Appendix C: Concepts in Statistics

C.1   Measures of Central Tendency and Dispersion

Mean, Median, and Mode

In many real-life situations, it is helpful to describe data by a single number that is most representative of the entire collection of numbers. Such a number is called a **measure of central tendency.** The most commonly used measures are as follows.

1. The **mean,** or **average,** of n numbers is the sum of the numbers divided by n.

2. The **median** of n numbers is the middle number when the numbers are written in numerical order. If n is even, then the median is the average of the two middle numbers.

3. The **mode** of n numbers is the number that occurs most frequently. If two numbers tie for most frequent occurrence, then the collection has two modes and is called **bimodal.**

EXAMPLE 1   Comparing Measures of Central Tendency

An analyst states that the average annual income of a company's 25 employees is $60,849. The actual annual incomes of the 25 employees are shown below. What are the mean, median, and mode of the incomes?   *(Spreadsheet at LarsonPrecalculus.com)*

DATA				
$17,305	$478,320	$45,678	$18,980	$17,408
$25,676	$28,906	$12,500	$24,540	$33,450
$12,500	$33,855	$37,450	$20,432	$28,956
$34,983	$36,540	$250,921	$36,853	$16,430
$32,654	$98,213	$48,980	$94,024	$35,671

Solution

The mean of the incomes is

$$\text{Mean} = \frac{17{,}305 + 478{,}320 + 45{,}678 + 18{,}980 + \cdots + 35{,}671}{25}$$

$$= \frac{1{,}521{,}225}{25}$$

$$= \$60{,}849.$$

To find the median, order the incomes as follows.

$12,500	$12,500	$16,430	$17,305	$17,408
$18,980	$20,432	$24,540	$25,676	$28,906
$28,956	$32,654	$33,450	$33,855	$34,983
$35,671	$36,540	$36,853	$37,450	$45,678
$48,980	$94,024	$98,213	$250,921	$478,320

From this list, you can see that the median income is $33,450. You can also see that $12,500 is the only income that occurs more than once. So, the mode is $12,500. ■

In Example 1, was the analyst telling the truth about the annual incomes? Technically, the person was telling the truth because the average is (generally) defined to be the mean. However, of the three measures of central tendency—*mean:* $60,849, *median:* $33,450, *mode:* $12,500—it seems clear that the median is most representative. The mean is inflated by the two highest salaries.

Choosing a Measure of Central Tendency

Which of the three measures of central tendency is most representative of a particular data set? The answer depends on the distribution of the data *and* the way in which you plan to use the data.

For instance, in Example 1, the mean salary of $60,849 would not seem very representative to a potential employee (the mean is inflated by the two highest salaries). A better use of the mean is for estimating income tax. For instance, to estimate a 1% income tax for the 25 employees, a tax collector would use 1% of the *mean* annual income of the employees.

EXAMPLE 2   Choosing a Measure of Central Tendency

Which measure of central tendency is most representative of the data given in each frequency distribution?

a.

Number	1	2	3	4	5	6	7	8	9
Frequency	7	20	15	11	8	3	2	0	15

b.

Number	1	2	3	4	5	6	7	8	9
Frequency	9	8	7	6	5	6	7	8	9

c.

Number	1	2	3	4	5	6	7	8	9
Frequency	6	1	2	3	5	5	8	3	0

Solution

a. For this data set, the mean is 4.23, the median is 3, and the mode is 2. Of these, the median or mode is probably the most representative measure.

b. For this data set, the mean and median are each 5, and the modes are 1 and 9 (the distribution is bimodal). Of these, the mean or median is the most representative measure.

c. For this data set, the mean is 4.88, the median is 5, and the mode is 7. Of these, the mean or median is the most representative measure.

Technology Tip

Most graphing utilities have *mean* and *median* features that can be used to find the means and medians of data sets. Enter the data from Example 2(a) in the *list editor* of a graphing utility. Then use the *mean* and *median* features to verify the solution to Example 2(a), as shown below.

```
mean(L1,L2)
              4.234567901
median(L1,L2)
                        3
```

For instructions on how to use the *list*, *mean*, and *median* features, see Appendix A; for specific keystrokes, go to this textbook's *Companion Website*.

Variance and Standard Deviation

Very different sets of numbers can have the same mean. You will now study two **measures of dispersion,** which give you an idea of how much the numbers in a set differ from the mean of the set. These two measures are called the *variance* of the set and the *standard deviation* of the set.

Definitions of Variance and Standard Deviation

Consider a set of numbers $\{x_1, x_2, \ldots, x_n\}$ with a mean of $\bar{x}$. The **variance** of the set is

$$v = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}$$

and the **standard deviation** of the set is $\sigma = \sqrt{v}$ (σ is the lowercase Greek letter sigma).

The standard deviation of a set is a measure of how much a typical number in the set differs from the mean. The greater the standard deviation, the more the numbers in the set vary from the mean. For instance, each of the sets

$$\{5, 5, 5, 5\}, \quad \{4, 4, 6, 6\}, \quad \text{and} \quad \{3, 3, 7, 7\}$$

has a mean of 5. The standard deviations of the sets are 0, 1, and 2, respectively.

$$\sigma_1 = \sqrt{\frac{(5 - 5)^2 + (5 - 5)^2 + (5 - 5)^2 + (5 - 5)^2}{4}} = 0$$

$$\sigma_2 = \sqrt{\frac{(4 - 5)^2 + (4 - 5)^2 + (6 - 5)^2 + (6 - 5)^2}{4}} = 1$$

$$\sigma_3 = \sqrt{\frac{(3 - 5)^2 + (3 - 5)^2 + (7 - 5)^2 + (7 - 5)^2}{4}} = 2$$

EXAMPLE 3   Estimations of Standard Deviation

Consider the three frequency distributions represented by the histograms in Figure C.1. Which data set has the least standard deviation? Which has the greatest?
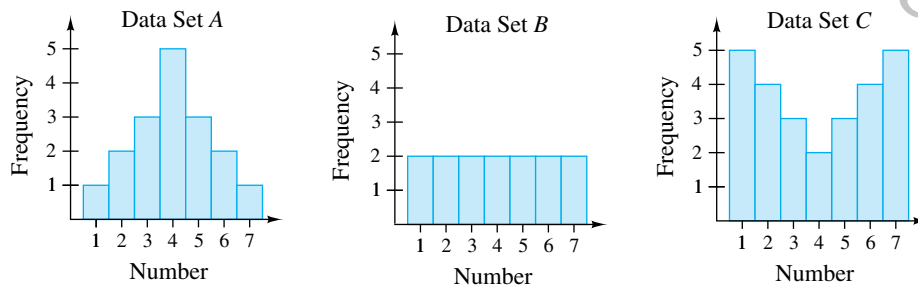

Figure C.1

Solution

Of the three data sets, the numbers in data set *A* are grouped most closely to the center and the numbers in data set *C* are the most dispersed. So, data set *A* has the least standard deviation and data set *C* has the greatest standard deviation.

<div style="background:#3a8dc6;color:white;padding:4px 12px;display:inline-block;">**EXAMPLE 4**</div>   **Finding a Standard Deviation**

Find the standard deviation of each data set shown in Example 3.

Solution

Because of the symmetry of each histogram, you can conclude that each has a mean of

$$\bar{x} = 4.$$

The standard deviation of data set A is

$$\sigma = \sqrt{\frac{(-3)^2 + 2(-2)^2 + 3(-1)^2 + 5(0)^2 + 3(1)^2 + 2(2)^2 + (3)^2}{17}}$$

$$\approx 1.53.$$

The standard deviation of data set B is

$$\sigma = \sqrt{\frac{2(-3)^2 + 2(-2)^2 + 2(-1)^2 + 2(0)^2 + 2(1)^2 + 2(2)^2 + 2(3)^2}{14}}$$

$$= 2.$$

The standard deviation of data set C is

$$\sigma = \sqrt{\frac{5(-3)^2 + 4(-2)^2 + 3(-1)^2 + 2(0)^2 + 3(1)^2 + 4(2)^2 + 5(3)^2}{14}}$$

$$\approx 2.22.$$

These values confirm the results of Example 3. That is, data set A has the least standard deviation and data set C has the greatest.

The following alternative formula provides a more efficient way to compute the standard deviation.

Alternative Formula for Standard Deviation

The standard deviation of $\{x_1, x_2, \ldots, x_n\}$ is given by

$$\sigma = \sqrt{\frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n} - \bar{x}^2}.$$

Because of lengthy computations, this formula is difficult to verify. Conceptually, however, the process is straightforward. It consists of showing that the expressions

$$\sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}}$$

and

$$\sqrt{\frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n} - \bar{x}^2}$$

are equivalent. Try verifying this equivalence for the set $\{x_1, x_2, x_3\}$ with

$$\bar{x} = \frac{x_1 + x_2 + x_3}{3}.$$

Technology Tip

Most graphing utilities have *statistical* features that can be used to find different statistical values of data sets, such as the standard deviation. Enter data set A from Example 3 in the *list editor* of a graphing utility. Then use the *one-variable statistics* feature to verify the solution in Example 4, as shown below.

```
1-Var Stats
x̄=4
Σx=68
Σx²=312
Sx=1.58113883
σx=1.533929978
↓n=17
```

In the figure above, the standard deviation is represented as σx, which is about 1.53. For instructions on how to use the *one-variable statistics* feature, see Appendix A; for specific keystrokes, go to this textbook's *Companion Website*.

EXAMPLE 5   Using the Alternative Formula

Use the alternative formula for standard deviation to find the standard deviation of the following set of numbers.

5, 6, 6, 7, 7, 8, 8, 8, 9, 10

Solution

Begin by finding the mean of the set, which is 7.4. So, the standard deviation is

$$\sigma = \sqrt{\frac{5^2 + 2(6^2) + 2(7^2) + 3(8^2) + 9^2 + 10^2}{10} - (7.4)^2}$$

$$= \sqrt{\frac{568}{10} - 54.76}$$

$$\approx 1.43.$$

A well-known theorem in statistics, called *Chebychev's Theorem*, states that at least

$$1 - \frac{1}{k^2}$$

of the numbers in a distribution must lie within k standard deviations $(k > 1)$ of the mean. So, at least $\frac{3}{4}$ or 75% of the numbers in a collection must lie within two standard deviations of the mean, and at least $\frac{8}{9}$ or $88\frac{8}{9}\%$ of the numbers must lie within three standard deviations of the mean. For most distributions, these percents are low. For instance, in all three distributions shown in Example 3, 100% of the numbers lie within two standard deviations of the mean.

EXAMPLE 6   Describing a Distribution

The table at the right shows the estimated numbers of registered nurses (in thousands) in each state and the District of Columbia in 2013. Find the mean and standard deviation of the numbers. What percent of the numbers lie within two standard deviations of the mean? (Source: U.S. Bureau of Labor Statistics)

Solution

Begin by entering the numbers in a graphing utility. Then use the *one-variable statistics* feature to obtain

$$\bar{x} \approx 52.2 \quad \text{and} \quad \sigma \approx 52.0.$$

The interval that contains all numbers that lie within two standard deviations of the mean is

$$[52.2 - 2(52.0), 52.2 + 2(52.0)] \quad \text{or} \quad [-51.8, 156.2].$$

From the table, you can see that all but four of the numbers (about 92%) lie in this interval—all but the numbers that correspond to the numbers of registered nurses in California, Florida, New York, and Texas.

AK	6	MT	9
AL	44	NC	88
AR	23	ND	8
AZ	46	NE	20
CA	253	NH	12
CO	42	NJ	75
CT	35	NM	15
DC	11	NV	17
DE	10	NY	170
FL	163	OH	124
GA	66	OK	26
HI	10	OR	28
IA	32	PA	125
ID	12	RI	12
IL	109	SC	42
IN	60	SD	12
KS	27	TN	58
KY	42	TX	190
LA	41	UT	19
MA	79	VA	60
MD	46	VT	7
ME	14	WA	53
MI	92	WI	57
MN	58	WV	18
MO	65	WY	5
MS	28		

Spreadsheet at LarsonPrecalculus.com

Box–and–Whisker Plots

Standard deviation is the measure of dispersion that is associated with the mean. **Quartiles** measure dispersion associated with the median.

> **Definition of Quartiles**
>
> Consider an ordered set of numbers whose median is m. The **lower quartile** is the median of the numbers that occur before m. The **upper quartile** is the median of the numbers that occur after m.

EXAMPLE 7 Finding Quartiles of a Set

Find the lower and upper quartiles for the following data set.

34, 14, 24, 16, 12, 18, 20, 24, 16, 26, 13, 27

Solution

Begin by ordering the set.

12, 13, 14,	16, 16, 18,	20, 24, 24,	26, 27, 34
1st 25%	2nd 25%	3rd 25%	4th 25%

The median of the entire set is 19. The median of the six numbers that are before 19 is 15. So, the lower quartile is 15. The median of the six numbers that are after 19 is 25. So, the upper quartile is 25.

Quartiles are represented graphically by a **box-and-whisker plot,** as shown in Figure C.2. In the plot, notice that five numbers are given: the least number, the lower quartile, the median, the upper quartile, and the greatest number. Also notice that the numbers are spaced proportionally, as though they were on a real number line.
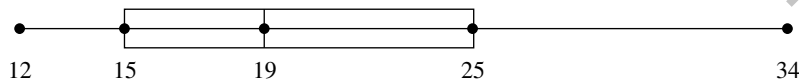

12 15 19 25 34

Figure C.2

Technology Tip

You can use a graphing utility to graph the box-and-whisker plot in Figure C.2. After entering the data in the graphing utility's *list editor*, use the *statistical plotting* and *ZoomStat* features to display the box-and-whisker plot, as shown in Figure C.3. For instructions on how to use the *list editor* and *statistical plotting* features, see Appendix A; for specific keystrokes, go to this textbook's *Companion Website*.
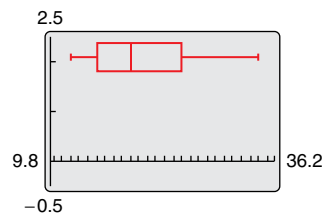

Figure C.3

The next example shows how to find quartiles when the number of elements in a set is not divisible by 4.

EXAMPLE 8   Sketching Box-and-Whisker Plots

Sketch a box-and-whisker plot for each data set.

a. 82, 82, 83, 85, 87, 89, 90, 94, 95, 95, 96, 98, 99

b. 11, 13, 13, 15, 17, 17, 20, 24, 24, 27

Solution

a. This set has 13 numbers. The median is 90 (the seventh number). The lower quartile is 84 (the median of the first six numbers). The upper quartile is 95.5 (the median of the last six numbers). See Figure C.4.
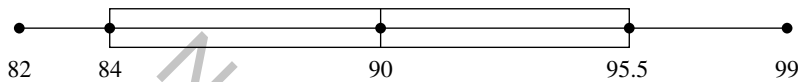

82   84   90   95.5   99

Figure C.4

b. This set has 10 numbers. The median is 17 (the average of the fifth and sixth numbers). The lower quartile is 13 (the median of the first five numbers). The upper quartile is 24 (the median of the last five numbers). See Figure C.5.

11   13   17   24   27

Figure C.5

C.1   Exercises

Vocabulary and Concept Check

In Exercises 1–4, fill in the blank(s).

1. A single number that is the most representative of a data set is called a _____ of _____ .

2. If two numbers are tied for the most frequent occurrence, then the collection has two _____ and is called _____ .

3. Two measures of dispersion are called the _____ and the _____ of a data set.

4. _____ measure dispersion associated with the median.

Procedures and Problem Solving

Comparing Measures of Central Tendency   **In Exercises 5–10, find the mean, median, and mode of the data set.**

5. 4, 12, 7, 16, 8, 9, 7, 5

6. 30, 37, 32, 39, 33, 34, 32

7. 4, 12, 7, 32, 8, 9, 7, 5

8. 17, 37, 32, 39, 33, 34, 32

9. 21, 4, 18, 6, 25, 3

10. 2, 10, 19, 3, 7, 5, 3, 10, 31, 11

11. **Exploration** Compare your answers in Exercise 5 with those in Exercise 7. Which of the measures of central tendency is sensitive to extreme measurements? Explain your reasoning.

12. **Exploration**

(a) Add 6 to each number in Exercise 8 and find the mean, median, and mode of the revised data set. How are the measures of central tendency changed?

(b) When a constant k is added to each number in a set of data, how will the measures of central tendency change?

13. **Sales** A car rental company kept the following record of the numbers of miles a rental car was driven. What are the mean, median, and mode of the data?

Monday	410	Tuesday	260
Wednesday	320	Thursday	320
Friday	460	Saturday	150

14. **Demographics** A study was done on families having six children. The table shows the numbers of families in the study with the indicated numbers of girls. Determine the mean, median, and mode of the data.

DATA	Number of girls	0	1	2	3	4	5	6
	Frequency	0	24	45	54	50	19	7

Spreadsheet at LarsonPrecalculus.com

15. **Athletics** The table shows the bowling scores of a three-member team for a three-game series.

DATA	Team member	Game 1	Game 2	Game 3
	Jay	181	222	196
	Hank	199	195	205
	Buck	202	251	235

Spreadsheet at LarsonPrecalculus.com

(a) Find the mean for each team member.

(b) Find the mean of the nine scores.

(c) Find the median of the nine scores.

(d) Which measure of central tendency best describes the nine scores?

16. **Sales** The selling prices of 12 new homes built in one subdivision are listed.

$525,000	$375,000	$425,000	$550,000
$385,000	$500,000	$550,000	$425,000
$475,000	$500,000	$350,000	$450,000

(a) Find the mean, mode, and median of the prices.

(b) Which measure of central tendency best describes the prices? Explain.

17. **Think About It** Construct a collection of numbers that has the following properties. If this is not possible, explain why.

Mean $= 6$, median $= 4$, mode $= 4$

18. **Think About It** Construct a collection of numbers that has the following properties. If this is not possible, explain why.

Mean $= 6$, median $= 6$, mode $= 4$

19. **Education** An English professor records the following scores for a 100-point exam.

99, 64, 80, 77, 59, 72, 87, 79, 92, 88, 90, 42, 20, 89, 42, 100, 98, 84, 78, 91

Which measure of central tendency best describes these test scores?

20. **Sales** A salesperson sold eight pairs of men's brown dress shoes. The sizes of the eight pairs were

$10\frac{1}{2}$, 8, 12, $10\frac{1}{2}$, 10, $9\frac{1}{2}$, 11, and $10\frac{1}{2}$.

Which measure (or measures) of central tendency best describes (describe) the typical shoe size for these data?

Finding Mean and Standard Deviation **In Exercises 21 and 22, line plots of data sets are given. Determine the mean and standard deviation of each data set.**

21. (a)

(b)

(c)

(d)

22. (a)

(b)

(c)

(d)
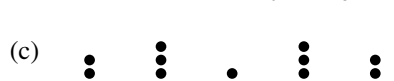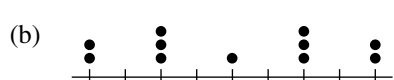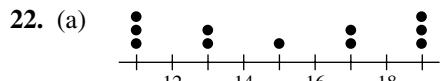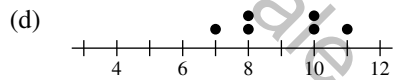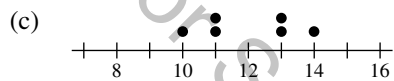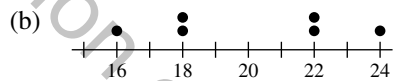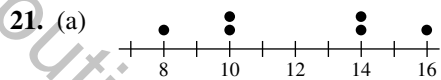
Finding Mean, Variance, and Standard Deviation   In Exercises 23–30, find the mean ($\bar{x}$), variance (v), and standard deviation (σ) of the data set.

23. 4, 10, 8, 2

24. 3, 15, 6, 9, 2

25. 1, 1, 1, 5, 5, 5

26. 2, 2, 2, 2, 2, 2

27. 0, 1, 1, 2, 2, 2, 3, 3, 4

28. 1, 2, 3, 4, 5, 6, 7

29. 49, 62, 40, 29, 32, 70

30. 1.5, 0.4, 2.1, 0.7, 0.8

Using the Alternative Formula   In Exercises 31–34, use the alternative formula to find the standard deviation of the data set.

31. 2, 4, 6, 6, 13, 5

32. 246, 336, 473, 167, 219, 359

33. 8.1, 6.9, 3.7, 4.2, 6.1

34. 9.0, 7.5, 3.3, 7.4, 6.0

35. Think About It   Without calculating the standard deviation, explain why the set {4, 4, 20, 20} has a standard deviation of 8.

36. Think About It   When the standard deviation of a set of numbers is 0, what does this imply about the set?

37. Education   An instructor adds five points to each student's exam score. Will this change the mean or standard deviation of the exam scores? Explain.

38. Exploration

(a) Multiply each number in Exercise 28 by 3 and find the mean, variance, and standard deviation of the revised data set. Compare these measures with what you found in Exercise 28. How did the measures change?

(b) When each number in a data set is multiplied by a constant k, how will the mean, variance, and standard deviation change?

39. *Why you should learn it*   (p. C1) The following data represent the average prices of gold (in dollars per ounce) for the years 1990 through 2013. Use a graphing utility to find the mean, variance, and standard deviation of the data. What percent of the data lie within two standard deviations of the mean?   (Source: Austin Rare Coins, Inc.)

384,	362,	344,	360,	384,	384,
388,	331,	294,	279,	279,	271,
310,	363,	410,	445,	603,	695,
872,	972,	1225,	1568,	1669,	1531

40. Athletics   The total number of points scored by the 437 players who played in the 2013–2014 NBA regular season had a mean and standard deviation of 563 and 468, respectively. Use Chebychev's Theorem to determine the intervals containing at least $\frac{3}{4}$ and at least $\frac{8}{9}$ of the point totals. How do the intervals change when the standard deviation is 225?   (Source: National Basketball Association)

Sketching a Box-and-Whisker Plot   In Exercises 41–44, (a) find the lower and upper quartiles of the data and (b) sketch a box-and-whisker plot for the data without using a graphing utility.

41. 23, 15, 14, 23, 13, 14, 13, 20, 12

42. 11, 10, 11, 14, 17, 16, 14, 11, 8, 14, 20

43. 46, 48, 48, 50, 52, 47, 51, 47, 49, 53

44. 25, 20, 22, 28, 24, 28, 25, 19, 27, 29, 28, 21

Creating a Box-and-Whisker Plot   In Exercises 45–48, use a graphing utility to create a box-and-whisker plot for the data.

45. 19, 12, 14, 9, 14, 15, 17, 13, 19, 11, 10, 19

46. 9, 5, 5, 5, 6, 5, 4, 12, 7, 10, 7, 11, 8, 9, 9

47. 20.1, 43.4, 34.9, 23.9, 33.5, 24.1, 22.5, 42.4, 25.7, 17.4, 23.8, 33.3, 17.3, 36.4, 21.8

48. 78.4, 76.3, 107.5, 78.5, 93.2, 90.3, 77.8, 37.1, 97.1, 75.5, 58.8, 65.6

49. Think About It   The histograms represent the test scores of two classes of a college course in mathematics. Which histogram has the smaller standard deviation?
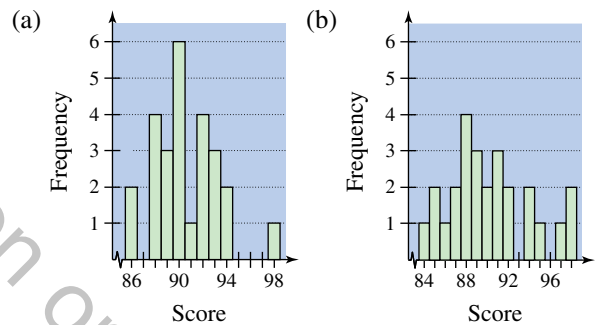
50. Manufacturing   A company has redesigned a product in an attempt to increase the lifetime of the product. The two sets of data list the lifetimes (in months) of 20 units with the original design and 20 units with the new design. Create a box-and-whisker plot for each set of data, and then comment on the differences between the plots.

Original Design

15.1	78.3	56.3	68.9	30.6
27.2	12.5	42.7	72.7	20.2
53.0	13.5	11.0	18.4	85.2
10.8	38.3	85.1	10.0	12.6

New Design

55.8	71.5	25.6	19.0	23.1
37.2	60.0	35.3	18.9	80.5
46.7	31.1	67.9	23.5	99.5
54.0	23.2	45.5	24.8	87.8

C.2 Least Squares Regression

In many of the examples and exercises in this text, you have been asked to use the *regression* feature of a graphing utility to find mathematical models for sets of data. The *regression* feature of a graphing utility uses the **method of least squares** to find a mathematical model for a set of data. As a measure of how well a model fits a set of data points

$$\{(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)\}$$

you can add the squares of the differences between the actual y-values and the values given by the model to obtain the **sum of the squared differences.** For instance, the table shows several values of x and y. The table also shows the values of a linear model $y^* = 0.21x - 2.0$ for each x-value. The sum of the squared differences for the model is 28.0178.

x	70	72	75	76	77	78	80
y	10.9	11.2	11.7	12	12.2	12.3	12.6
y^*	12.70	13.12	13.75	13.96	14.17	14.38	14.80
$(y - y^*)^2$	3.24	3.6864	4.2025	3.8416	3.8809	4.3264	4.84

The model that has the *least* sum of the squared differences is the **least squares regression** line for the data. The least squares regression line for the data in the table is $y = 0.18x - 1.5$. The sum of the squared differences is 0.4032.

To find the least squares regression line $y = ax + b$ for the points $\{(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)\}$ algebraically, you need to solve the following system for a and b.

$$\begin{cases} nb + \left(\sum_{i=1}^{n} x_i\right)a = \sum_{i=1}^{n} y_i \\ \left(\sum_{i=1}^{n} x_i\right)b + \left(\sum_{i=1}^{n} x_i^2\right)a = \sum_{i=1}^{n} x_i y_i \end{cases}$$

In the system,

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$$

$$\sum_{i=1}^{n} y_i = y_1 + y_2 + \cdots + y_n$$

$$\sum_{i=1}^{n} x_i^2 = x_1^2 + x_2^2 + \cdots + x_n^2$$

$$\sum_{i=1}^{n} x_i y_i = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n.$$

What you should learn

► Use the sum of squared differences to determine a least squares regression line.
► Find a least squares regression line for a set of data.
► Find a least squares regression parabola for a set of data.

Why you should learn it

The method of least squares provides a way of creating a mathematical model for a set of data, which can then be analyzed.

Technology Tip

Recall from Section 1.7 that when you use the *regression* feature of a graphing utility, the program may output a correlation coefficient, r. When $|r|$ is close to 1, the model is a good fit for the data.

EXAMPLE 1   Finding a Least Squares Regression Line

Find the least squares regression line for $(-3, 0)$, $(-1, 1)$, $(0, 2)$, and $(2, 3)$.

Solution

Begin by constructing a table, as shown below.

x	y	xy	x^2
-3	0	0	9
-1	1	-1	1
0	2	0	0
2	3	6	4
$\sum_{i=1}^{n} x_i = -2$	$\sum_{i=1}^{n} y_i = 6$	$\sum_{i=1}^{n} x_i y_i = 5$	$\sum_{i=1}^{n} x_i^2 = 14$

Applying the system for the least squares regression line with $n = 4$ produces

$$\begin{cases} nb + \left(\sum_{i=1}^{n} x_i\right)a = \sum_{i=1}^{n} y_i \\ \left(\sum_{i=1}^{n} x_i\right)b + \left(\sum_{i=1}^{n} x_i^2\right)a = \sum_{i=1}^{n} x_i y_i \end{cases} \implies \begin{cases} 4b - 2a = 6 \\ -2b + 14a = 5 \end{cases}.$$

Solving this system of equations produces $a = \frac{8}{13}$ and $b = \frac{47}{26}$. So, the least squares regression line is $y = \frac{8}{13}x + \frac{47}{26}$, as shown in Figure C.6.
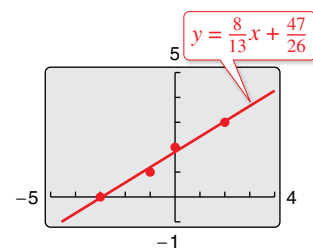

$y = \frac{8}{13}x + \frac{47}{26}$

Figure C.6

The least squares regression parabola $y = ax^2 + bx + c$ for the points

$$\{(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)\}$$

is obtained in a similar manner by solving the following system of three equations in three unknowns for a, b, and c.

$$\begin{cases} nc + \left(\sum_{i=1}^{n} x_i\right)b + \left(\sum_{i=1}^{n} x_i^2\right)a = \sum_{i=1}^{n} y_i \\ \left(\sum_{i=1}^{n} x_i\right)c + \left(\sum_{i=1}^{n} x_i^2\right)b + \left(\sum_{i=1}^{n} x_i^3\right)a = \sum_{i=1}^{n} x_i y_i \\ \left(\sum_{i=1}^{n} x_i^2\right)c + \left(\sum_{i=1}^{n} x_i^3\right)b + \left(\sum_{i=1}^{n} x_i^4\right)a = \sum_{i=1}^{n} x_i^2 y_i \end{cases}$$

C.2   Exercises

Finding a Least Squares Regression Line   In Exercises 1–4, find the least squares regression line for the points. Verify your answer with a graphing utility.

1. $(-4, 1)$, $(-3, 3)$, $(-2, 4)$, $(-1, 6)$

2. $(0, -1)$, $(2, 0)$, $(4, 3)$, $(6, 5)$

3. $(-3, 1)$, $(-1, 2)$, $(1, 2)$, $(4, 3)$

4. $(0, -1)$, $(2, 1)$, $(3, 2)$, $(5, 3)$