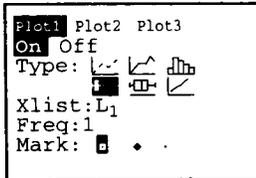
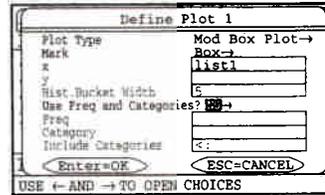


TECHNOLOGY TOOLBOX *Calculator boxplots and numerical summaries (continued)*

TI-83

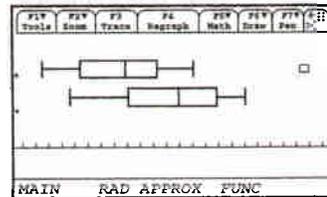
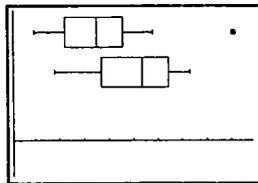


TI-89



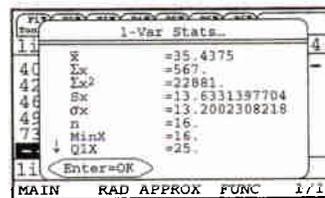
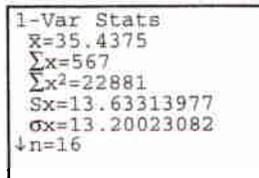
3. Use the calculator's zoom feature to display the side-by-side boxplots.

- Press **ZOOM** and select 9:ZoomStat.
- Press **F5** (ZoomData).



4. Calculate numerical summaries for each set of data.

- Press **STAT** (CALC) and select 1:1-Var Stats
- Press **F4** (Calc) and choose 1:1-Var Stats.
- Press **ENTER**. Now press **2nd** **1** (L1) and **ENTER**.
- Type list1 in the list box. Press **ENTER**.



5. Notice the down arrow on the left side of the display. Press **▼** to see Bonds's other statistics. Repeat the process to find the Babe's numerical summaries.

EXERCISES

1.36 SSHA SCORES Here are the scores on the Survey of Study Habits and Attitudes (SSHA) for 18 first-year college women:

154 109 137 115 152 140 154 178 101 103 126 126 137 165 165 129 200 148

and for 20 first-year college men:

108 140 114 91 180 115 126 92 169 146 109 132 75 88 113 151 70 115 187 104

- (a) Make side-by-side boxplots to compare the distributions.

- (b) Compute numerical summaries for these two distributions.
 (c) Write a paragraph comparing the SSHA scores for men and women.

1.37 HOW OLD ARE PRESIDENTS? Return to the data on presidential ages in Table 1. (page 19). In Example 1.6, we constructed a histogram of the age data.

- (a) From the shape of the histogram (Figure 1.7, page 20), do you expect the mean to be much less than the median, about the same as the median, or much greater than the median? Explain.
 (b) Find the five-number summary and verify your expectation from (a).
 (c) What is the range of the middle half of the ages of new presidents?
 (d) Construct by hand a (modified) boxplot of the ages of new presidents.
 (e) On your calculator, define Plot 1 to be a histogram using the list named PREZ that you created in the Technology Toolbox on page 22. Define Plot 2 to be a (modified) boxplot also using the list PREZ. Use the calculator's zoom command to generate a graph. To remove the overlap, adjust your viewing window so that Ymin = -6 and Ymax = 22. Then graph. Use TRACE to inspect values. Press the up and down cursor keys to toggle between plots. Is there an outlier? If so, who was it?

1.38 Is the interquartile range a resistant measure of spread? Give an example of small data set that supports your answer.

1.39 SHOPPING SPREE, III Figure 1.19 displays computer output for the data on amount spent by grocery shoppers in Exercise 1.11 (page 18).

- (a) Find the total amount spent by the shoppers.
 (b) Make a boxplot from the computer output. Did you check for outliers?

DataDesk

```
Summary of spending
No Selector
Percentile 25
Count 50
Mean 34.7022
Median 27.8550
StdDev 21.6974
Min 3.11000
Max 93.3400
Lower 10th %tile 19.2700
Upper 10th %tile 45.4000
```

Minitab

```
Descriptive Statistics
Variable N Mean Median TrMean StDev SEMean
spending 50 34.70 27.85 32.92 21.70 3.07
Variable Min Max Q1 Q3
spending 3.11 93.34 19.06 45.72
```

FIGURE 1.19 Numerical descriptions of the unrounded shopping data from the Data Desk and Minitab software.

Measuring spread: the standard deviation

The five-number summary is not the most common numerical description of a distribution. That distinction belongs to the combination of the mean to measure center and the *standard deviation* to measure spread. The standard deviation measures spread by looking at how far the observations are from their mean.

THE STANDARD DEVIATION s

The **variance** s^2 of a set of observations is the average of the squares of the deviations of the observations from their mean. In symbols, the variance of n observations x_1, x_2, \dots, x_n is

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1}$$

or, more compactly,

$$s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$$

The **standard deviation** s is the square root of the variance s^2 :

$$s = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$$

In practice, use software or your calculator to obtain the standard deviation from keyed-in data. Doing a few examples step-by-step will help you understand how the variance and standard deviation work, however. Here is such an example.

EXAMPLE 1.14 METABOLIC RATE

A person's metabolic rate is the rate at which the body consumes energy. Metabolic rate is important in studies of weight gain, dieting, and exercise. Here are the metabolic rates of 7 men who took part in a study of dieting. (The units are calories per 24 hours. These are the same calories used to describe the energy content of foods.)

1792	1666	1362	1614	1460	1867	1439
------	------	------	------	------	------	------

The researchers reported \bar{x} and s for these men.

First find the mean:

$$\bar{x} = \frac{1792 + 1666 + 1362 + 1614 + 1460 + 1867 + 1439}{7} = \frac{11,200}{7} = 1600 \text{ calories}$$

To see clearly the nature of the variance, start with a table of the deviations of the observations from this mean.

Observations x_i	Deviations $x_i - \bar{x}$	Squared deviations $(x_i - \bar{x})^2$
1792	$1792 - 1600 = 192$	$192^2 = 36,864$
1666	$1666 - 1600 = 66$	$66^2 = 4,356$
1362	$1362 - 1600 = -238$	$(-238)^2 = 56,644$
1614	$1614 - 1600 = 14$	$14^2 = 196$
1460	$1460 - 1600 = -140$	$(-140)^2 = 19,600$
1867	$1867 - 1600 = 267$	$267^2 = 71,289$
1439	$1439 - 1600 = -161$	$(-161)^2 = 25,921$
	sum = 0	sum = 214,870

The variance is the sum of the squared deviations divided by one less than the number of observations:

$$s^2 = \frac{214,870}{6} = 35,811.67$$

The standard deviation is the square root of the variance:

$$s = \sqrt{35,811.67} = 189.24 \text{ calories}$$

Compare these results for s^2 and s with those generated by your calculator or computer.

Figure 1.20 displays the data of Example 1.14 as points above the number line, with their mean marked by an asterisk (*). The arrows show two of the deviations from the mean. These deviations show how spread out the data are about their mean. Some of the deviations will be positive and some will be negative because observations fall on each side of the mean. In fact, *the sum of the deviations of the observations from their mean will always be zero*. Check that this is true in Example 1.14. So we cannot simply add the deviations to get an overall measure of spread. Squaring the deviations makes them all nonnegative, so that observations far from the mean in either direction will have large positive squared deviations. The variance s^2 is the average squared deviation. The variance is large if the observations are widely spread about their mean; it is small if the observations are all close to the mean.

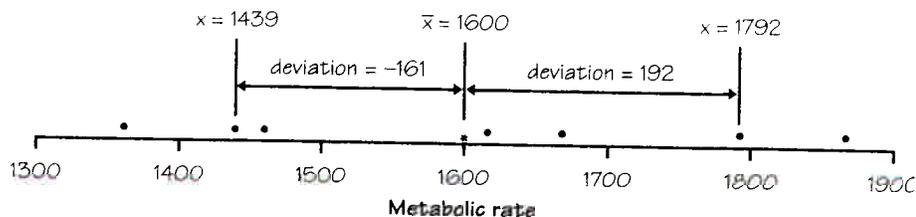


FIGURE 1.20 Metabolic rates for seven men, with their mean (*) and the deviations of two observations from the mean.

Because the variance involves squaring the deviations, it does not have the same unit of measurement as the original observations. Lengths measured in centimeters, for example, have a variance measured in squared centimeters. Taking the square root remedies this. The standard deviation s measures spread about the mean in the original scale.

If the variance is the average of the squares of the deviations of the observations from their mean, why do we average by dividing by $n - 1$ rather than n ? Because the sum of the deviations is always zero, the last deviation can be found once we know the other $n - 1$ deviations. So we are not averaging n unrelated numbers. Only $n - 1$ of the squared deviations can vary freely, and we average by dividing the total by $n - 1$. The number $n - 1$ is called the *degrees of freedom* of the variance or of the standard deviation. Many calculators offer a choice between dividing by n and dividing by $n - 1$, so be sure to use $n - 1$.

degrees of freedom

Leaving the arithmetic to a calculator allows us to concentrate on what we are doing and why. What we are doing is measuring spread. Here are the basic properties of the standard deviation s as a measure of spread.

PROPERTIES OF THE STANDARD DEVIATION

- s measures spread about the mean and should be used only when the mean is chosen as the measure of center.
- $s = 0$ only when there is *no spread*. This happens only when all observations have the same value. Otherwise, $s > 0$. As the observations become more spread out about their mean, s gets larger.
- s , like the mean \bar{x} , is not resistant. Strong skewness or a few outliers can make s very large. For example, the standard deviation of Barry Bonds's home run counts is 13.633. (Use your calculator to verify this.) If we omit the outlier, the standard deviation drops to 9.573.

You may rightly feel that the importance of the standard deviation is not yet clear. We will see in the next chapter that the standard deviation is the natural measure of spread for an important class of symmetric distributions, the normal distributions. The usefulness of many statistical procedures is tied to distributions of particular shapes. This is certainly true of the standard deviation.

Choosing measures of center and spread

How do we choose between the five-number summary and \bar{x} and s to describe the center and spread of a distribution? Because the two sides of a strongly *skewed distribution have different spreads, no single number such as s* describes the spread well. The five-number summary, with its two quartiles and two extremes, does a better job.

CHOOSING A SUMMARY

The five-number summary is usually better than the mean and standard deviation for describing a skewed distribution or a distribution with strong outliers. Use \bar{x} and s only for reasonably symmetric distributions that are free of outliers.

Do remember that a graph gives the best overall picture of a distribution. Numerical measures of center and spread report specific facts about a distribution, but they do not describe its entire shape. Numerical summaries do not disclose the presence of multiple peaks or gaps, for example. **Always plot your data.**

EXERCISES

1.40 PHOSPHATE LEVELS The level of various substances in the blood influences our health. Here are measurements of the level of phosphate in the blood of a patient, in milligrams of phosphate per deciliter of blood, made on 6 consecutive visits to a clinic:

5.6	5.2	4.6	4.9	5.7	6.4
-----	-----	-----	-----	-----	-----

A graph of only 6 observations gives little information, so we proceed to compute the mean and standard deviation.

- Find the mean from its definition. That is, find the sum of the 6 observations and divide by 6.
- Find the standard deviation from its definition. That is, find the deviations of each observation from the mean, square the deviations, then obtain the variance and the standard deviation. Example 1.14 shows the method.
- Now enter the data into your calculator to obtain \bar{x} and s . Do the results agree with your hand calculations? Can you find a way to compute the standard deviation without using one-variable statistics?

1.41 ROGER MARIS New York Yankee Roger Maris held the single-season home run record from 1961 until 1998. Here are Maris's home run counts for his 10 years in the American League:

15	28	16	39	61	33	23	26	8	13
----	----	----	----	----	----	----	----	---	----

- Maris's mean number of home runs is $\bar{x} = 26.2$. Find the standard deviation s from its definition. Follow the model of Example 1.14.
- Use your calculator to verify your results. Then use your calculator to find \bar{x} and s for the 9 observations that remain when you leave out the outlier. How does the outlier affect the values of \bar{x} and s ? Is s a resistant measure of spread?

1.42 OLDER FOLKS, III In Exercise 1.12 (page 22), you made a histogram displaying the percentage of residents aged 65 or older in each of the 50 U.S. states. Do you prefer the five-number summary or \bar{x} and s as a brief numerical description? Why? Calculate your preferred description.

1.43 This is a standard deviation contest. You must choose four numbers from the whole numbers 0 to 10, with repeats allowed.

- Choose four numbers that have the smallest possible standard deviation.
- Choose four numbers that have the largest possible standard deviation.
- Is more than one choice possible in either (a) or (b)? Explain.

Changing the unit of measurement

The same variable can be recorded in different units of measurement. Americans commonly record distances in miles and temperatures in degrees Fahrenheit. Most of the rest of the world measures distances in kilometers and temperatures in degrees Celsius. Fortunately, it is easy to convert from one unit of measurement to another. In doing so, we perform a *linear transformation*.

LINEAR TRANSFORMATION

A linear transformation changes the original variable x into the new variable x_{new} given by an equation of the form

$$x_{\text{new}} = a + bx$$

Adding the constant a shifts all values of x upward or downward by the same amount.

Multiplying by the positive constant b changes the size of the unit of measurement.

EXAMPLE 1.15 LOS ANGELES LAKERS' SALARIES

Table 1.8 gives the approximate base salaries of the 14 members of the Los Angeles Lakers basketball team for the year 2000. You can calculate that the mean is $\bar{x} = \$4.14$ million and that the median is $M = \$2.6$ million. No wonder professional basketball players have big houses!

TABLE 1.8 Year 2000 salaries for the Los Angeles Lakers

Player	Salary	Player	Salary
Shaquille O'Neal	\$17.1 million	Ron Harper	\$2.1 million
Kobe Bryant	\$11.8 million	A. C. Green	\$2.0 million
Robert Horry	\$5.0 million	Devean George	\$1.0 million
Glen Rice	\$4.5 million	Brian Shaw	\$1.0 million
Derek Fisher	\$4.3 million	John Salley	\$0.8 million
Rick Fox	\$4.2 million	Tyronne Lue	\$0.7 million
Travis Knight	\$3.1 million	John Celestand	\$0.3 million

Figure 1.21(a) is a stemplot of the salaries, with millions as stems. The distribution is skewed to the right and there are two high outliers. The very high salaries of Kobe Bryant and Shaquille O'Neal pull up the mean. Use your calculator to check that the mean is \$4.76 million, and that the five-number summary is

\$0.3 million \$1.0 million \$2.6 million \$4.5 million \$17.1 million

(a) Suppose that each member of the team receives a \$100,000 bonus for winning the NBA Championship (which the Lakers did in 2000). How will this affect the shape, center, and spread of the distribution?

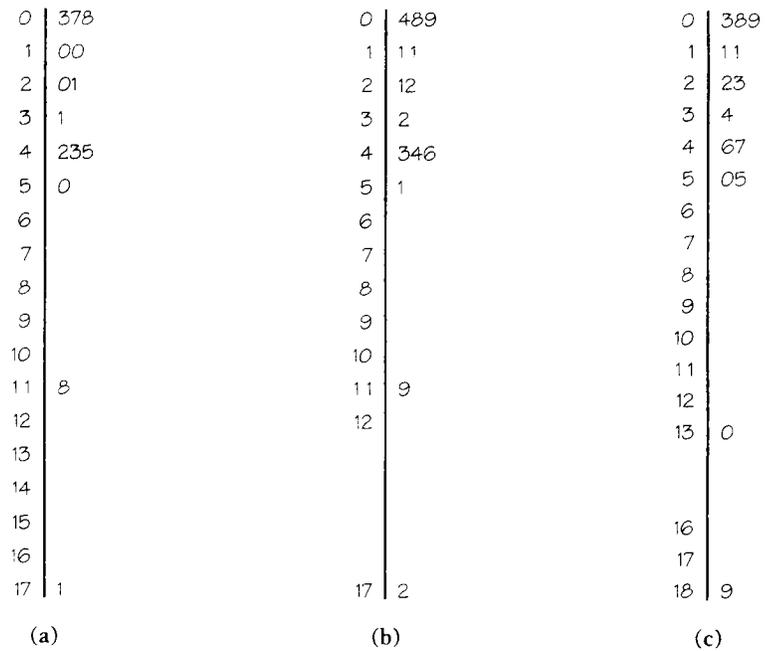


FIGURE 1.21 Stemplots of the salaries of Los Angeles Lakers players, from Table 1.8.

Since \$100,000 = \$0.1 million, each player's salary will increase by \$0.1 million. This linear transformation can be represented by $x_{\text{new}} = 0.1 + 1x$, where x_{new} is the salary after the bonus and x is the player's base salary. Increasing each value in Table 1.8 by 0.1 will also increase the mean by 0.1. That is, $\bar{x}_{\text{new}} = \$4.2$ million. Likewise, the median salary will increase by 0.1 and become $M = \$2.7$ million.

What will happen to the spread of the distribution? The standard deviation of the Lakers' salaries after the bonus is still $s = \$4.76$ million. With the bonus, the five-number summary becomes

\$0.4 million	\$1.1 million	\$2.7 million	\$4.6 million	\$17.2 million
---------------	---------------	---------------	---------------	----------------

Both before and after the salary bonus, the *IQR* for this distribution is \$3.5 million. Adding a constant amount to each observation does not change the spread. The shape of the distribution remains unchanged, as shown in Figure 1.21(b).

(b) Suppose that, instead of receiving a \$100,000 bonus, each player is offered a 10% increase in his base salary. John Celestand, who is making a base salary of \$0.3 million, would receive an additional $(0.10)(\$0.3 \text{ million}) = \0.03 million . To obtain his new salary, we could have used the linear transformation $x_{\text{new}} = 0 + 1.10x$, since multiplying the current salary (x) by 1.10 increases it by 10%. Increasing all 14 players' salaries in the same way results in the following list of values (in millions):

\$0.33	\$0.77	\$0.88	\$1.10	\$1.10	\$2.20	\$2.31
\$3.41	\$4.62	\$4.73	\$4.95	\$5.50	\$12.98	\$18.81

Use your calculator to check that $\bar{x}_{\text{new}} = \$4.55 \text{ million}$, $s_{\text{new}} = \$5.24 \text{ million}$, $M_{\text{new}} = \$2.86 \text{ million}$, and the five-number summary for x_{new} is

\$0.33	\$1.10	\$2.86	\$4.95	\$18.81
--------	--------	--------	--------	---------

Since $\$4.14(1.10) = \4.55 and $\$2.6(1.10) = \2.86 , you can see that both measures of center (the mean and median) have increased by 10%. This time, the spread of the distribution has increased, too. Check for yourself that the standard deviation and the *IQR* have also increased by 10%. The stemplot in Figure 1.21(c) shows that the distribution of salaries is still right-skewed.

Linear transformations do not change the shape of a distribution. As you saw in the previous example, changing the units of measurement can affect the center and spread of the distribution. Fortunately, the effects of such changes follow a simple pattern.

EFFECT OF A LINEAR TRANSFORMATION

To see the effect of a linear transformation on measures of center and spread, apply these rules:

- Multiplying each observation by a positive number b multiplies both measures of center (mean and median) and measures of spread (standard deviation and *IQR*) by b .
- Adding the same number a (either positive or negative) to each observation adds a to measures of center and to quartiles but does not change measures of spread.

EXERCISES

1.44 COCKROACHES! Maria measures the lengths of 5 cockroaches that she finds at her school. Here are her results (in inches):

1.4	2.2	1.1	1.6	1.2
-----	-----	-----	-----	-----

- Find the mean and standard deviation of Maria's measurements.
- Maria's science teacher is furious to discover that she has measured the cockroach lengths in inches rather than centimeters. (There are 2.54 cm in 1 inch.) She gives Maria two minutes to report the mean and standard deviation of the 5 cockroaches in centimeters. Maria succeeded. Will you?
- Considering the 5 cockroaches that Maria found as a small sample from the population of all cockroaches at her school, what would you estimate as the average length of the population of cockroaches? How sure of your estimate are you?

1.45 RAISING TEACHERS' PAY A school system employs teachers at salaries between \$30,000 and \$60,000. The teachers' union and the school board are negotiating the form of next year's increase in the salary schedule. Suppose that every teacher is given a flat \$1000 raise.

- How much will the mean salary increase? The median salary?
- Will a flat \$1000 raise increase the spread as measured by the distance between the quartiles?
- Will a flat \$1000 raise increase the spread as measured by the standard deviation of the salaries?

1.46 RAISING TEACHERS' PAY, II Suppose that the teachers in the previous exercise each receive a 5% raise. The amount of the raise will vary from \$1500 to \$3000, depending on present salary. Will a 5% across-the-board raise increase the spread of the distribution as measured by the distance between the quartiles? Do you think it will increase the standard deviation?

Comparing distributions

An experiment is carried out to compare the effectiveness of a new cholesterol-reducing drug with the one that is currently prescribed by most doctors. A survey is conducted to determine whether the proportion of males who are likely to vote for a political candidate is higher than the proportion of females who are likely to vote for the candidate. Students taking AP Calculus AB and AP Statistics are curious about which exam is harder. They have information on the distribution of scores earned on each exam from the year 2000. In each of these situations, we are interested in comparing distributions. This section presents some of the more common methods for making statistical comparisons.