrated that fingerprints do not change over the course of an individual's lifetime, and that no two fingerprints are exactly the same. Even identical twins, triplets, and quadruplets have completely different prints. According to Galton's calculations, the odds of two individual fingerprints being the same are 1 in 64 billion. For more information on fingerprint identification, see the weblinks at www.keymath.com/DAA .



Developed by researchers in Atsugi, Japan, a microchip scans a finger to identify a fingerprint with 99% accuracy in half a second.

Review

13. Paul, Kenyatta, and Rosanna each took one national language exam. Paul took the French exam and scored 88. Kenyatta took the Spanish exam and scored 84. Rosanna took the Mandarin exam and scored 91. The national means and standard deviations for the tests are as follows:

French: $\mu = 72$, $\sigma = 8.5$
Spanish: $\mu = 72$, $\sigma = 5.8$
Mandarin: $\mu = 85$, $\sigma = 6.1$

a. Can you determine which test is most difficult? Why or why not?

b. Which test had the widest range of scores nationally? Explain your reasoning.

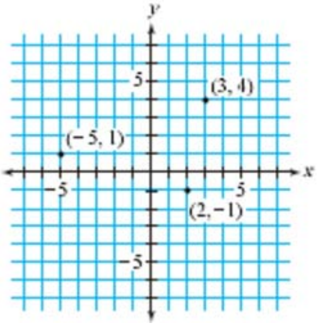
c. Which of the three friends did best when compared to the national norms? Explain your reasoning.

14. Mr. Hamilton gave his history class an exam in which a student must choose 3 out of 6 parts and complete 2 out of 4 questions in each part selected. How many different ways are there to complete the exam?

15. In the expansion of $(2x + y)^{12}$, what is the coefficient of the term containing y^7?

16. Find the equation of a conic section that passes through the three points given at right if the conic section is

a. A parabola. **b.** A circle.



EXPLORATION Fathom™

Normally Distributed Data

What kinds of data are normally distributed? In this exploration you'll use census data and Fathom Dynamic Statistics software to explore what attributes of the population of the United States are distributed normally.

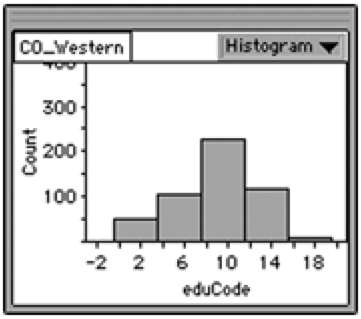
Activity

Is This Normal?

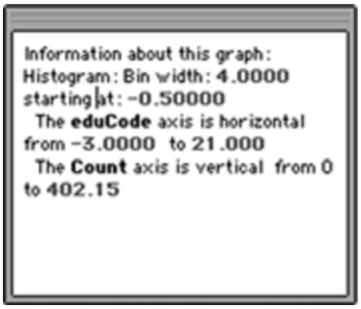
Step 1 Start Fathom. From the File menu choose **Open**. Open one of the census data files in the **Sample Documents** folder. You'll see a box of gold balls, called a collection, that holds data about a group of individuals.

Step 2 Click on the collection, and then choose **Case Table** from the Insert menu. Scroll through the table. What numerical attributes are included? Which ones do you think might be normally distributed?

Step 3 A histogram can show whether a set of data is approximately normally distributed. To create a histogram, choose **Graph** from the Insert menu. Drag and drop an attribute onto the horizontal axis, and then choose **Histogram** from the pull down menu in the corner of the graph window.

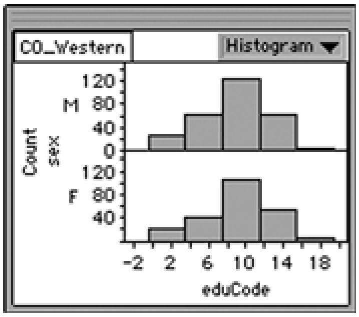


To experiment with different bin widths, double-click on any number on the horizontal axis. You will see a new window describing the bin width and axes scales. Click on any of the numbers in blue to modify them. What bin width makes it easiest to see patterns for each of the numerical attributes?



Information about this graph:
Histogram: Bin width: 4.0000
starting at: -0.50000
The **eduCode** axis is horizontal from -3.0000 to 21.000
The **Count** axis is vertical from 0 to 402.15

Step 4 Experiment with different attributes, bin widths, and regional data files to find data that are approximately normally distributed. You may want to filter your data. For example, income will not appear to be normally distributed because there are many people who have an income of $0. Many of these people are younger than 20 or older than 65. Choose **Add Filter** from the Data menu. Type "(age > 20) and (age < 65)". Click **Apply**. People younger than 20 and older than 65 will be removed from the case table and histogram. You may also want to separate histograms into categories, such as male and female. To do this, drag and drop a second attribute onto the vertical axis of the histogram.



What normally distributed data do you find? What data are not normally distributed that you thought might be? What regional differences are there?

Questions

1. Make conjectures about why specific attributes in the census data that you explored are or are not normally distributed.
2. For census data that are not normally distributed, what histogram shape is most common?

© 2004 Key Curriculum Press

LESSON 13.3

z-Values and Confidence Intervals

The way to do research is to attack the facts at the point of greatest astonishment.

CELIA GREEN

In Lesson 13.2, you learned about normally distributed populations and the probability that a randomly chosen item from such a population will fall in various intervals. That is, you knew something about the population and you saw how to find information about a sample.

In most real-life situations, however, you have statistics from one or more samples and want to estimate parameters of the population, which can be quite large. For example, suppose you know the mean height and standard deviation of 50 students that you survey, and you want to know the mean height and standard deviation of the entire population of students in your school. In this lesson you'll see how to describe some population parameters based on sample statistics.

First, it will be useful to learn a new method for describing a data value in a normal distribution. Knowing how a value relates to the mean value is important, but it does not tell you how typical the value is. For instance, to say that a penny's weight is 0.4 g less than the mean does not tell you whether this measurement is a rare event or a common event. But if you state how many standard deviations a value is from the mean, you have a much better idea of how unusual the value is.

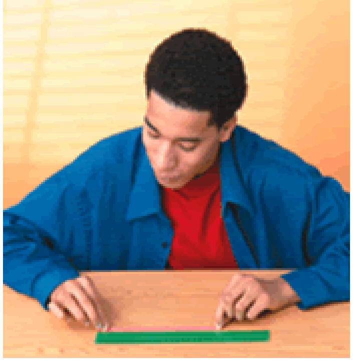
Chinese-American artist Diana Ong (b 1940) titled this watercolor *So Very Crowded.*

Investigation
Areas and Distributions

You will need

- a piece of rope
- a meterstick or tape measure

Any measurement of an object's length is an approximation of the actual length. Typically, the measurements made by several people will be normally distributed. You'll use this idea to explore areas under the normal curve.



Step 1 Measure a length of rope, accurate to 0.1 cm. Assume that your measurement is the mean of all measurements and the standard deviation is 0.8 cm. Sketch a normal curve based on your measurement.

© 2004 Key Curriculum Press

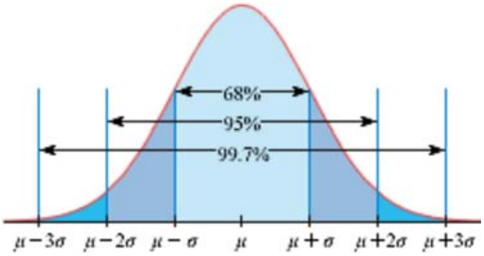
Step 2 Use the area under your normal curve to find the probability that a new measurement will be

a. within one standard deviation of your length. (That is, find the area between *your length* – 0.8 and *your length* + 0.8.)
b. within two standard deviations of your length.
c. within three standard deviations of your length.

Step 3 There is a rule in statistics known as the "68-95-99.7 rule." Compare your results from Step 2 with those of your group members, and write a rule that might go by this name.

To say that a penny's weight is 0.4 g less than the mean does not tell you whether or not this is a rare event. But to say that a penny's weight is 2.86 standard deviations from the mean does indicate that this measurement is a rare event.

In the investigation, you focused on adding and subtracting some number of standard deviations to or from the mean. The number of standard deviations that a normally distributed variable x is from the mean is called its **z-value.** In terms of z-values, the investigation asked for the probabilities that a new measurement would have a z-value between –1 and 1, between –2 and 2, and between –3 and 3. The 68-95-99.7 rule says that answers to these questions are about 68%, 95%, and 99.7%.



By the 68-95-99.7 rule, 68% of the area under a normal curve falls within one standard deviation of the mean, 95% within two standard deviations, and 99.7% within three standard deviations.

You can think of the z-value of x as the image of x under a transformation that translates and either shrinks or stretches the normal distribution to the standard normal distribution $n(x)$, with mean $\mu = 0$ and standard deviation $\sigma = 1$. This transformation from x-value to z-value is called **standardizing the variable** and can be calculated with the equation $z=\frac{x-\mu}{\sigma}$. The following example illustrates how to standardize values of a variable.

EXAMPLE A The heights of a large group of men are distributed normally with mean 70 in. and standard deviation 2.5 in.

a. Find the z-values for 67.5 in. and 72.5 in.

b. What is the probability that a randomly chosen member of this group has height x between 65 and 75 in.?
c. Find an interval of x-values, symmetric about the mean, that contains 90% of the heights.

© 2004 Key Curriculum Press

► Solution

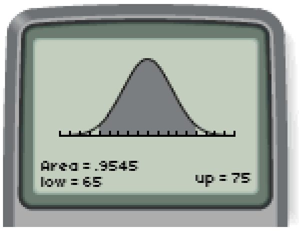
For this population, $\mu = 70$ and $\sigma = 2.5$.

a. Use the formula $z = \frac{x-\mu}{\sigma}$ to standardize the variable.

$$z = \frac{67.5-70}{2.5} = -1 \quad \text{and} \quad z = \frac{72.5-70}{2.5} = 1$$

In this distribution, 67.5 in. corresponds to a z-value of –1, which means the value is one standard deviation below the mean. The height 72.5 in. corresponds to a *z*-value of 1, or one standard deviation above the mean.

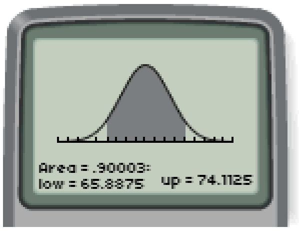
b. The *z*-value of 65 in. is $\frac{65-70}{2.5}$, or –2, and the *z*-value of 75 in. is $\frac{75-70}{2.5}$, or 2. There is a 95% probability that a randomly chosen value is within two standard deviations of the mean. A calculator graph confirms this prediction. [► See Calculator Note 13C to recall how to draw a normal curve with an area shaded. ◄]



```
Area = .9545
low = 65          up = 75
```

[61, 79, 1, –0.05, 0.2, 0.05]

c. You already know that 68% of the heights fall in the interval $-1 \leq z \leq 1$, which corresponds to $67.5 \leq x \leq 72.5$. You also know that 95% of the heights fall in the interval $65 \leq x \leq 75$. So, you can guess that an interval of about $66 \leq x \leq 74$ will contain 90% of the heights. Use your calculator and trial and error to obtain more precise endpoints. The calculator screen below shows that the interval $65.8875 \leq x \leq 74.1125$ contains 90.003% of the data.



```
Area = .90003
low = 65.8875     up = 74.1125
```

[61, 79, 1, -0.05, 0.2, 0.05]

Do *z*-values and the 68-95-99.7 rule help you learn about a population from a normally distributed sample taken from that population? Can you conclude, for example, that the probability is 68% that a rope's actual length-the mean of the population-is within one standard deviation of the sample mean? Not really. The population mean either is or isn't in this interval, so the probability that it is there is either 1 or 0; you just don't know which. But you can describe how confident you are that the population mean lies in a particular interval.

Clothing sizes are usually normally distributed.

© 2004 Key Curriculum Press

Confidence Interval

Suppose a sample from a normally distributed population has size n and mean $\bar{x}$, and the population standard deviation is σ. Then the p% **confidence interval** is an interval about $\bar{x}$ in which you can be p% confident that the population mean, μ, lies. If z is the number of standard deviations from the mean within which p% of normally distributed data lie, the p% confidence interval is

$$\bar{x} - \frac{z\sigma}{\sqrt{n}} < \mu < \bar{x} + \frac{z\sigma}{\sqrt{n}}$$

The confidence interval may also be expressed as $\mu = \bar{x} \pm \frac{z\sigma}{\sqrt{n}}$ or $\mu = \left(\bar{x} - \frac{z\sigma}{\sqrt{n}}, \bar{x} + \frac{z\sigma}{\sqrt{n}}\right)$.

In real-world situations, you may not know the population standard deviation. However, if the sample size is large enough, generally $n > 30$, you may use the sample standard deviation, s, in place of σ when calculating confidence intervals.

For example, suppose your class obtained a sample of 30 measurements of a rope's length, with mean 32.4 cm and standard deviation 0.8 cm. You can't say that the population mean is exactly 32.4 cm. But you can describe your confidence that the population mean (the rope's actual length) lies in an interval. For instance, knowing that 95% of normally distributed data have a z-value between –2 and 2, you can say that you're 95% confident that the population mean lies in the interval $\left(32.4 - \frac{2(0.8)}{\sqrt{30}}, 32.4 + \frac{2(0.8)}{\sqrt{30}}\right)$, or about (32.1, 32.7).

The 68-95-99.7 rule is useful in cases like these, but sometimes you want to be confident by a percentage other than 68%, 95%, or 99.7%. You can find the associated z-values by experimenting with a normal curve and trying to find an area symmetric to the mean that has the desired percentage of the total area. Here are the z-values associated with some other commonly used confidence intervals:

Confidence interval	90%	99%	99.9%
***z*-value**	1.645	2.576	3.291

In the next example, confidence intervals are calculated using z-values from this table.

EXAMPLE B

Jackson is training for the 100 m race. His coach timed his last run at 11.47 s. Experience in previous training sessions indicates that the standard deviation for timing this race is 0.28 s.

a. Find the 95% confidence interval.

b. What confidence interval corresponds to ±2.3 standard deviations?

c. Find the 90% confidence interval.

▸ Solution

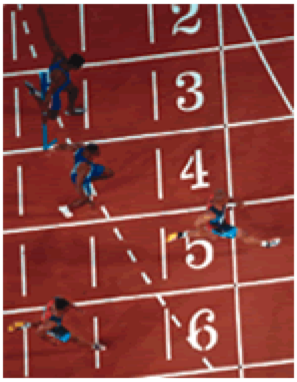
Because only one run time is known, the value for n is 1. A confidence interval is given as $\left(11.47 - \frac{z(0.28)}{\sqrt{1}}, 11.47 + \frac{z(0.28)}{\sqrt{1}}\right)$. This is the interval with endpoints $11.47 \pm 0.28z$.

© 2004 Key Curriculum Press

a. By the 68-95-99.7 rule, the *z*-value for 95% is about 2 standard deviations. The endpoints of the confidence interval, then, are $11.47 \pm 0.28(2)$. The coach is 95% confident that the actual time is between approximately 10.91 and 12.03 s.

b. By guess-and-check, the interval with endpoints $11.47 \pm 0.28(2.3)$, that is, or between 10.826 and 12.114 s, has a probability of $N(10.826, 12.114, 11.47, 0.28)$, or .9786. That is, a *z*-value of 2.3 standard deviations corresponds to about a 98% confidence interval.

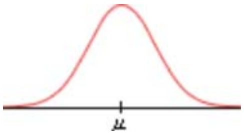
c. Using the table on page 748, the coach is 90% confident that the actual time is in the interval $11.47 \pm 0.28(1.645)$, or between 11.01 and 11.93 s.


Men race in the 100-meter final at the 2000 Olympic Games in Sydney, Australia.

In Example B, the coach had to rely on only one time measurement. If four people, instead of only one, had timed the run, and the mean of those times had been 11.47 s, then the 95% interval would have had endpoints $11.47 \pm \frac{2(0.28)}{\sqrt{4}}$, making it between 11.19 and 11.75 s. In general, the larger the sample size, the narrower the interval in which you can be confident that the population mean lies.

EXERCISES

Practice Your Skills



1. Trace the normal curve at right onto your paper. Add vertical lines demonstrating the 68-95-99.7 rule.

2. A set of normally distributed data has mean 63 and standard deviation 1.4. Find the *z*-value for each of these data values.

a. 64.4 **b.** 58.8 **c.** 65.2 **d.** 62

3. A set of normally distributed data has mean 125 and standard deviation 2.4. Find the data value for each of these *z*-values.

a. $z = -1$ **b.** $z = 2$ **c.** $z = 2.9$ **d.** $z = -0.5$

4. Complete these statements.

a. 95% of all data values in a normal distribution are within _?_ standard deviations of the mean.
b. 90% of all data values in a normal distribution are within _?_ standard deviations of the mean.
c. 99% of all data values in a normal distribution are within _?_ standard deviations of the mean.

Keymath.com More Practice Your Skills

© 2004 Key Curriculum Press

Reason and Apply

Created by American sculptor Richard Beyer (b 1925), *Waiting for the Interurban* is a cluster of statues at a bus stop in Seattle, Washington. Local residents frequently change the statues' clothes. In this photo, they are wearing Batman costumes.

5. The mean travel time between two bus stops is 58 min with standard deviation 4.5 min.

a. Find the z-value for a trip that takes 66.1 min.
b. Find the z-value for a trip that takes 55 min.
c. Find the probability that the bus trip takes between 55 and 66.1 min.

6. A set of normally distributed data has mean 47 s and standard deviation 0.6 s. Find the percentage of data within these intervals:

a. between 45 and 47 s
b. greater than 1.5 s above or below the mean

7. A sample has mean 3.1 and standard deviation 0.14. Find each confidence interval, and round values to the nearest 0.001. Assume $n = 30$ in each case.

a. 90% confidence interval **b.** 95% confidence interval **c.** 99% confidence interval

8. Repeat Exercise 7 assuming $n = 100$.

9. Make a statement about the change in size of each new confidence interval.

a. If you increase the size of your sample, then the confidence interval will $\underline{?}$.
b. If you increase your confidence from 90% to 99%, then the interval will $\underline{?}$.
c. If your sample has a larger mean, then the interval will $\underline{?}$.
d. If your sample has a larger standard deviation, then the interval will $\underline{?}$.

10. APPLICATION Fifty recent tests of an automobile's mileage indicate it averages 31 mi/gal with standard deviation 2.6 mi/gal. Assuming the distribution is normal, find the 95% confidence interval.

11. APPLICATION A commercial airline finds that, over the last 60 days, a mean of 207.5 ticketed passengers actually show up for a particular 7:24 A.M. flight. The standard deviation of their data is 12 passengers.

a. Assuming this distribution is normal, find the 95% confidence interval.
b. If the plane seats 225 passengers, what is the probability the plane will be overbooked?

12. APPLICATION The BB Manufacturing Company mass-produces ball bearings. The optimum diameter of a bearing is 45 mm, but records show that the diameters follow a normal distribution with mean 45 mm and standard deviation 0.05 mm. A diameter between 44.95 and 45.05 mm is acceptable.

a. What percentage of the output is acceptable?
b. What percentage of the output is unacceptable?
c. After retooling some of the equipment, the standard deviation is cut in half. What percentage of the output is now unacceptable?
d. If the engineers at the company want a 99.7% acceptability rate, what should their target standard deviation be?

© 2004 Key Curriculum Press

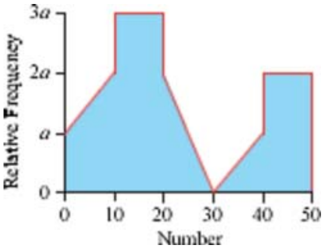
Consumer CONNECTION

Most manufacturers practice quality control. The strength of materials, sizes of parts, and reliability of the product are constantly monitored. In mass production and assembly lines, companies measure randomly chosen samples and compare the measurements to the desired mean and standard deviation. Using sampling, manufacturers can determine whether or not they are operating consistently, and they can determine when adjustments or improvements are necessary.

Quality control in plastics manufacturing

Review

13. A random-number generator selects a real number between 0 and 50, inclusive, according to the probability distribution at right. Find each value described.



a. *a*

b. *P*(a number is less than 30)

c. *P*(a number is between 20 and 40)

d. *P*(a number is 30)

e. *P*(a number is 15)

f. the median value

14. Five hundred integer values, –3 through 3, are randomly selected (and replaced) from a hat containing tens of thousands of integers. The frequency table at right lists the results. Find these values.

Value	Frequency
– 3	32
– 2	60
– 1	153
0	92
1	45
2	90
3	28

a. *P*(–3)

b. *P*(less than 0)

c. *P*(not 2)

d. the expected sum of the next ten selected values

15. You are about to sign a long-term rental agreement for an apartment. You are given two options:

Plan 1: Pay $400 the first month with a $4 increase each month.

Plan 2: Pay $75 the first month with a 2.5% increase each month.

a. Write a function to model the accumulated total you will pay over time for each plan.

b. Use your calculator to graph both functions on the same screen.

c. Which rental plan would you choose? Explain your reasoning.

LESSON 13.4

The Central Limit Theorem

You can use all the quantitative data you can get, but you still have to distrust it and use your own intelligence and judgment.

ALVIN TOFFLER

In the previous lesson, you saw how to use sample statistics to estimate some parameters of a normally distributed population. But what if you don't know whether or not the population is normally distributed? Can you still learn about population parameters from samples? In this lesson you'll explore this question.

Perhaps it will surprise you to learn that you can check how well a single sample predicts population parameters by imagining what would happen if you took lots of samples. For example, to determine the yield of a new variety of corn, a seed company runs a test with an experimental group of farmers. It collects a random sample of ears of corn from each farm and finds the mean number of kernels on the ears from each sample. These means help determine the mean and standard deviation of the entire population of corn, no matter how that population is skewed.

Investigation
Means of Samples

In this investigation you'll explore how the mean of a sample compares to the mean of a population.

Step 1 Each person in your group will create a population. Each group member's population should have a different type of distribution, as listed below. Make a histogram of your population to check that it is distributed appropriately. [► See Calculator Note 13D to create a list of 200 values, each between 20 and 50, for your population. ◄]

a. uniform **b.** normal **c.** skewed left **d.** skewed right

Step 2 Calculate the mean, μ, and standard deviation, σ, of your population.

Step 3 Devise a way to choose values randomly from your list. Select three values from your population, and calculate the mean of this small sample. Then select two more values, add them to the sample, and recalculate the mean. Add another two values and recalculate. How do these sample means compare to the actual mean?

© 2004 Key Curriculum Press

Step 4 Graph the equations $y = \mu$, $y = \mu - \frac{2\sigma}{\sqrt{x}}$, and $y = \mu + \frac{2\sigma}{\sqrt{x}}$, where μ and σ are the values of your population mean and standard deviation. Use the graphing window [0, 50, 5, $\mu - 2\sigma$, $\mu + 2\sigma$, 5]. These graphs will help you see a trend in your sample means from Step 3.

Step 5 Create a recursive routine that adds one randomly chosen value at a time to a sample from your population and plots the mean of the new sample. Plot each point in the form (*number sampled, mean*). [► See Calculator Note 13E for help with this routine. ◄] Make a rough sketch of what you see.

Step 6 Clear the graph, reset the counters *N* and *T* to 0, and repeat Step 5 three more times. What do you notice?

Step 7 Compare your results to those of your group members who used a different type of population distribution. In general, explain how the means of your samples compare to the mean of the entire population.

Your work on the investigation shows that, although the mean number of kernels per ear is different on each farm, the mean of a sample approximates the population mean, and the approximation is better for larger samples. In fact, given several different samples from a population, the sample means themselves are normally distributed, even if the population is not. Moreover, from the sample means, you can even predict the standard deviation of the population. These observations are summarized by the **Central Limit Theorem.**

The Central Limit Theorem

If several samples, each containing *n* data values, are taken from a population (with any distribution):

1. The means of the samples form a distribution that is approximately normal.
2. The population mean is approximately the mean of the distribution of sample means.
3. The standard deviation of the sample means is approximately the population's standard deviation divided by the square root of *n*, or $\frac{\sigma}{\sqrt{n}}$.

Each approximation is better for larger values of *n*.

In most real-life situations, you have only one sample rather than many. But you can still use information about the theoretical possibilities of many samples. For instance, many statements are made about "average" American teens-the amount of money they spend weekly, their consumption of various foods, the amount of time they spend on the Internet. These "averages" usually come not from asking every teen in the country but, instead, from sampling them. The Central Limit Theorem says that the sample mean is not much different from the mean of the population, no matter how skewed the population distribution is. If pollsters choose a large enough sample or several samples, they can confidently estimate the parameters of the entire population.

The Central Limit Theorem can also be applied to evaluate a claim about a population mean, as in the next example.

© 2004 Key Curriculum Press

Career CONNECTION

Demographers collect statistics to determine how the number of people in a location changes from year to year. They also study factors that cause population changes and use this information to predict future population trends. Having a detailed understanding of a population's development helps people working in government and public and private organizations to initiate policies on education, health, unemployment, and other community services.

A demographic study can determine what health issues are of concern in a particular community, allowing resources to be used in the most effective way.

EXAMPLE A

A pharmaceutical company claims that its antacid contains an average of 324 mg of its active ingredient in each tablet. Students in a chemistry class analyzed 25 tablets to determine the amount of active ingredient. Their results, in milligrams, are at right. If the company's claim is correct, what is the probability of getting these results?

314	338	330	328	319
326	307	319	313	315
335	351	308	333	316
318	317	306	300	321
294	325	314	317	335

▶ Solution

The mean of the students' sample is 319.96 mg. You want to know the probability that a sample will have a mean of 319.96 mg or less if the population mean is 324 mg.

The first part of the Central Limit Theorem says that, if you took many samples, their means would form a normal distribution. Therefore, if you knew the mean and standard deviation of that distribution of means, you could use properties of the normal distribution to determine the probability of a range of means, such as 319.96 mg or less.

The second part of the Central Limit Theorem says that the mean of the means is the same as the population mean, which the company says is 324 mg. And the third part of the Central Limit Theorem says that the standard deviation of the distribution of means is the population standard deviation divided by the square root of the sample size, or $\frac{\sigma}{\sqrt{n}}$. You don't know the standard deviation of the population, but you can assume it's the same as the standard deviation of the sample, which is 12.75 mg. Then the standard deviation of the distribution of means is $\frac{12.75}{\sqrt{25}}$, or 2.55 mg.

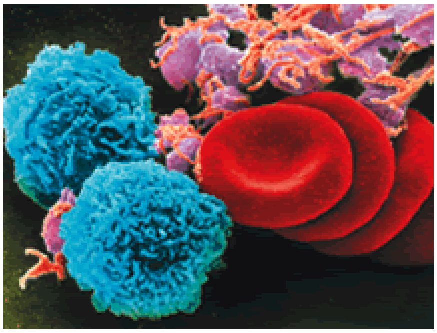
Putting all this together, you see that the probability of a mean of 319.96 mg or less is $N(0, 319.96, 324, 2.55)$, or .057. There is only a 5.7% chance that a sample's mean would be 319.96 mg or less. So, actually collecting a sample with a mean of 319.96 mg is a fairly rare event, and the company's claim of 324 mg is open to question.

© 2004 Key Curriculum Press

Example A illustrates **Inference.** Inference involves creating a hypothesis about one or more population parameters ("The mean is 324 mg"), deciding on what would make the hypothesis "improbable," collecting data, and either rejecting the hypothesis or letting it stand, based on probabilities. This **hypothesis testing** is the basis of research in many areas, including the medical field.

EXAMPLE B

A routine company medical exam performed on a worker suggests that a new product used for cleaning might be causing a reduction in the amount of white blood cells in the blood. How should the company proceed?



This magnified photo of blood cells uses color scanning to show white blood cells (blue), red blood cells (red), and platelets (pink). White blood cells function in the immune system, red blood cells carry oxygen through the body, and platelets help heal wounds.

Solution

The company can follow these steps to decide whether they should be concerned about the effect of the cleaning product.

Step 1

With statistics you cannot decide for sure whether the cleaning product has an effect on the number of white blood cells. You can decide only whether or not it "probably" has an effect. Actually, you want to decide whether or not a resulting sample statistic is "improbable." To do that, you state as your hypothesis that the cleaning product has *no* effect; this is called the **null hypothesis.** Then you decide what would make the hypothesis "improbable." Suppose you decide that if the mean number of white blood cells you get from sample data is less than 5% probable, then you'll reject the null hypothesis.

Step 2

You take a random sample of 36 workers and find they have a mean white blood cell count of 7075 cells per cubic millimeter (cells/mm^3). You look up medical information and find that a population of healthy adults has a mean white blood cell count of 7500 cells/mm^3 and a standard deviation of 1250 cells/mm^3.

Step 3

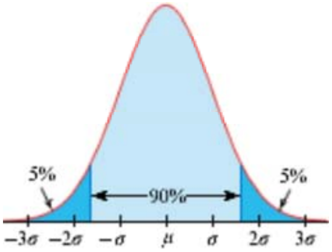
By the central limit theorem, the mean of many sample means would be 7500 cells/mm^3 with a standard deviation of $\frac{1250}{\sqrt{36}}$, or 208 cells/mm^3. The probability that a sample will have a mean of 7075 or less is therefore $N(0, 7075, 7500, 208)$, or .0205. This means that there is a 2% chance that, if the population mean were 7500 cells, a sample would have a mean of 7075 cells or less.

Step 4

Because the probability of the hypothesis is below your cutoff of 5%, the null hypothesis-that the cleaning product has no effect-is improbable and you reject it. The company should stop using the cleaning product and consider doing more conclusive testing.

© 2004 Key Curriculum Press

If you had decided in advance that you would reject the null hypothesis if the statistic were less than 1% probable, you couldn't reject the hypothesis here. Not being able to reject the null hypothesis doesn't mean the hypothesis is true, only that it's not what you decided to call false. Most hypothesis testing rejects null hypotheses if the sample statistic has a probability less than 5% or 10%. When the null hypothesis is rejected, statisticians call it a "significant event."



As you learned on page 748, 90% of the data in a normal curve falls within 1.645σ of the mean. So 5% is less than 1.645σ below the mean and 5% is more than 1.645σ above the mean.

In this chapter the problems and examples have noted that the data collected were from a random sample. A sample in which not only each person is equally likely but all groups of persons are equally likely is called a **simple random sample.** For instance, suppose a class consists of 20 girls and 10 boys. If you put everyone's name in a hat and randomly draw any 6 names, you have a simple random sample. In contrast, if you separate the names by gender and randomly select 4 girls and 2 boys, you do not have a simple random sample because not all groups of 6 persons are equally likely. For example, a group of 6 girls will never be selected.

The random samples throughout this book are actually simple random samples. In particular, the samples in the Central Limit Theorem must be simple random samples. The probability formulas you learned in Chapter 12 also depend on working with a simple random sample.

If you do not have a simple random sample, then you need to find different formulas that do not assume simple random sampling, accept the fact that your answer is approximate at best, or realize that you can draw no conclusion. You'll get the most benefit from starting with a simple random sample.

EXERCISES

Practice Your Skills

1. From a population with mean 200 and standard deviation 12, find the probability of

a. A value of 195 or less.

b. A sample of size $n = 4$ with mean 195 or less.

c. A sample of size $n = 9$ with mean 195 or less.

d. A sample of size $n = 36$ with mean 195 or less.

2. For each population and sample, determine what value of sample means would indicate a significant event two standard deviations from the population mean.

Keymath.com More Practice Your Skills

a. $\mu = 80$, $\sigma = 10$, sample size $n = 25$

b. $\mu = 130$, $\sigma = 6$, sample size $n = 36$

c. $\mu = 18$, $\sigma = 2$, sample size $n = 64$

d. $\mu = 0.52$, $\sigma = 0.1$, sample size $n = 100$

© 2004 Key Curriculum Press

Reason and Apply

3. Penny Adler has worked for many years as an actuary in the same office. By her calculations, it takes her an average of 23 min to get to work every day, with a standard deviation of 4.1 min. As she leaves her home one day, she notes that she must be at the office in 25 min. What is the probability that she will be late?

4. The "You Gotta Be Nutz" candy bar has mean weight 75.3 g and standard deviation 4.7 g. The manufacturer wants to avoid complaints that any single candy bar weighs far too little, so it decides to advertise a "minimum guaranteed weight."

a. What weight should the manufacturer advertise if they want 80% of the candy bars to meet or exceed the minimum weight? (Sketch a normal curve with shading, and write a complete sentence using your numerical answer.)

b. What weight should the manufacturer advertise if they want 90% of the candy bars to meet or exceed the minimum weight? (Sketch a normal curve with shading, and write a complete sentence using your numerical answer.)

c. What weight should the manufacturer advertise if they want 95% of the candy bars to meet or exceed the minimum weight? (Sketch a normal curve with shading, and write a complete sentence using your numerical answer.)

5. A random sample of Brand X medication has a mean of 230 mg of its active ingredient. The standard deviation of the sample is 12 mg. Using the sample standard deviation as the population standard deviation, make a prediction of the actual population mean with a 95% confidence interval given these sample sizes:

a. $n = 16$ **b.** $n = 100$ **c.** $n = 144$

6. A cookie company boasts an average of 20 chips per chocolate chip cookie. You doubt this claim, so you decide to test it. To start, you collect a random sample of 30 cookies and count the number of chips in each cookie. Your results are shown at right.

13	24	13	12	16	18
21	16	20	19	17	24
16	20	19	17	24	20
10	16	17	17	9	15
17	17	19	21	13	20

a. Make a mathematical statement of what you are trying to disprove. This is your null hypothesis.

b. Find the mean and standard deviation of your sample.

c. Use the standard deviation of the sample as a population parameter and find the probability that a population with mean 20 would produce a sample mean less than or equal to your statistic in 6b.

d. Use numbers and context to state a conclusion. If the probability from 6c is small (less than 5%), then state a rejection of the null hypothesis. If the probability is larger (more than 5%), then state that you fail to reject the hypothesis.

7. A biologist takes 300 water samples from a lake. He uses an indicator solution to find that 225 of the samples are in the pH range between 5.5 and 6.5. The mean pH is calculated to be 6.0. Estimate the standard deviation of the samples. Then sketch a graph of the pH distribution of the lake.

© 2004 Key Curriculum Press

8. The number of automobile accidents per week in a small city were recorded for the first half of the year. The data collected are:

$$\{4, 2, 0, 2, 3, 2, 0, 10, 3, 1, 2, 3, 1, 6, 1, 1, 2, 2, 3, 0, 1, 4, 0, 0, 9, 3\}$$

a. Calculate the mean and standard deviation of these data. Is the accident data normally distributed? If not, how is it skewed?

b. What are the mean, median, and mode of a normal distribution with the mean and standard deviation you found in 8a?

c. How do the mean, median, and mode of the accident data compare to your answers to 8b? Will the mean and median always be in this order for a distribution that is skewed right?

9. **APPLICATION** A real estate development corporation wants to demolish an abandoned factory and build condominiums in its place. Before proceeding with this plan, the corporation tests the site for toxic contaminants. Fifty core samples are randomly collected from various locations around the site and analyzed for cadmium (Cd). If the average concentration of cadmium in the soil is 0.8 mg/kg or higher, the site is declared contaminated and the developer will scrap its plans because the project will be too costly. The results of the sampling are shown in the table at right.

Concentration of Cd (mg/kg)					
0.89	0.4	0.56	0.29	0.51	0.30
0.53	0.71	0.79	0.61	0.77	0.36
0.24	0.62	0.55	0.63	0.40	0.43
0.63	0.89	0.47	0.33	0.32	0.22
0.71	0.57	0.50	0.98	0.66	0.72
0.43	0.54	0.65	0.95	0.11	
0.78	0.75	0.45	0.69	0.77	
0.62	0.88	0.36	0.71	0.58	
0.36	0.61	0.37	0.39	0.53	

a. How many samples are contaminated?

b. Find the mean and standard deviation of these data.

c. If the true concentration of cadmium in the soil is 0.8 mg/kg, what is the probability of these results?

d. Will the developer build the condominiums?

Environmental CONNECTION

Soil samples are often tested for toxic chemicals, but soil can also be tested for soil pH, nutrient or element breakdown, and particular physical characteristics. Many laboratories-public, private, and university-will carry out soil testing for farmers and home gardeners. Using the test results, farmers and gardeners can improve plant production by implementing a soil treatment plan.

An employee of an energy company records a core sampling test of the soil.

Review

10. Graph the equation $y = {}_{20}C_x\,(0.5)^x(0.5)^{(20-x)}$ on your calculator in a graphing window with friendly x-values, and turn off the axes. Describe this graph.

11. Graph the equation $y = {}_{20}C_x\,(0.3)^x(0.7)^{(20-x)}$ on your calculator in a graphing window with friendly x-values, and turn off the axes. Describe this graph.

12. Consider the linear function $f(x) = 16.8x + 405$.

a. Find $f(-19.5)$.

b. Find x such that $f(x) = 501.096$.

c. Write the equations of the two parallel lines that are 2.4 units above and below the line $y = f(x)$.

13. Find the median-median line for these data, and determine the root mean square error.

x	1	2	3.5	5	5.5	7	8.5	9.5	10
y	17	23	32	30	36	52	57	55	70

14. Every student in a large school measures the distance from the front door of the school to the flagpole using a meterstick. The results are normally distributed with mean 12.45 m and standard deviation 0.36 m.

a. Find the z-value that corresponds to a measurement of 12.00 m.

b. What is the probability that a randomly chosen measurement is between 12.30 and 12.60 m?

c. Nine students, who were absent on measurement day, measured the distance the next day. What is the probability that the mean of their measurements is between 12.30 and 12.60 m?

d. Find the 95% confidence interval for a single measure.

e. Find the 95% confidence interval for the mean of nine measurements.

IMPROVING YOUR VISUAL THINKING SKILLS

Acorns

Acorns fall around the base of a particular tree in an approximately normal distribution with standard deviation 20 ft away from the base.

1. What is the probability that any 1 acorn will land more than 10 ft from the tree?

2. What is the probability that any 4 randomly chosen acorns will land an average of more than 10 ft from the tree?

3. What is the probability that any 16 randomly chosen acorns will land an average of more than 10 ft from the tree?

© 2004 Key Curriculum Press

EXPLORATION

Confidence Intervals for Binary Data

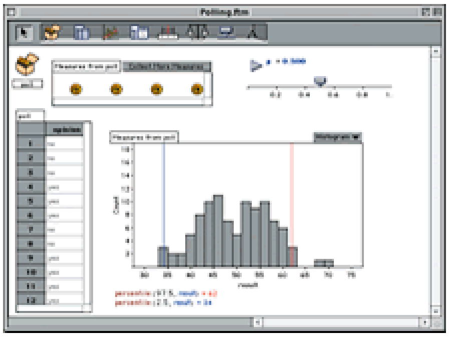
Suppose you take a poll of 50 randomly chosen voters in your state and find that 53% would vote "yes" on a proposition that would increase taxes in order to support schools. Is it possible that the proposition will fail? How likely is it that the proposition will pass? (If a proposition is to pass, at least 50% of the population must vote "yes.") What if your poll shows that 78% of your sample supports the proposition? Then how certain can you be that the proposition will pass?

You learned about confidence intervals in Lesson 13.4, and explored continuous variable data, such as age and height. In this voting situation, however, the data are *binary*. That is, there are two possible values for each data value, "yes" or "no." In this exploration you'll use Fathom to simulate the results of 100 random polls of 50 voters and make hypotheses about what the results of a sample poll allow you to conclude about a population's opinion.

Activity

Polling Voters

Step 1 Start Fathom. From the File menu, choose **Open,** and open the file **Polling.ftm.** You will see a window similar to the one shown below.



There are two collections: **poll,** which contains 50 opinions (either "yes" or "no"), and **Measures from poll,** which contains the results (the percentage who said "yes") from each of 100 polls of a random sample of 50 voters. The case table on the left shows the results of one poll, and the histogram shows the distribution of the percentage results from the 100 polls. The histogram also shows the 2.5th percentile and the 97.5th percentile of the 100 results, so 95% of the results lie between the blue and red lines-the equivalent of the 95% confidence interval.

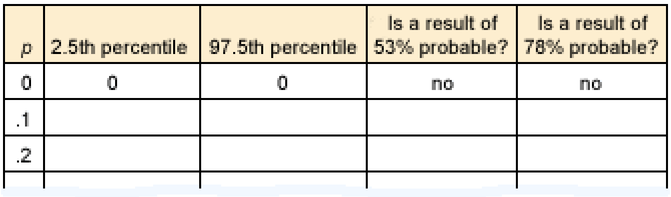
© 2004 Key Curriculum Press

The slider, *p*, controls the true percentage of voters in the population who support the proposal.

To use the simulation, change the value of *p* by using the slider or typing in a value. Then click **Collect More Measures** to simulate conducting 100 polls of 50 randomly chosen voters. You should see the case table, the histogram, and the percentile values change.

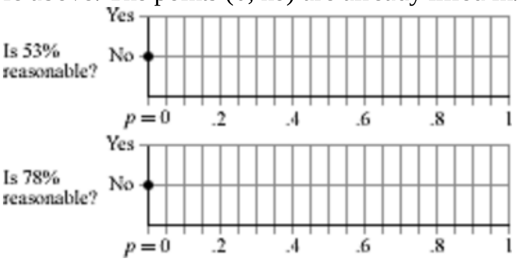
Step 2 Look at the Fathom window on page 760, which shows results for a population with *p* =.5. Is a polling result of 53% probable if exactly 50% of the population support the proposal? Would a result of 78% be probable? (For this exploration, consider a result probable if it falls within the 95% confidence interval.)

Step 3 Work with a partner to estimate the 2.5th and 97.5th percentile values for *p* = 0, .1, .2, and so on, to 1. For each value of *p*, set the slider, collect a new set of measures, and enter the percentile values into a table like the one below. Then enter “yes” or “no” in the third and fourth columns, depending on whether or not results of 53% and 78% fall within the 95% confidence interval.



p	2.5th percentile	97.5th percentile	Is a result of 53% probable?	Is a result of 78% probable?
0	0	0	no	no
.1				
.2				

Step 4 Copy the graphs below for results of 53% and 78%. Make one point for each row in the table above. The points (0, no) are already filled in.



Step 5 Analyze your graphs from Step 4 and figure out, as accurately as possible, where the points change from “no” to “yes” and back again. Explore *p*-values between those in your table, such as .35, if necessary. Plot any additional points on your graphs. For which values of *p* are poll results of 53% and 78% probable?

Step 6 You’ve now examined what kind of poll results are probable, given that you know the actual percentage of the population that will vote “yes.” Use your observations to make a reverse hypothesis. That is, is a proposition likely to pass if a poll shows 53% support? 78% support? Explain your reasoning.

© 2004 Key Curriculum Press

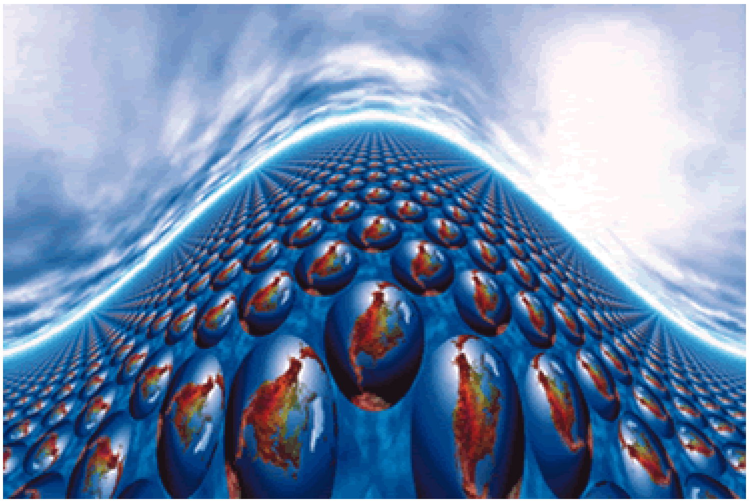
Questions

1. If you use **Collect More Measures** several times, while holding *p* constant, the values of the 2.5 and 97.5 percentiles change slightly. Why does this happen? How might you determine more precisely the 2.5 and 97.5 percentiles?

2. For binary data, the formula that determines a 95% confidence interval is

$$CI = \hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where $\hat{p}$ is the sample percentage and *n* is the number in the sample. Calculate the confidence interval for sample results of 53% and 78% from a survey of 50 voters. How do these calculations compare to your experimental results in Step 5?



© 2004 Key Curriculum Press

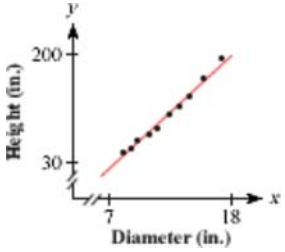
LESSON 13.5

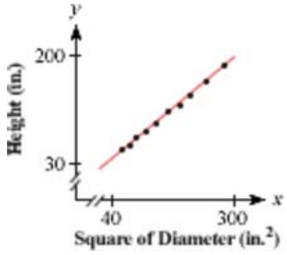
Bivariate Data and Correlation

It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.

SIR ARTHUR CONAN DOYLE

Dr. Aviles and Dr. Scott collected data on tree diameters and heights. Dr. Aviles thought that the height was closely associated with the diameter, but Dr. Scott claimed that the height was more closely associated with the square of the diameter. Which model is better? Each researcher plotted data and found a good line of fit. Their graphs are shown here.





Does it appear that a line is the appropriate model? If so, is there a better linear relationship in Dr. Aviles's data or in Dr. Scott's data? How good is the fit for each line? In this lesson you'll learn how to answer questions like these.

In Lesson 13.4, you learned how to make predictions about population parameters from sample statistics. You can also predict associations between parameters, such as height and diameter or height and the squares of diameter, for a large population, such as all trees. You can even apply this method to populations that are infinitely large. The process of collecting data on two possibly related variables is called **bivariate sampling.** How can you measure the strength of the association in the sample?

A commonly used statistical measure of linear association is called the **correlation coefficient.** A linear association between two variables is called **correlation.** In the investigation you'll use a calculator to explore properties of the correlation coefficient for a bivariate sample.

A giant Sequoia tree in Yosemite National Park, California.

© 2004 Key Curriculum Press

Investigation

Looking for Connections

Step 1 Work with your group to create a survey with five questions that are all answered with a number, or use the sample survey here.

1. How many minutes of homework did you do last night?
2. How many minutes did you spend talking, calling, e-mailing, or writing to friends?
3. How many minutes did you spend just watching TV or listening to music?
4. At what time did you go to bed?
5. How many academic classes do you have?

Step 2 Conjecture with your group about the strengths of correlations between pairs of variables. For example, you may decide that the number of minutes of homework is strongly correlated with the number of academic classes. Consider each of the ten pairs of variables and identify which combinations you believe will have

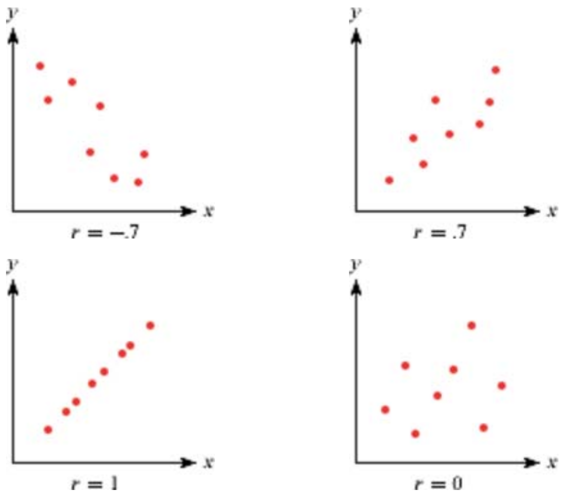
- **i.** A positive correlation (as one increases, the other tends to increase).
- **ii.** A negative correlation (as one increases, the other tends to decrease).
- **iii.** A weak correlation.

Step 3 Gather data from each student in your class. Then enter the data into five calculator lists. Plot points for each pair of lists, and find the correlation coefficients. [► See Calculator Note 13F to learn how to find the correlation coefficient. ◄] You may want to divide this work among members of your group. Describe the relationship between the appearance of the graph and the value of the correlation coefficient.

Step 4 Write a paragraph describing the correlations you discover. Include any pairs that are not correlated that you find surprising. You have collected a small and not very random sample; do you think these relationships would still be present if you collected answers from a random sample of your entire school population?

In 1896, English mathematician Karl Pearson (1857–1936) proposed the correlation coefficient, now abbreviated *r*. To compute the correlation coefficient, Pearson replaced each *x*- and *y*-value in a data set with its corresponding *z*-value. If a particular *x*- or *y*-value is larger than the mean value for that variable, then its *z*-value is positive. And if a particular *x*- or *y*-value is smaller than the mean value for that variable, then its *z*-value is negative. Pearson then found the product of z_x and z_y for each data point, and summed these products. In a data set that is generally increasing, the products of z_x and z_y are positive. This is because for every point, either both *x* and *y* are above the mean, or both *x* and *y* are below the mean. In a data set that is generally decreasing, usually either z_x is positive and z_y is negative, or z_x is negative and z_y is positive. Therefore, the products will be negative. After summing the products, Pearson divided by $(n - 1)$ to get a number between –1 and 1. So, he defined the correlation coefficient as $\frac{\sum z_x z_y}{n-1}$. But what do values of this coefficient mean?

You may have noticed in the investigation that values of r can range from –1 to 1. A value of 1 means the x-values are positively correlated with the y-values in the strongest possible way. That is, as x-values increase, y-values increase proportionally. A value of –1 means the x-values are negatively correlated with the y-values in the strongest possible way. That is, as x-values increase, y-values decrease proportionally. A value of 0 means there’s no linear correlation between the values of x and y.



If data are highly correlated, a straight line will model the data points well.

For Dr. Aviles’s data in the tree example at the beginning of this lesson, the correlation coefficient is .992. So, his data are very close to linear. For Dr. Scott’s data, the correlation coefficient is .999. That’s even better. This means that for the trees measured, the squares of diameters are better predictors of the heights than are the diameters themselves.

The Correlation Coefficient

The correlation coefficient, r, can be calculated with the formula

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

A value of r close to ± 1 indicates a strong correlation, whereas a value of r close to 0 indicates no correlation.

Note that the definition of the correlation coefficient includes no reference to any particular line, though it describes how well a line fits a bivariate data set. In contrast, the root mean square error you studied in Chapter 3 describes how well a *particular* line fits a data set.

Often, bivariate data are collected from a study or an experiment in which one variable represents some condition and the other represents measurements based on that condition, as shown in the next example. In statistics, the x- and y-variables are often called the **explanatory** and **response** variables instead of the independent and dependent variables.

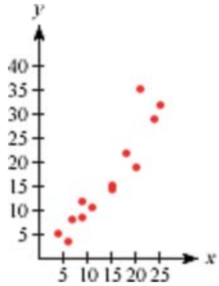
EXAMPLE A

Kiane belongs to many committees and notices that different groups take different amounts of time to make decisions. She wonders if the time it takes to make a decision is linearly related to the size of the committee. So, she collects some data. Find the correlation coefficient of this data set and interpret your result.

Size (people)	4	6	7	9	9	11	15	15	18	20	21	24	25
Time (min)	5.2	3.8	8.2	8.5	12.0	10.8	14.7	15.5	22.0	19.1	35.3	29.2	32.1

Solution

In this instance, it makes sense to let the explanatory variable, x, represent the committee size and the response variable, y, represent the time it takes to make a decision. When plotted, the data show an approximately linear pattern.



You can find the value of r directly using a calculator, but you should also try to use the formula—assisted by a calculator, of course. First, enter the data into lists and verify these statistics:

$\bar{x} \approx 14.15$ people, $s_x \approx 7.0456$ people
$\bar{y} \approx 16.65$ minutes, $s_y \approx 10.302$ minutes

Then use the lists in the formula:

$$\frac{\sum(x-14.15)(y-16.65)}{(7.0456)(10.302)(13-1)} \approx .9380$$

A correlation coefficient of $r \approx .9380$ means there is a strong positive correlation between the size of a committee and the time it takes to reach a decision. This means that as the size of a committee increases, the time it takes to reach a decision increases proportionately.

Be careful that you don't confuse the ideas of correlation and causation. A strong correlation may exist between two sets of data, but this does not necessarily imply a causal relationship. For instance, in Example A, Kiane found a strong correlation between committee size and decision-making time. But this does not necessarily mean that the size of the committee *caused* the decision to take longer. Whether it did or did not can be proved only by a carefully controlled experiment.

EXAMPLE B

The director of a summer camp has collected data for two weeks on both daily ice cream sales from the camp store and visits to the camp nurse for treatment of sunburn. What conclusions, if any, can you make?

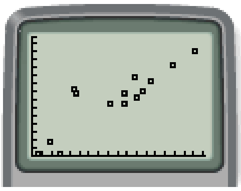
Ice cream sales	\$245.10	\$45.25	\$17.85	\$205.00	\$276.35	\$428.25	\$312.15
Visits to nurse	66	17	1	65	72	131	93

Ice cream sales	\$288.25	\$267.95	\$74.10	\$111.50	\$371.55	\$244.45	\$115.75
Visits to nurse	81	99	2	84	113	78	79

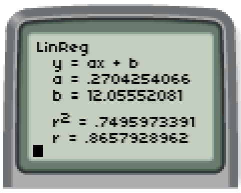
© 2004 Key Curriculum Press

Solution

Graph the data and calculate the correlation coefficient. The graph of the data shows a clear upward trend. The correlation coefficient of .866 indicates a fairly strong correlation.



[0, 450, 25, 0, 150, 10]



```
LinReg
 y=ax+b
 a=.2704254066
 b=12.05552081
 r²=.7495973391
 r=.8657928962
```

So, can you conclude that buying ice cream causes a sunburn? Or does getting a sunburn cause ice cream buying? There is most likely another variable causing both of these effects. Perhaps the daily temperature might be a **lurking variable** behind both of these results.

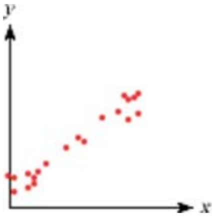
So, you can conclude that sunburn and ice cream sales are correlated, but not that one of these occurrences causes the other.

EXERCISES

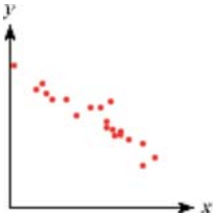
Practice Your Skills

1. Approximate the correlation coefficient for each data set.

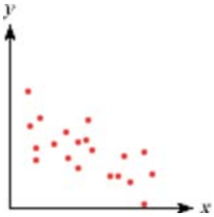
a.



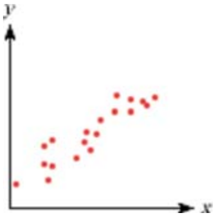
b.



c.



d.



© 2004 Key Curriculum Press

2. Sketch a graph of a data set with approximately these correlation coefficients.

a. $r = -.8$ **b.** $r = -.4$

c. $r = .4$ **d.** $r = .8$

3. Copy and complete the table at right. Then answer 3a–d to calculate the correlation coefficient.

x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$
12	8			
14	8			
18	6			
19	5			
22	3			

a. What is the sum of the values for $(x - \bar{x})(y - \bar{y})$?

b. What are $\bar{x}$ and s_x?

c. What are $\bar{y}$ and s_y?

d. Calculate $r = \dfrac{\sum(x - \bar{x})(y - \bar{y})}{s_x s_y(n - 1)}$.

e. What does this value of r tell you about the data?

f. Draw a scatter plot to confirm your conclusion in 3e.

4. In each study described, identify the explanatory and response variables.

a. Doctors measured how well students learned finger-tapping patterns after various amounts of sleep.

b. Scientists investigated the relationship between the weight of a mammal and the weight of its brain.

c. A university mathematics department collected data on the number of students enrolled each year in the school and the number of students who signed up for a basic algebra class.

5. For each research finding, decide whether there is evidence of causation, correlation, or both. If it is only a correlation, name a possible lurking variable that may be the cause of the results.

a. As the sales of television sets has increased, so has the number of overweight adults. Does television cause weight gain?

Keymath.com More Practice Your Skills

b. A study in an elementary school found that children with larger shoe sizes were better readers than those with smaller shoe sizes. Do big feet make children read better?

c. The more firefighters sent to a fire, the longer it takes to put out the fire. Does sending more firefighters cause a fire to burn longer?

Reason and Apply

6. An environmental science class conducted a research project to determine whether there was a relationship between the soil pH, x, and the percent dieback of new growth, y, for a particular type of tree. The table below contains the data the class collected.

x	3.3	3.4	3.5	3.6	3.7	3.8	3.9	4.0	4.1	4.2	4.3	4.5	5.0	5.1
y	7.3	10.8	10.4	9.3	12.4	11.2	6.6	10.0	9.2	12.4	2.3	4.3	1.6	1.0

a. Make a scatter plot of the data.

b. Describe the relationship between the two variables.

© 2004 Key Curriculum Press

c. Find the correlation coefficient. Does this confirm or refute your answer to 6b? How can you tell?

d. Can you conclude that higher soil pH causes less dieback of new growth?

Environmental CONNECTION

Trees take in carbon dioxide from the air which helps offset carbon dioxide emissions caused by cars and industry. Destruction of forests has contributed to an increase in the amount of carbon dioxide in the atmosphere, leading to global warming and other problems. Although nature can eventually reforest most areas on its own, we can speed the process by improving soil quality and selecting the best species for a particular area.

Deforestation from logging and pasture clearing over the past 30 years have reduced the size of the Amazon rain forest in Brazil by 15 percent.

7. The table contains information from a selection of four-year colleges and universities for the 2000–2001 school year. Describe the correlation between the number of students and the number of faculty.

Four-Year Colleges

College	Number of students	Number of faculty
Alfred University	2,433	208
Brandeis University	4,753	461
Brown University	7,723	737
Bryn Mawr College	1,784	167
California College of Arts & Crafts	1,213	326
Carleton College	1,936	214
College of William & Mary	7,530	718
DePauw University	2,225	234
Drake University	5,126	293
Duquesne University	9,667	847
Gallaudet University	1,661	221
Hampshire College	1,172	114
Illinois Wesleyan University	2,102	183
Lehigh University	6,509	457
Maryland Institute, College of Art	1,302	220
Miami University of Ohio	16,290	975
Mills College	1,070	154
Morehouse College	2,970	235
Mt. Holyoke College	2,089	228
Princeton University	6,547	914
Rhode Island School of Design	2,086	399
Rhodes College	1,554	156
Saint John’s University	2,020	190
St. Olaf College	3,014	315
Spelman College	1,897	147
Swarthmore College	1,428	201
Syracuse University	14,478	1,395
Tufts University	8,933	1,097
Tuskegee University	2,826	262
University of Tulsa	4,158	414
Wesleyan University	3,158	308
Wheaton College	2,827	244

(*The World Almanac and Book of Facts 2002*)

8. These data list the number of bound volumes, the annual circulation (the number of books that are checked out), and the annual operating cost for randomly selected public libraries in 2002. Each number in the table is in thousands.

Public Libraries in Selected Cities

City	Volumes	Circulation	Cost ($)
Atlanta, GA	1,960	2,704	19,500
Baton Rouge, LA	1,232	2,449	13,000
Boston, MA	6,582	3,500	30,100
Buffalo, NY	5,241	8,998	29,000
Cleveland, OH	5,380	3,782	48,600
Dallas, TX	3,000	3,807	19,200
Denver, CO	4,224	7,448	23,700
Detroit, MI	2,808	1,513	27,400
Kansas City, MO	2,215	2,224	14,800
Memphis, TN	1,849	3,702	14,700
Omaha, NE	917	2,474	9,090
Philadelphia, PA	7,892	6,067	46,800
Seattle,WA	1,777	4,580	24,500
St. Paul, MN	1,051	2,401	7,800

(*www.infoplease.com*)

Library: Homage to Marcel Proust is a mixed media installation box constructed by French artist Charles Matton (b 1933). The miniature work (15.8 in. wide) evokes the world of French writer Marcel Proust (1871–1922).

a. What is the correlation coefficient for number of volumes and operating cost, and for circulation and operating cost?

b. Is number of volumes or circulation more strongly correlated with operating cost? Explain your reasoning.

9. ***Mini-Investigation*** For each data set given in 9a–c, draw a scatter plot and find the correlation coefficient, *r*. State what this value of *r* implies about the data, and note any surprising results you find.

a. {(0.5, 1), (0.6, 0.9), (0.7, 0.8), (0.8, 0.7), (0.9, 0.6)}

b. {(0.5, 1), (0.6, 0.9), (0.7, 0.8), (0.8, 0.7), (0.9, 0.6), (1.9, 1.9)}

c. {(0.5, 1), (0.6, 0.9), (0.7, 0.8), (0.8, 0.7), (0.9, 0.6), (1.9, 0.9)}

d. Based on your answers to 9a–c, do you think the correlation coefficient is strongly affected by outliers?

10. ***Mini-Investigation*** For each data set given in 10a and b, draw a scatter plot and find the correlation coefficient, *r*. State what this value of *r* implies about the data.

a. {(0.3, 0.9), (0.4, 0.6), (0.6, 0.4), (1, 0.3), (1.4, 0.4), (1.6, 0.6), (1.7, 0.9), (0.3, 1.1), (0.4, 1.4), (0.6, 1.6), (1, 1.7), (1.4, 1.6), (1.6, 1.4), (1.7, 1.1)}

b. {(0.4, 1.4), (0.6, 1.1), (0.8, 0.8), (1, 0.5), (1.2, 0.8), (1.4, 1.1), (1.6, 1.4)}

c. For the data sets in 10a and b, does there appear to be a relationship between *x*- and *y*-values? Is this reflected in the values you found for *r*?

Review

11. What is the slope of the line that passes through the point (4, 7) and is parallel to the line $y = 12(x - 5) + 21$?

12. The data at right give the reaction times of ten people who were administered different dosages of a drug. Find the median-median line for these data and determine the root mean square error.

Dosage (mg)	Reaction time (s)
85	0.5
89	0.6
90	0.2
95	1.2
95	1.6
103	0.6
107	1.0
110	1.8
111	1.0
115	1.5

13. Find an equation of the line passing through (4, 0) and (6, – 3).

14. Graph the equation $y = \log_5 x$.

15. On a car trip, your speed averages 50 km/h as you drive to your destination. You return by the same route and average 75 km/h. What's your average speed for the entire trip?

16. David wants to send his nephew a new 5-foot fishing pole. David wraps up the pole and takes it to Bob's Courier Service. But Bob's has a policy of not accepting any parcels longer than 4 feet. David returns an hour later with the fishing pole wrapped and sends it with no problem. How did he do it? (Assume the pole is not broken into pieces.)

Project

CORRELATION VS. CAUSATION

Think of a relationship someone claims involves causation, but you think might only involve correlation. Your claim can be about anything—science, popular beliefs, sociology—but it must be something that can be tested. First, research data related to the claim and determine whether or not the data seem to show a correlation between the two variables. Then, think about whether or not one event really causes the other. What other factors might be involved? Might the data you found be misleading in some way? If you can, find the data for any other factors and see how these data are related to your claim. Write a report on your findings.

Your project should include

- The claim you researched and the data and analyses you found.
- Any graphs, tables, or equations that you used while analyzing the data.
- A summary of other factors that might be involved.
- Your own conclusion on the relationship of the data.

Fathom™

With Fathom Dynamic Statistics you can plot data related to different pairs of variables. You can also compare the fit of different equations through your data points.

LESSON 13.6

The Least Squares Line

In Lesson 13.5, you saw how the correlation coefficient could be used to determine how closely two variables in a sample are linearly related. In this lesson you'll learn how to use the correlation coefficient of a sample to determine a line of fit, from which you can make predictions about the population.

History CONNECTION

In the late 1800s, English anthropologist Francis Galton studied correlations among various measurements, including heights of fathers and sons. He found that sons' heights tended to be closer to the mean height for men than their fathers' heights were. This phenomenon is known today as "regression toward the mean." The term **regression analysis** now refers to finding a model with which to make predictions about one variable from another.

From a family of politicians, brothers John, Robert, and Edward Kennedy stand together in 1960.

In Chapter 3, you saw one line of fit for bivariate data, the median-median line. In this lesson you'll learn about the **least squares line.** You find the least squares line by first standardizing both variables—the x-values and the y-values—which gives the bivariate data center (0, 0) and standard deviation 1 in both the horizontal and vertical directions. Then fit a line that passes through the origin and has slope equal to the correlation coefficient, r. In terms of z-values for x and y, the equation of the least squares line is $z_y = rz_x$. In the extreme case where the data are all perfectly linear, the value of r is +1 or –1, so the equation is $z_y = z_x$ or $z_y = -z_x$. In the other extreme, when the data are very scattered, the value of r is 0 and the equation is $z_y = 0$, a horizontal line through the origin.

In practice, you don't want to standardize every piece of sample data. Instead, you can rewrite the equation using means and standard deviations. By the definition of z-values, the equation $z_y = rz_x$ is equivalent to $\frac{y-\bar{y}}{s_y} = r\left(\frac{x-\bar{x}}{s_x}\right)$, or, solving for y, $y = \bar{y} + r\left(\frac{s_y}{s_x}\right)(x - \bar{x})$. Notice that this equation represents a translation of the center from the origin to $(\bar{x}, \bar{y})$.

Finding a Least Squares Line

1. Find the values of r, s_x, s_y, $\bar{x}$, and $\bar{y}$ for the data set.

2. Calculate the slope, $b = r\left(\frac{s_y}{s_x}\right)$.

3. Substitute values of b, $\bar{x}$, and $\bar{y}$ to write the equation for the least squares line, $\hat{y} = \bar{y} + b(x - \bar{x})$.

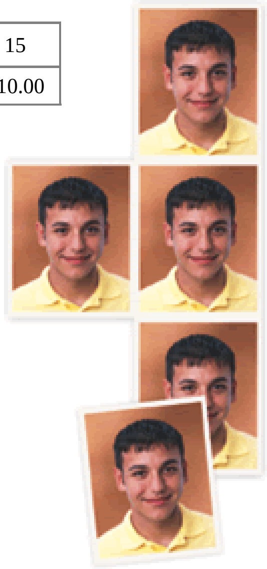
© 2004 Key Curriculum Press

EXAMPLE A

A photography studio offers several packages to students who pose for yearbook photos.

Number of pictures (x)	44	31	24	15
Total cost (y)	\$19.00	\$16.00	\$13.00	\$10.00

Find an equation of the least squares line. Use your line to decide how much a package of two photographs should cost.



Solution

Begin by finding the mean and standard deviation of both the x- and y-values.

$\bar{x} = 28.5 \quad \bar{y} = 14.5 \quad s_x = 12.23 \quad s_y = 3.873$

Then create a table to calculate values of $(x-\bar{x})(y-\bar{y})$.

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$
44	19	15.5	4.5	69.75
31	16	2.5	1.5	2.25
24	13	−4.5	−1.5	6.75
15	10	−13.5	−4.5	60.75

So, $\sum(x-\bar{x})(y-\bar{y})$ is 141. Use the formula $r = \dfrac{\sum(x-\bar{x})(y-\bar{y})}{s_x s_y(n-1)}$ to find $r \approx .992$.

The slope of the least squares line is $r\left(\dfrac{s_y}{s_x}\right) = \dfrac{.992 \cdot 3.873}{12.23} = 0.314$ dollar per photo.

So, the equation of the least squares line is $\hat{y} = 14.5 + 0.314\,(x - 28.5)$, or $\hat{y} = 5.55 + 0.314x$.

To find the cost of two photographs, substitute 2 for x. You get $y = 6.178$, so according to the model the package should cost \$6.18.

The least squares line has some interesting properties. You'll discover some of them in the investigation.

Investigation
Relating Variables

You will need

- rulers or metersticks

Your class will collect data on the measurements listed below and look for correlations. Then you'll explore a property of the least squares line.

Step 1 Take these measurements in centimeters. Choose one member of your group to post the measurements for your group.

hand span
foot length
length of little finger
height

length of cubit (from tip of middle finger to elbow)
length of lower leg (from knee to floor)
length of upper arm (from shoulder to elbow)
width of thumbnail

© 2004 Key Curriculum Press

Step 2 As a group, decide on two of the measurements that you think might be linearly related. Enter the data for these measurements into list L1 and list L2 on your calculator, and find the correlation coefficient. Are the data linearly related? If not, try another pair of measurements until you find data that are related linearly.

Step 3 Find the equation of the least squares line for your data, and graph the line with your data. Does the line appear to be a good fit?

Step 4 Run the LSL program. [► See Calculator Note 13G. ◄] Adjust the line produced by your calculator until the sum of the squares of the residuals is as small as possible. Then answer these questions.

a. How does the line you found using the LSL program compare to the least squares line you found in Step 3?

b. What is the sum of the residuals?

c. What property of the line do you think gives it its name?

d. Write a sentence or two describing the relationship between the measurements you analyzed.

You can measure the accuracy of the least squares line using the typical spread of the residuals, as calculated by the root mean square error.

EXAMPLE B

In Chapter 3, you estimated a line of fit for data on the concentration of CO_2 in the atmosphere around Mauna Loa in Hawaii as a function of time. Refer back to the table on page 129 to find

a. The least squares line for the data given.

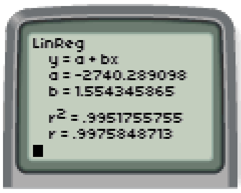
b. The median-median line for the data given.

c. The root mean square error of both models.

► Solution

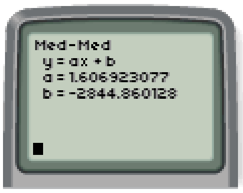
Enter the data into lists in your calculator and find each model. [► See Calculator Note 13H to learn how to find the equation of the least squares line. ◄]

a. The equation for the least squares line is $\hat{y} = -2740.289 + 1.544x$. Notice that $r \approx .998$, so a linear model is a good fit for these data.



```
LinReg
 y=a+bx
 a=-2740.289098
 b=1.554345865
 r²=.9951755755
 r=.9975848713
```

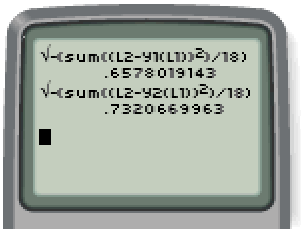
b. The equation of the median-median line is $\hat{y} = -2844.86 + 1.607x$.



```
Med-Med
 y=ax+b
 a=1.606923077
 b=-2844.860128
```

© 2004 Key Curriculum Press

c. To calculate the root mean square error, find the differences between the *y*-values in the data and the *y*-values predicted by the model. Then square the differences and sum the squares. Next, divide by two less than the number of data values, and finally, take the square root. Enter your equations for the least squares line and the median-median line into your calculator as Y1 and Y2. Then calculate the root mean square error as shown.



```
√(sum((L2-Y1(L1))²)/18)
        .6578019143
√(sum((L2-Y2(L1))²)/18)
        .7320669963
```

The root mean square error for the least squares line is 0.658 ppm and for the median-median line is 0.732 ppm. The root mean square error is smaller for the least squares line, so predictions you make with this model are likely to have smaller errors than predictions you make with the median-median line. In all cases, though, you should be careful not to predict too far in the future. Just because the trend has been quite linear in the past does not mean it will be so in the future.

There are many procedures to find lines of fit for linear data. Some, such as the median-median line, distill the data into a few points and are relatively unaffected by one or two outliers. Others, like the least squares line, place equal importance on each point. The least squares procedure produces the line that has the smallest sum of squares of errors between data points and predictions from the line. It is often called the "best-fit line" because of this property. However, when fitting a line to data, always check the line visually; you may sometimes find that a procedure that ignores outliers gives a line that models the overall trend better than the least squares line.

EXERCISES

Practice Your Skills

1. Use these data to find the values specified.

Year x	1950	1960	1970	1980	1990	2000
Percentage y	29.6	33.4	38.1	42.5	45.3	52.0

a. $\bar{x}$ **b.** $\bar{y}$ **c.** s_x **d.** s_y **e.** r

2. Use the following sets of statistics to calculate the least squares line for each data set described.

a. $\bar{x} = 18$, $s_x = 2$, $\bar{y} = 54$, $s_y = 5$, $r = .8$ **b.** $\bar{x} = 0.31$, $s_x = 0.04$, $\bar{y} = 5$, $s_y = 1.2$, $r = -.75$

c. $\bar{x} = 88$, $s_x = 5$, $\bar{y} = 6$, $s_y = 2$, $r = -.9$ **d.** $\bar{x} = 1975$, $s_x = 18.7$, $\bar{y} = 40$, $s_y = 7.88$, $r = .9975$

© 2004 Key Curriculum Press

3. Use the data from Exercise 1 and the equation $\hat{y} = -818.13 + 0.434571x$ to calculate

a. the residuals

b. the sum of the residuals

c. the squares of the residuals

d. the sum of the squares of the residuals

e. the root mean square error

4. Use the equation $\hat{y} = -818.13 + 0.434571x$ to predict a y-value for each x-value.

a. $x = 1954$ **b.** $x = 1978$ **c.** $x = 1989$ **d.** $x = 2004$

Reason and Apply

5. **APPLICATION** Carbon tetrachloride is an ozone-depleting chemical found in the atmosphere. The table below shows the concentration of the chemical in parts per trillion (ppt) measured in the European Union.

Year x	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989
Carbon tetrachloride (ppt) y	87.3	90.7	92.1	93.6	94.3	95.5	96.5	97.8	99.3	100.3	102.0	103.3

a. Make a scatter plot of the data, and find the least squares line to fit the data.

b. Use your model to predict the amount of carbon tetrachloride present in 2005.

c. In 1987, 22 countries and the European Economic Community agreed on the Montreal Protocol to reduce ozone-depleting chemicals in the atmosphere. In 1989, the protocol went into effect. The data below represent the levels of carbon tetrachloride from 1990 to 1997. Make a scatter plot of the data, and find the least squares line to fit the data.

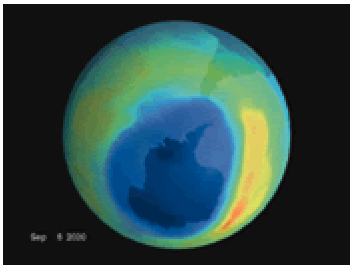
Year	1990	1991	1992	1993	1994	1995	1996	1997
Carbon tetrachloride (ppt)	104.2	103.2	102.9	102.1	101.3	100.8	99.5	98.6

(http://dataservice.eea.eu.int/dataservice/)

d. Use the model you found in 5c to predict the amount of carbon tetrachloride in 2005. How does this compare to your answer in 5b?

Environmental CONNECTION

Ozone depletion is a worldwide environmental concern that has been addressed by several international agreements. The ozone in the stratosphere (from 11 to 50 km above Earth's surface) protects us by blocking the Sun's ultraviolet (UV) radiation. Exposure to too much UV radiation has been linked to skin cancer, eye problems, and immune-system suppression. When ozone is depleted in the stratosphere, it can build up closer to the Earth's surface, where it acts as a pollutant and contributes to lung damage. For more information on ozone depletion, see the links at www.keymath.com/DAA .



This color-coded image of Earth shows a hole in the ozone layer above Antarctica. Blue and purple indicate low ozone levels, and yellow and orange indicate higher ozone.

6. In the 1990s, an upward trend was noticed in mean SAT math scores of college-bound seniors.

Year	1991	1992	1993	1995	1996	1997	1998	1999	2000
Score	500	501	503	506	508	511	512	511	514

(The World Almanac and Book of Facts 2002)

a. Find the equation of the least squares line. Let *x* represent the year, and let *y* represent the mean SAT score.

b. Is a linear model a good fit for these data? Justify your answer.

c. Verify that the least squares line passes through the mean *x*-value and the mean *y*-value.

d. What does the equation predict for the year 1994? How does this compare with the actual mean score of 504?

e. What does the equation predict for the year 2010? How reasonable is this prediction?

7. These data give the average daily maximum temperatures in April for various cities in North America, and the corresponding latitudes in degrees and minutes north.

City	**Lat.**	**Temp. (°F)**
Acapulco, Mex.	16°51′	87
Bakersfield, CA	35°26′	73
Caribou, ME	46°52′	50
Charleston, SC	32°54′	74
Chicago, IL	41°59′	55
Dallas, TX	32°54′	75
Denver, CO	39°46′	54
Duluth, MN	46°50′	52
Great Falls, MT	47°29′	56
Juneau, AK	58°18′	39
Kansas City, MO	39°19′	59
Los Angeles, CA	33°56	69

City	**Lat.**	**Temp. (°F)**
Mexico City, Mex.	19°25′	78
Miami, FL	25°49′	81
New Orleans, LA	29°59′	77
New York City, NY	40°47′	60
Ottawa, Ont.	46°26′	51
Phoenix, AZ	33°26′	83
Quebec, Que.	46°48′	45
Salt Lake City, UT	40°47′	58
San Francisco, CA	37°37′	65
Seattle, WA	47°27′	56
Vancouver, BC	49°18′	58
Washington, DC	38°51′	64

a. Find the equation of the least squares line. Let *x* represent the latitude, and let *y* represent the temperature. (You will need to convert the latitudes to decimal degrees; for example, $35°26' = 35\frac{26}{60}° \approx 35.43°$.)

b. What is an appropriate domain for this model?

c. Which cities do not appear to follow the pattern? Give a reason for each of these cases.

d. Choose two cities not on the list, and find the latitude of each. Use your model to predict the average daily maximum temperature in April for each city. Compare your result with the official average April temperature for the city. (An almanac is a good source for this information.)

© 2004 Key Curriculum Press

8. This table shows the percentage of females in the U.S. labor force at various times.

Year	1950	1960	1970	1980	1990	2000
Percentage	29.6	33.4	38.1	42.5	45.3	52.0

a. Find the least squares line for these data. Let x represent the year, and let y represent the percentage. (Use 1900 as the reference year.)

b. What is the real-world meaning of the slope? Of the y-intercept?

c. According to your model, what percentage of the current labor force is female? Check an almanac to see how accurate your prediction is.

A firefighter manages a controlled fire in New Jersey.

9. Name at least two major differences between the median-median method and the least squares method for finding a line of fit.

10. Explain how the root mean square error is related to the sum of the squares that is minimized by the least squares procedure. If the root mean square error is minimized, is the sum of the squares minimized as well?

Review

11. Solve each equation for y.

a. $\log y = 3$

b. $\log x + 2\log y = 4$

12. This table shows the number of daily newspapers in the United States, the daily circulation, and the total U.S. population, for selected years.

Daily Newspapers in the United States

Year	Number of daily papers	Daily circulation (millions)	Total population (millions)
1900	2226	15.1	76.2
1920	2042	27.8	106.0
1940	1878	41.1	132.1
1960	1763	58.9	179.3
1980	1745	62.2	226.5
2000	1480	55.8	248.7

(The New York Times Almanac 2002)

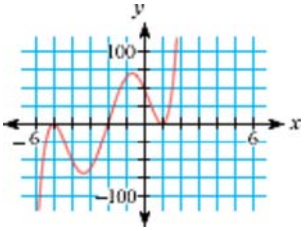
a. What is the correlation coefficient between the number of daily newspapers and the population? What does this mean?

b. What is the correlation between daily circulation and the total population?

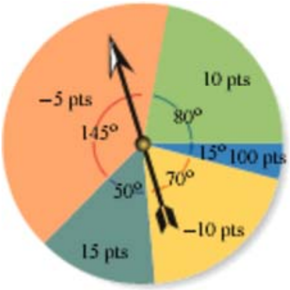
c. For each year, what percentage of the population reads a daily paper? What does this mean? What are the trends? What are the implications of the data?

© 2004 Key Curriculum Press

13. Write an equation that will produce the graph shown at right, with intercepts (−5, 0), (−2, 0), (1, 0), and (0, 60).



14. If you spin this spinner ten times, what is your expected score?

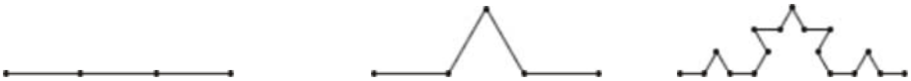


15. The Koch curve is a famous fractal introduced in 1906 by Swedish mathematician Niels Fabian Helge von Koch (1870-1924). To create a Koch curve, follow these steps:

i. Draw a segment and divide it into thirds.

ii. Make a bottomless triangle on the middle portion, with each side the length of the missing bottom.

iii. Repeat steps i and ii with each shorter segment.



If the original segment is 18 cm and the process continues through infinitely many steps, how long will the Koch curve become?

IMPROVING YOUR REASONING SKILLS

A Set of Weights

You have a 40-ounce bag of sand, a balance scale, and many blocks of wood with unknown weights. How can you divide the sand into four smaller bags so that you can then use the smaller bags to weigh any block of wood with a whole-number weight between 1 and 40 ounces?

© 2004 Key Curriculum Press

LESSON 13.7

Nonlinear Regression

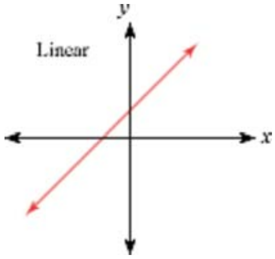
Telling the future by looking at the past assumes that conditions remain constant. This is like driving a car by looking in the rearview mirror.

HERB BRODY

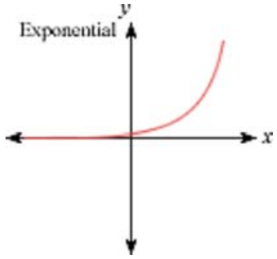
You have seen how to do regression analysis when bivariate sample data have a strong linear correlation. But what about bivariate data with a clear trend that isn't linear? Can you find the equation of a parabolic or exponential graph such that the sum of the squares of the residuals is as small as possible? In this lesson you'll see how to modify some data like these so that you can apply linear regression techniques.

The first task is to decide what type of function best fits the shape of the data. Standard shapes you've seen in this course fall into two categories.

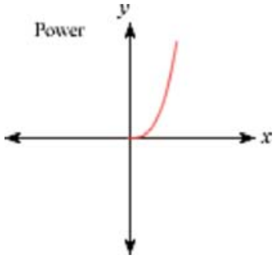
In the first type of model, the variable, *x*, appears only once. Here are the most common of these functions:



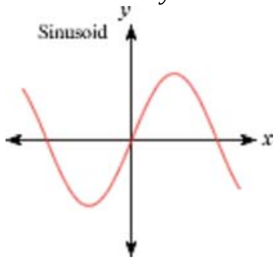
General form: $y = a + bx$



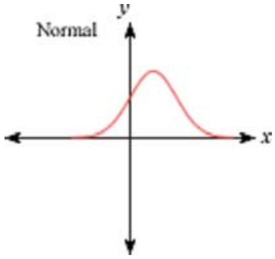
General form: $y = ab^x$



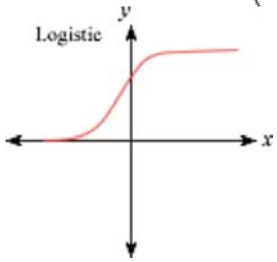
General form: $y = ax^b$



General form: $y = a\sin\left(\frac{x-h}{b}\right) + k$



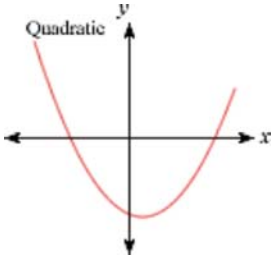
General form: $y = ae-^{((x-c)/b)^2}$



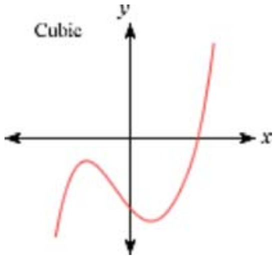
General form: $y = \frac{c}{1+ab^x}$

© 2004 Key Curriculum Press

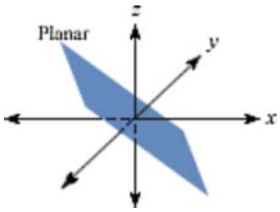
In the second type of model, there is more than one independent variable or the variable appears more than one time. Here are some examples of these functions:



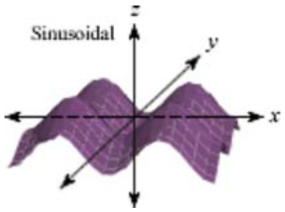
General form: $y = ax^2 + bx + c$



General form: $y = ax^3 + bx^2 + cx + d$



General form: $z = ax + by + c$



General form: $z = a \sin(bx + cy)$

Some equations of the second type can be transformed into equations of the first type; others will be left to later courses. Be aware that you may not yet have all the tools you need to model every data set in the best way possible.

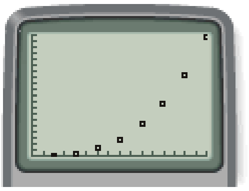
EXAMPLE A

What kind of functions might fit these data?

Time (s) x	2	4	6	8	10	12	14	16
Distance (cm) y	7.5	48.5	151	337.5	631	1059	1626	2360

▸ Solution

By looking at a graph of the data, you can determine that an exponential function, a power function, or a polynomial function (such as a quadratic function) might provide the best fit.



[0, 16, 1, 0, 2400, 100]

At this point, you might try using the finite difference method to see whether a polynomial function fits the data. You would find that a set of constant finite differences never occurs, so a polynomial function is not a good fit.

If your data seem to fit exponential and power models, you can check whether or not these are good models by **linearizing** the data. If the fit is good, then you can extend the linear regression techniques of Lesson 13.6.

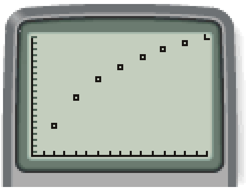
EXAMPLE B

What exponential or power function provides the best fit for the data in Example A?

▶ Solution

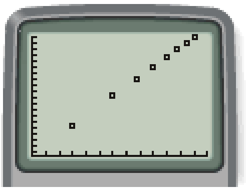
If the data are exponential, an equation in the form $y = ab^x$ will be a good fit. You can use the rules of logarithms to show that $y = ab^x$ is equivalent to $\log y = \log a + x \log b$. Because a and b are constants, $\log a$ and $\log b$ are also constants. So the equation $\log y = (\log a) + x(\log b)$ is a linear equation, where the variables are x and $(\log y)$. Thus, the graph of $(x, \log y)$ should be linear.

A graph shows that $(x, \log y)$ isn't linear, so an exponential function must not be a good fit. Can you fit the data better with a power function?



[0, 16, 1, 0, 3.4, 0.1]

If a power function, $y = ax^b$, fits these data, then by rules of logarithms, $\log y = \log a + b \log x$. Because $\log a$ and b are constants, this is a linear equation with variables $(\log x)$ and $(\log y)$. Thus, a graph of $(\log x, \log y)$ should be linear if the data can be modeled by a power function.

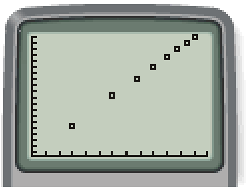


[0, 1.3, 0.1, 0, 3.4, 0.1]

The plot of $(\log x, \log y)$ looks linear. The equation of the least squares line for these data is $\hat{y} = 2.77x + 0.03$.

Now replace $\hat{y}$ with $(\log y)$ and x with $(\log x)$, then solve for y.

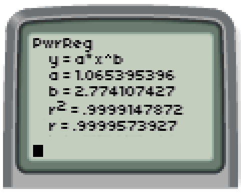
$$\begin{aligned}
\log y &= 2.77 \log x + 0.03\\
y &= 10^{2.77 \log x + 0.03}\\
&= 10^{2.77 \log x} \cdot 10^{0.03}\\
&= (10^{\log x})^{2.77} \cdot 10^{0.03}\\
&= x^{2.77} \cdot 10^{0.03}\\
y &= 1.07x^{2.77}
\end{aligned}$$



[0, 16, 1, 0, 2400, 100]

The function that fits the data is the power function $\hat{y} = 1.07x^{2.77}$. Using this equation, the root mean square error is approximately 21.7, which is small considering the size of the data values.

The linearization process you use to fit power or exponential curves is built into most graphing technology. [▶ See Calculator Note 13I. ◀] You can use a calculator to verify your answer to Example B.



```
PwrReg
  y=a*x^b
  a=1.065395396
  b=2.774107427
  r²=.9999147872
  r=.9999573927
```

Your calculator can also find polynomial equations that fit data with the smallest sum of the residuals squared. You can use the finite differences method to find a polynomial curve that passes through *selected* data points, but regression techniques will find a curve that is a good fit for the whole data set.

© 2004 Key Curriculum Press

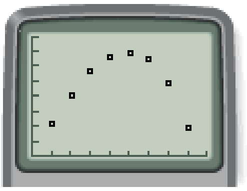
EXAMPLE C

Enter these data into lists and graph them. Then identify what types of curves may be good models.

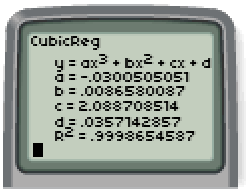
Base (in.)	1	2	3	4	5	6	7	8
Area (in.^{2})	2.1	4.0	5.6	6.6	6.9	6.4	4.8	1.9

▸ Solution

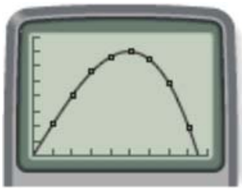
The data do not appear symmetric, so a quadratic function is not a good option. It appears that a cubic function may be a good fit.



Use the cubic regression command to find a third-degree polynomial that is a good fit.



```
CubicReg
  y=ax³+bx²+cx+d
  a=-.0300505051
  b=.0086580087
  c=2.088708514
  d=.0357142857
  R²=.9998654587
```



The equation that models these data is $\hat{y} = -0.030x^3 + 0.009x^2 + 2.089x + 0.036$.

In this investigation you will collect data and find a function that fits the data.

Investigation

A Leaky Bottle Experiment

You will need

- a plastic water container with a hole near the bottom
- a metric ruler
- tape
- a timing device
- water (with a bit of food coloring if available)

Procedure Note

1. Assign each group member a role: bottle holder, timekeeper, water-level reader, or recorder.
2. Attach the ruler to the container with tape so that 0 cm is at the bottom of the bottle.
3. Fill the container with water to the 15 cm mark, keeping a finger on the hole.
4. When the timekeeper begins timing, the bottle holder removes his or her finger from the hole and lets the water run out freely.
5. The timekeeper calls out the time every ten seconds.
6. The water-level reader reads aloud the water level to the nearest millimeter.
7. The recorder records the data.
8. Stop measuring the water level before it reaches the curved bottom of the bottle.

Step 1 Follow the procedure note to collect data.

Step 2 Sketch a graph of the data. Look at the graph and make a conjecture about the type of functions that might fit the data.

Step 3 Find equations of several different types that fit the data well.

Step 4 Create a separate graph of the residuals of each equation, and calculate the root mean square error.

Step 5 Select the best equation, and add its graph to your data plot from Step 2.

Step 6 Use your model to predict when the container would be empty.

If your model fits the data very well, then the residuals will not increase or decrease in any noticeable pattern. But sometimes it can be difficult to fit a curve to data even when you've guessed a good function. In science and in industry, it may be sufficient to find a relatively simple function that produces results that are "close enough." For example, it is common in industry to use polynomial models-even though they may not provide the best fit, they are usually quick to find and fit well enough to predict data in the near future.

EXERCISES

Practice Your Skills

1. Sketch scatter plots of transformed data in the form specified.

Time (h) x	1	2	3	4	5
Percentage y	65.0	50.0	42.5	38.0	35.0

a. (x, y) **b.** $(\log x, y)$ **c.** $(x, \log y)$ **d.** $(\log x, \log y)$

e. Which data appear to be the most linear?

© 2004 Key Curriculum Press

2. The data from Exercise 1 were gathered again the next day. After the experiment had run 24 hours, the y-values had dropped by nearly 20%. Subtract 20 from each y-value and create a scatter plot of

a. (x, y) **b.** $(\log x, y)$ **c.** $(x, \log y)$ **d.** $(\log x, \log y)$

e. Which data appear to be the most linear?

3. Solve each equation for $\hat{y}$, and rewrite the function in one of these forms:

$\hat{y} = a + bx$ $\hat{y} = ab^x + c$ $\hat{y} = ax^b + c$ $\hat{y} = a + b \log x$

a. $\hat{y} - 20 = 47.7 - 7.2x$

b. $\hat{y} - 20 = 44 - 43.25 \log x$

c. $\log(\hat{y} - 20) = 1.7356 - 0.227690x$

d. $\log(\hat{y} - 20) = 1.66586 - 0.68076 \log x$

4. Use the data from Exercise 1 and find

a. A quadratic (2nd degree) regression.

b. A cubic (3rd degree) regression.

c. A quartic (4th degree) regression.

d. The root mean square error for each regression.

Reason and Apply

5. APPLICATION A cylindrical tanker truck has a volume of 50 m^3 when it is full. The driver can use a stick to find the depth of the contents. The following information is known:

Depth (m)	0	0.5	1.0	1.5	2.0	2.5
Volume (m^3)	0	7.12	18.68	31.32	42.88	50

a. Find a cubic model to estimate the volume for different depths.

b. What is the root mean square error for the cubic model from 5a?

c. Predict the volume when the depth is 0.75 m.

d. What kind of accuracy do you expect for this value?

6. APPLICATION An additive to puppy food is shown to increase weight gain in underweight puppies when it is mixed with standard food. What mixture results in highest weight gain for the average puppy? Data are collected from a study of eight puppies fed with different percentages of additive.

Percentage additive x	20%	20%	40%	40%	60%	60%	80%	80%
Weight gain (kg) y	4.1	6.2	6.5	7.3	3.1	4.8	0.5	1.2

a. Find quadratic and cubic models for these data.

b. Use each model to find the predicted percentage that produces the greatest weight gain.

c. How much difference is there in each of these predictions?

© 2004 Key Curriculum Press

7. ***Mini-Investigation*** For nonlinear data that cannot be linearized (such as polynomial functions), you cannot calculate the coefficient of correlation, r. Instead, you can calculate the **coefficient of determination**, R^2. A value of R^2 close to ±1 indicates a good fit. Use the data and answers from Exercise 6.

a. Calculate the mean y-value of the data.

b. Calculate the sum of the squares of the deviations for the y-values, $\sum(y_i - \bar{y})^2$.

c. Calculate the sum of the squares of the residuals predicted by the quadratic model, $\sum(y_i - \hat{y})^2$.

d. Find the proportion of change in these values:

$$R^2 = \frac{\sum(y_i - \bar{y})^2 - \sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

e. Repeat these calculations for the cubic model. Which is a better fit?

f. Find a linear model for the data, and repeat the calculations to find R^2. How does the value of R^2 for the linear model compare with the value of r^2?

8. **APPLICATION** The speed of a computer has increased with time in a consistent way. In 1965, Dr. Gordon Moore (b 1929), a co-founder of Intel, noted that the number of transistors per square inch on integrated circuits seems to double approximately every 18 months. The table below shows the year of introduction of a new chip and the number of transistors on that chip.

a. Make a scatter plot of the data. (Let $x = 0$ represent 1970.) Is the data linear?

b. Make a scatter plot of $(\log x, y)$, $(x, \log y)$, and $(\log x, \log y)$. Which is most linear?

c. Find a least squares line to model the most linear data you found in 8b.

d. Using the least squares line in 8c, write an equation to model the data.

e. Use your model to make a prediction about the number of transistors on a microchip in 2011.

Year of introduction x	Number of transistors y
1971	2,250
1972	2,500
1974	5,000
1978	29,000
1982	120,000
1985	275,000
1989	1,180,000
1993	3,100,000
1997	7,500,000
1999	24,000,000
2000	42,000,000

(*www.intel.com*)

Technology CONNECTION

As microchips are made smaller and smaller, more and more transistors can fit onto a chip. The chip's speed increases because the distance between the transistors is decreased, resulting in an increase in computer performance. Is Moore's Law limitless? In 1997, Moore said that the physical limitations of silicon chips could be reached by 2017. Researchers are now experimenting with replacing silicon transistors with carbon nanotubes to fit more transistors on a chip.

A 1984 IBM AT computer (left) compared to a 2000 Sharp handheld computer shows how computer sizes have decreased with time and technological advances.

© 2004 Key Curriculum Press

Review

9. A set of data with a mean of 83 and a standard deviation of 3.2 is normally distributed. Find each value.

a. one standard deviation above the mean

b. one standard deviation below the mean

c. two standard deviations above the mean

10. A poll shows 47% of voters favor Proposition B. What is the probability that exactly 11 of 20 voters end up voting in favor of the proposal?

11. These data show numbers of country radio stations and numbers of oldies radio stations for several years. Assuming the trends in the data continue, answer the questions that follow.

Year	1994	1995	1996	1997	1998	1999	2001
Country	2642	2613	2525	2491	2368	2306	2190
Oldies	714	710	738	755	799	766	785

(*The World Almanac and Book of Facts 2002*)

a. How many country stations will there be in 2005?

b. How many oldies stations will there be in 2005?

c. When the number of country stations reaches 1500, how many oldies stations are there likely to be?

12. In your class, if three students are picked at random, what is the probability that

a. all were born on a Wednesday?

b. at least one was born on a Wednesday?

c. each was born on a different day of the week?

Singer Patsy Cline's (1932-1963) (top) recordings are considered some of the greatest in country music. Between 1998 and 2003, the Dixie Chicks (bottom) recorded three albums and have released seven number-one singles.

Project

MAKING IT FIT

Find any bivariate data that you think might be related. You might look in an almanac, or search the library or the internet. Then use techniques from this chapter to find a function that fits the data well.

Your project should include

- Your data and its source.
- A graph of your data with the equation you found to model it.
- A description of your process and an analysis of how well your curve fits the data.

CHAPTER 13 REVIEW

In this chapter you saw some statistical tools for estimating **parameters** of a very large (perhaps infinite) population from **statistics** of samples taken from that population. Many large populations can be described by **probability distributions** of **continuous random variables**. The area under the graph of a probability distribution is always 1. When using any probability distribution, you do not find the probability of a single exact value; you find the probability of a range of values.

The probability distributions of many sets of data are **normal**. Their graphs, called **normal curves**, are bell-shaped. To write an equation of a normal distribution curve, you need to know the mean and standard deviation of the data set. To make predictions about a population based on a sample, you can make a nonstandard normal distribution standard by using **z-values**, and you can predict things about the population mean with a **confidence interval**.

Even if the population is not normal, the **Central Limit Theorem** allows you to discuss its mean and standard deviation based on sample statistics. This theorem also allows you to perform **hypothesis testing** about a population parameter.

You can also make predictions by collecting **bivariate data** and analyzing the relationship between two variables. The **correlation coefficient**, r, tells whether or not the variables have a linear relationship. If two variables are linearly related, the **least squares line** of fit, which passes through $(\bar{x}, \bar{y})$ and has slope $r\left(\frac{s_y}{s_x}\right)$, can help you make predictions about the population. You can also find curves to fit some nonlinear data by linearizing those data and finding a least squares line, or by using your calculator to perform other **regression analysis** techniques.

EXERCISES

Practice Your Skills

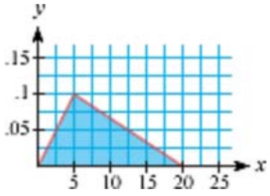
1. A graph of a probability distribution consists of two segments. The first segment connects the points (0, 0) and (5, .1), and the second segment connects the points (5, .1) and (20, 0), as shown.

   a. Verify that the area under the segments is equal to 1.

   b. Find the median.

   c. What is the probability that a data value will be less than 3?

   d. What is the probability that a data value will be between 3 and 6?



© 2004 Key Curriculum Press

2. Suppose a probability distribution has the shape of a semicircle. What is the radius of the semicircle?

3. The 68% confidence interval for the weight of a house cat in northern Michigan is between 8.4 and 12.7 lb.

   a. What are the mean and standard deviation for the weight of a house cat in northern Michigan?
   b. What is the 95% confidence interval for the weight of a house cat in northern Michigan?

4. At a maple tree nursery, a grower selects a random sample of 5-year-old trees. He measures their heights to the nearest inch.

Height (in.)	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75
Frequency	1	1	4	6	8	10	9	8	13	10	5	6	6	3	6	3

   a. Find the mean and the standard deviation of the heights. Make a statement about the meaning of the standard deviation in this problem.
   b. Make a histogram of these data and describe the shape of the distribution.

5. Recently, Adam put larger wheels on his skateboard and noticed it would coast farther. He decided to test this relationship. He collected some skateboards, attached wheels of various sizes to them, and rolled them to see how far they would roll on their own. He collected the data shown in the table.

Wheel diameter (in.) x	**Rolling distance (in.)** y
1	17
2	23
3.5	32
5	30
5.5	36
7	52
8.5	57
9.5	55
10	70

Dirtboards use wheels up to 10 inches in diameter.

   a. Make a scatter plot of these data.
   b. Does there appear to be a linear relationship between the variables? Find the correlation coefficient, and use this value to justify your answer.
   c. Find the equation of the least squares line that models these data.
   d. Describe the real-world meaning of the slope and the y-intercept of your line.
   e. Use your model to determine the size of wheel of a skateboard that rolls 50.5 in.

© 2004 Key Curriculum Press

6. Suppose the weights of all male baseball players who are 6 ft tall and between the ages of 18 and 24 are normally distributed. The mean is 175 lb, and the standard deviation is 14 lb.

a. What percentage of these males weigh between 180 and 200 lb?

b. What percentage of these males weigh less than 160 lb?

c. Find the 90% confidence interval.

d. Find the equation of a normal curve that provides a probability distribution for this information.

7. The length of an algebra book is measured by many people. The measurements have mean 284 mm and standard deviation 1.3 mm. If four students measure the book, what is the probability that the mean of their measurements will be less than 283 mm?

8. A 3 ft deep fish pond in the shape of a hemisphere has a volume of 56.549 ft^3 when it is full. These data have been collected relating depth to volume in the pond:

Depth (ft)	0	0.5	1.0	1.5	2.0	2.5	3.0
Volume (ft^3)	0	2.225	8.378	17.671	29.322	42.542	56.549

a. Find the equation of the least squares line that estimates the volume as a function of depth.

b. Plot your equation from 8a with the data points, and decide whether you think it is a good fit.

c. Calculate the values of r and r^2 for a linear model. What do they tell you about whether or not the least squares line is a good fit?

d. Find quadratic and cubic models to estimate the volume for different depths. Graph these curves with the data.

e. Which of the three regression models (linear, quadratic, or cubic) appears to fit the data best?

f. What are the values of R^2 for the quadratic and cubic models? How does this confirm or refute your answer to 8e?

MIXED REVIEW

9. Sketch a graph of each equation, and identify the shape formed.

a. $\frac{x^2}{12} - \frac{(y+3)^2}{9} = 1$

b. $\left(\frac{x-1}{5}\right)^2 + \left(\frac{y-1}{4}\right)^2 = 1$

c. $(y-3)^2 = \frac{x-4}{3}$

d. $-4x^2 - 24x + y^2 + 2y = 39$

10. If the probability of a snowstorm in July is .004, and the probability you will score an A in algebra is .75, then what is the probability of a snowstorm in July or an A in algebra?

11. Find the sums of each series.

a. Find the sum of the first 12 odd positive integers.

b. Find the sum of the first 20 odd positive integers.

c. Find the sum of the first n odd positive integers. (*Hint:* Try several choices for n until you see a pattern.)

12. **APPLICATION** A Detroit car rental business has a second outlet in Chicago. The company allows customers to make local rentals or one-way rentals to the other location. At the end of each month, one-eighth of the cars that start the month in Detroit will end up in Chicago, and one-twelfth of the cars that start the month in Chicago will end up in Detroit.

a. Write a transition matrix to represent this situation.

b. If there are 500 cars in each city at the start of operations, what would you expect the distribution to be four months later? In the long run?

13. **APPLICATION** A store in Yosemite National Park charges \$6.60 for a flashlight. Approximately 200 of them are sold each week. A survey indicates that the sales will decrease by 10 flashlights per week for each \$0.50 increase in price.

a. Write a function that describes the weekly revenue in dollars, y, as a function of selling price in dollars, x.

b. What selling price provides maximum weekly revenue? What is the maximum revenue?

Camping at Glacier Point, Yosemite National Park, California

14. Consider the function $y = \cos x$.

a. Write the equation of the image after the function is reflected across the x-axis, shrunk by a vertical scale factor of $\frac{1}{2}$, stretched by a horizontal scale factor of 2, and translated up 6 units.

b. What is the period of the image, in radians? What are the amplitude and phase shift?

c. Graph the function and its image on the same graph.

15. **APPLICATION** Lily and Philip both go to their doctor, complaining of the same symptoms. The doctor tests them for a rare disease. Data have shown that 20% of the people with these symptoms actually have the disease. The test the doctor uses is correct 90% of the time. Calculate the probabilities in the table below, and explain the meaning of the results.

		Test results	
		Accurate	**Inaccurate**
Patient's condition	Doesn't have the disease		
	Has the disease		

16. The population of Bombay, India, at various times is given in the table below.

Year	1950	1970	1990	2000
Population (in millions)	2.9	5.8	12.2	18.1

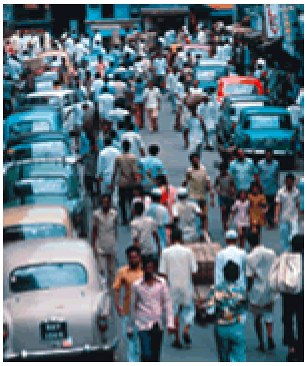
(*The New York Times Almanac 2002*)

a. The population roughly doubled in the 20 years between 1950 and 1970 and slightly more than doubled again between 1970 and 1990. What is a good estimate of the growth rate?

b. Find an exponential equation to model Bombay's population.

c. Use your model to predict the population in 2015.

d. *The New York Times Almanac 2002* predicts that the population in 2015 will be 26.1 million. How does this compare with your prediction?



Bombay, India

17. This table shows the number of seats on various types of airplanes, the planes' cruising speed, and their operating cost per hour.

Plane	**Seats**	**Speed (mi/h)**	**Operating cost ($/h)**
B747-400	369	537	8,158
B747-200/300	357	522	8,080
L-1011	339	493	8,721
DC-10-10	309	513	5,000
DC-10-40	284	491	6,544
DC-10-30	273	520	6,388
MD-11	270	525	7,474
B-777	266	525	4,878
A300-600	228	479	5,145
B767-300ER	207	499	3,823
B767-200ER	176	487	4,406
MD-90	149	441	2,590
B737-800	148	454	2,255
B727-200	147	439	3,435
A320-100/200	146	454	2,492
B737-400	141	407	2,948
MD-80	135	431	2,725
B737-300/700	131	408	2,417
DC-9-50	126	365	1,954
A319	122	445	1,987
B737-100/200	117	401	2,601
DC-9-40	111	380	1,845
B737-500	109	410	2,397
B717-200	106	374	2,212
DC-9-30	97	392	2,218
F-100	88	380	3,015
DC-9-10	69	389	2,227
CRJ-145	50	389	1,033
ERJ-145	50	362	1,151
ERJ-135	37	363	1,028

(*The World Almanac and Book of Facts 2003*)

a. How strongly is the number of seats related to operating cost? How strongly is the speed related to operating cost?

b. Does the number of seats or the speed have the stronger correlation to operating cost? Why might that correlation be stronger?

© 2004 Key Curriculum Press

18. Identify each sequence as arithmetic, geometric, or neither. Then write both a recursive and explicit formula to describe the pattern, if possible.

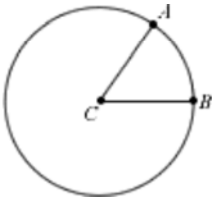
a. 3, 9, 27, 81, 243, . . .

b. –1, –3, –5, –7, –9, . . .

c. 2, 5, 10, 17, 26, . . .

d. $1, -\frac{1}{2}, \frac{1}{4}, -\frac{1}{8}, \frac{1}{16}, \ldots$

19. The circle at right has radius 4 cm, and the measure of central angle *ACB* is 55°.



a. What is the measure of $\angle ACB$ in radians?

b. What is the length of $\overset{\frown}{AB}$?

c. What is the area of sector *ACB*?

20. The lengths in feet of the main spans of 38 notable suspension bridges in North America are

{4260, 4200, 3800, 3500, 2800, 2800, 2310, 2300, 2190, 2150, 2000, 1850, 1800, 1750, 1632, 1600, 1600, 1600, 1595, 1550, 1500, 1500, 1470, 1447, 1400, 1380, 1207, 1200, 1150, 1108, 1105, 1080, 1060, 1059, 1057, 1050, 1030, 1010}

(*The World Almanac and Book of Facts 2003*)

a. What are the mean, median, and mode of these data?

b. Make a box plot of these data. Describe the shape.

c. What is the standard deviation?

New York's Manhattan Bridge was constructed from 1901 to 1909 over the East River.

21. Solve each system of equations.

a. $\begin{cases} 3x - y = -1 \\ 2x + y = 6 \end{cases}$

b. $\begin{cases} 2x + 4y = -9 \\ x - y = -6 \end{cases}$

22. Consider this series.

$$\frac{1}{10} + \frac{1}{30} + \frac{1}{90} + \frac{1}{270} + \ldots$$

a. What is the sum of the first five terms?

b. What is the sum of the first ten terms?

c. What is the sum of infinitely many terms?

© 2004 Key Curriculum Press

23. Consider the functions $f(x) = \sqrt{2x-3}$ and $g(x) = 6x^2$.

a. What are the domain and range of $f(x)$?
b. What are the domain and range of $g(x)$?
c. Find $f(2)$.
d. Find x such that $g(x) = 2$.
e. Find $g(f(3))$.
f. Find $f(g(x))$.

24. Two people begin 400 m apart and jog toward each other. One person jogs 2.4 m/s, and the other jogs 1.8 m/s. When they meet, they stop.

a. Write parametric equations to simulate the movement of the joggers. What range of t-values do you need?

b. Use your graph to find how far each person runs before they meet.

c. How long does it take for them to meet?

25. The heights of all adults in Bigtown are normally distributed with a mean of 167 cm and a standard deviation of 8.5 cm.

a. Sketch a graph of the normal distribution of these heights.

b. Shade the portion of that graph showing the percentage of people who are shorter than 155 cm.

c. What percentage of people are shorter than 155 cm?

TAKE ANOTHER LOOK

1. The least squares method minimizes the sum of the squares of the residuals. Why is this important? Try to think of another method you could use to find a line of fit. Explain what advantage or disadvantage the method of minimizing squares of residuals has that this other method does not have.

2. As you add to the degree of a polynomial function that models data, the value of the coefficient of determination, R^2, will increase. But increasing the degree of a polynomial doesn't necessarily *significantly* improve how well a function fits data. The formula below adjusts for the increase in accuracy that comes from increasing the degree of a polynomial function. The adjusted value of R^2, $R^2{}_A$, allows you to judge whether the model has a significant improvement. The variable n represents the number of data points, and p represents the number of parameters in the model. There are two parameters in a linear, exponential, or power model (a and b), there are three in a quadratic model (a, b, and c), four in a cubic model (a, b, c, and d), and so on.

$$R_A^2 = 1-(1-R^2)\left(\frac{n-1}{n-p}\right)$$

Consider these data, the depth of water in a leaking bucket at various times.

Time (s) x	0	10	20	30	40	50	60	70	80	90
Depth (cm) y	15	12.5	10.5	8.5	6.5	5.0	3.5	2.5	1.5	1.5

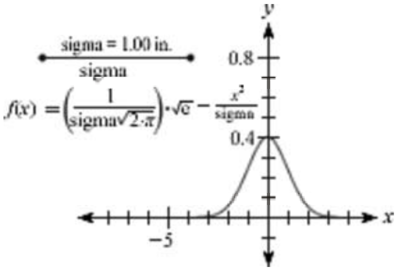
© 2004 Key Curriculum Press

Find several polynomial equations to model these data, state the value of R^2 that your calculator gives for each model, and use the formula to find $R^2{}_A$ for each model. What is the best model to describe these data?

3. Consider a normal distribution with mean 0. Use your graphing calculator or geometry software to explore how the standard deviation, σ, affects the equation of the normal curve,

$$y = \frac{1}{\sigma\sqrt{2\pi}}\left(\sqrt{e}\right)^{-(x/\sigma)^2}$$

Summarize how the normal curve changes as the value of σ changes.



4. Consider a normal distribution with mean 0. You've already seen that the equation of the normal curve is

$$y = \frac{1}{\sigma\sqrt{2\pi}}\left(\sqrt{e}\right)^{-(x/\sigma)^2}$$

To avoid using *e*, you can use another equation that approximates the normal curve:

$$y = \frac{1}{\sigma\sqrt{2\pi}}\left(1 - \frac{1}{2\sigma^2}\right)^{x^2}$$

Use your graphing calculator or geometry software to explore how the graphs of these two equations compare for different values of σ. For which values of σ is the second equation a good approximation, a poor approximation, or even undefined?

Assessing What You've Learned

WRITING TEST QUESTIONS Write a few test questions that reflect the topics of this chapter. You may want to include questions on probability distributions, confidence intervals, or bivariate data and correlation. Include detailed solutions.

ORGANIZE YOUR NOTEBOOK Make sure that your notebook has complete notes on all of the statistical tools and formulas that you have learned. Specify which statistical measures apply to populations and which apply to samples, and explain which tools allow you to make conclusions about a population based on a sample, and vice versa. Be sure you know when to use the various statistical measures and exactly what each one allows you to predict.

PERFORMANCE ASSESSMENT As a friend, family member, or teacher watches, solve a problem from this chapter that deals with fitting a line or curve to data and analyzing how well the function fits. Explain the various tools for analyzing how well a function fits data. Include a description of what values of *r* and/or R^2 tell you about a function's fit.

© 2004 Key Curriculum Press