Describing Distributions with Numbers

12

FRANCES M. ROBERTS/Newscom

CASE STUDY Does education pay? We are told that people with more education earn more on the average than people with less education. How much more? How can we answer this question?

Data on income can be found at the Census Bureau website. The data are estimates, for the year 2013, of the total incomes of 136,641,000 people aged 25 and over with earnings and are based on the results of the Current Population Survey in 2014. The website gives the income distribution for each of several education categories. In particular, it gives the number of people in each of several education categories who earned between $1 and $2499, between $2500 and $4999, up to between $97,500 and $99,999, and $100,000 and over. That is a lot of information. A histogram could be used to display the data, but are there simple ways to summarize the information with just a few numbers that allow us to make sensible comparisons?

In this chapter, we will learn several ways to summarize large data sets with a few numbers. By the end of this chapter, with these new methods for summarizing large data sets, you will be able to provide an answer to whether education really pays.

Baseball has a rich tradition of using statistics to summarize and characterize the performance of players. We begin by investigating ways to summarize the performance of the greatest home run hitters of all time.

In the summer of 2007, Barry Bonds shattered the career home run record, breaking the previous record set by Hank Aaron. Here are his home run counts for the years 1986 (his rookie year) to 2007 (his final season):

1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996
16	25	24	19	33	25	34	46	37	33	42
1997	**1998**	**1999**	**2000**	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**
40	37	34	49	73	46	45	45	5	26	28

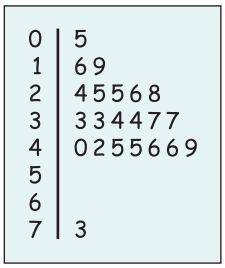


Figure 12.1 Stemplot of the number of home runs hit by Barry Bonds in his 22-year career.

The stemplot in Figure 12.1 displays the data. The shape of the distribution is a bit irregular, but we see that it has one high outlier, and if we ignore this outlier, we might describe it as slightly skewed to the left with a single peak. The outlier is, of course, Bonds's record season in 2001.

A graph and a few words give a good description of Barry Bonds's home run career. But words are less adequate to describe, for example, the incomes of people with a high school education. We need *numbers* that summarize the center and variability of a distribution.

Median and quartiles

A simple and effective way to describe center and variability is to give the **median** and the **quartiles**. The median is the midpoint, the value that separates the smaller half of the observations from the larger half. The first and third quartiles mark off the middle half of the observations. The quartiles get their name because with the median they divide the observations into quarters—one-quarter of the observations lie below the first quartile, half lie below the median, and three-quarters lie below the third quartile. That's the idea. To actually get numbers, we need a rule that makes the idea exact.

EXAMPLE 1 Finding the median

We might compare Bonds's career with that of Hank Aaron, the previous holder of the career record. Here are Aaron's home run counts for his 23 years in baseball:

13 27 26 44 30 39 40 34 45 44 24 32
44 39 29 44 38 47 34 40 20 12 10

To find the median, first arrange them in order from smallest to largest:

10 12 13 20 24 26 27 29 30 32 34 **34**
38 39 39 40 40 44 44 44 44 45 47

The bold 34 is the center observation, with 11 observations to its left and 11 to its right. When the number of observations n is odd (here $n = 23$), there is always one observation in the center of the ordered list. This is the median, $M = 34$.

How does this compare with Bonds's record? Here are Bonds's 22 home run counts, arranged in order from smallest to largest:

5	16	19	24	25	25	26	28	33	33	**34**
34	37	37	40	42	45	45	46	46	49	73

When n is even, there is no one middle observation. But there is a middle pair—the bold 34 and 34 have 10 observations on either side. We take the median to be halfway between this middle pair. So Bonds's median is

$$M = \frac{34 + 34}{2} = \frac{68}{2} = 34$$

There is a fast way to locate the median in an ordered list: count up $(n + 1)/2$ places from the beginning of the list. Try it. For Aaron, $n = 23$ and $(23 + 1)/2 = 12$, so the median is the 12th entry in the ordered list. For Bonds, $(n = 22)$ and $(22 + 1)/2 = 11.5$. This means "halfway between the 11th and 12th" entries, so M is the average of these two entries. This "$(n + 1)/2$ rule" is especially handy when you have many observations. The median of $n = 46{,}940$ incomes is halfway between the 23,470th and 23,471st in the ordered list. Be sure to note that $(n + 1)/2$ does *not* give the median M, just its position in the ordered list of observations.

The median *M*

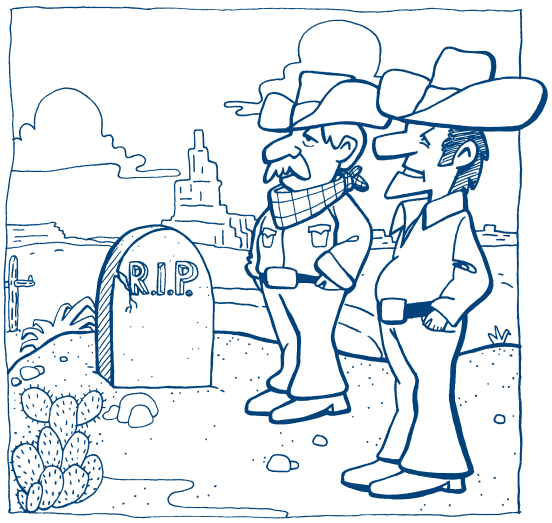
The **median** M is the midpoint of a distribution, the number such that half the observations are smaller and the other half are larger. To find the median of a distribution:

1. Arrange all observations in order of size, from smallest to largest.

2. If the number of observations n is odd, the median M is the center observation in the ordered list. Find the location of the median by counting $(n + 1)/2$ observations up from the bottom of the list.

3. If the number of observations n is even, the median M is the average of the two center observations in the ordered list. The location of the median is again $(n + 1)/2$ from the bottom of the list.

The Census Bureau website provides data on income inequality. For example, it tells us that in 2013 the median income of Hispanic



"Yup, Old Bob drowned due to being ignorant of statistics. He thought it was enough to know the average depth of the river."

households was $40,963. That's helpful but incomplete. Do most Hispanic households earn close to this amount, or are the incomes very variable? The simplest useful description of a distribution consists of both a measure of *center* and a measure of *variability*. If we choose the median (the midpoint) to describe center, the quartiles (in particular, the difference between the quartiles) provide natural descriptions of variability. Again, the idea is clear: find the points one-quarter and three-quarters up the ordered list of observations. Again, we need a rule to make the idea precise. The rule for calculating the quartiles uses the rule for the median.

The quartiles Q_1 and Q_3

To calculate the **quartiles**:

1. Arrange the observations in increasing order and locate the median M in the ordered list of observations.

2. The **first quartile** Q_1 is the median of the observations whose position in the ordered list is to the left of the location of the overall median. The overall median is not included in the observations considered to be to the left of the overall median.

3. The **third quartile** Q_3 is the median of the observations whose position in the ordered list is to the right of the location of the overall median. The overall median is not included in the observations considered to be to the right of the overall median.

EXAMPLE 2 Finding the quartiles

Hank Aaron's 23 home run counts are

10 12 13 20 24 26 27 29 30 32 32 **34** 38
39 39 40 40 44 44 44 44 45 47

Q_1 = 26, M = 34, Q_3 = 44

There is an odd number of observations, so the median is the one in the middle, the bold 34 in the list. To find the quartiles, ignore this central observation. The first quartile is the median of the 11 observations to the left of the bold 34 in the list. That's the sixth, so $Q_1 = 26$. The third quartile is the median of the 11 observations to the right of the bold 34. It is $Q_3 = 44$.

Barry Bonds's 22 home run counts are

5	16	19	24	25	25	26	28	33	33	**34**	**34**
					↑					↑	
					Q_1					M	

37	37	40	42	45	45	46	46	49	73
				↑					
				Q_3					

The median lies halfway between the middle pair. There are 11 observations to the left of this location. The first quartile is the median of these 11 numbers. That's the sixth, so $Q_1 = 25$. The third quartile is the median of the 11 observations to the right of the overall median's location, $Q_3 = 45$.

NOW IT'S YOUR TURN

12.1 Babe Ruth. Prior to Hank Aaron, Babe Ruth was the holder of the career record. Here are Ruth's home run counts for his 22 years in Major League Baseball, arranged in order from smallest to largest:

0	2	3	4	6	11	22	25	29	34	35
41	41	46	46	46	47	49	54	54	59	60

Find the median, first quartile, and third quartile of these counts.

You can use the $(n + 1)/2$ rule to locate the quartiles when there are many observations. The Census Bureau website tells us that there were 15,811,000 (rounded off to the nearest 1000) Hispanic households in the United States in 2013. If we ignore the roundoff, the median of these 15,811,000 incomes is halfway between the 7,905,500th and 7,905,501st in the list arranged in order from smallest to largest. So the first quartile is the median of the 7,905,500 incomes below this point in the list. Use the $(n + 1)/2$ rule with $n = 7{,}905{,}500$ to locate the quartile:

$$\frac{n+1}{2} = \frac{7{,}905{,}500 + 1}{2} = 3{,}952{,}750.5$$

The average of the 3,952,750th and 3,952,751st incomes in the ordered list falls in the range \$20,000 to \$24,999, and we estimate the first quartile to be \$21,621.

The third quartile is the median of the 7,905,500 incomes above the median. By the $(n + 1)/2$ rule with 7,905,500, this will be the average of the 3,952,750th and 3,952,751st incomes above the median in the ordered list. We find that this falls in the range \$65,000 to \$69,999 and we estimate the third quartile to be \$67,660.

In practice, people use statistical software to compute quartiles. Software can give results that differ from those you will obtain using the method described here. In fact, different software packages use slightly different rules for deciding how to divide the space between two adjacent values between which the quartile is believed to lie. We have chosen to select the point halfway between them, but other rules exist. Two different software packages can give two slightly different answers, depending on the rule employed.

The five-number summary and boxplots

The smallest and largest observations tell us little about the distribution as a whole, but they give information about the tails of the distribution that is missing if we know only the median and the quartiles. To get a quick summary of both center and variability, combine all five numbers.

The five-number summary

The **five-number summary** of a distribution consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest. In symbols, the five-number summary is

$$\text{Minimum} \quad Q_1 \quad M \quad Q_3 \quad \text{Maximum}$$

These five numbers offer a reasonably complete description of center and variability. The five-number summaries of home run counts are

$$10 \quad 26 \quad 34 \quad 44 \quad 47$$

for Aaron and

$$5 \quad 25 \quad 34 \quad 45 \quad 73$$

for Bonds. The five-number summary of a distribution leads to a new graph, the *boxplot*. Figure 12.2 shows boxplots for the home run comparison.

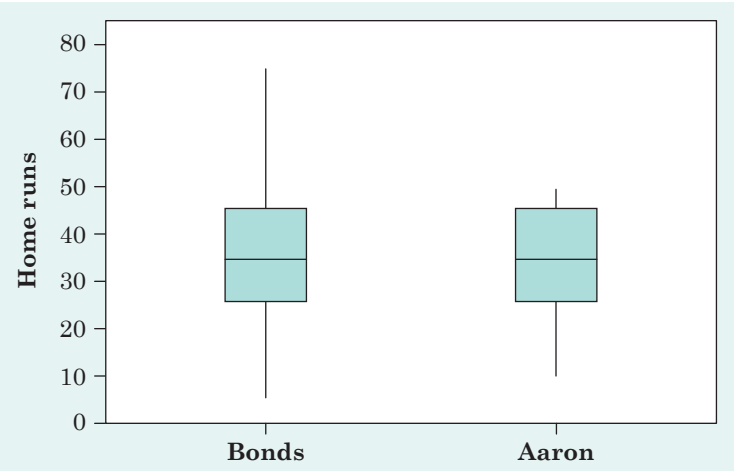


Figure 12.2 Boxplots comparing the yearly home run production of Barry Bonds and Hank Aaron.

Boxplot

A **boxplot** is a graph of the five-number summary.

- A central box spans the quartiles.
- A line in the box marks the median.
- Lines extend from the box out to the smallest and largest observations.

You can draw boxplots either horizontally or vertically. Be sure to include a numerical scale in the graph. When you look at a boxplot, first locate the median, which marks the center of the distribution. Then look at the variability. The quartiles (more precisely, the difference between the two quartiles) show the variability of the middle half of the data, and the extremes (the smallest and largest observations) indicate the variability of the entire data set. We see from Figure 12.2 that Bonds's usual performance, as indicated by the median and the box that marks the middle half of the distribution, is similar to that of Aaron. We also see that the distribution for Aaron is less variable than the distribution for Bonds.

NOW IT'S YOUR TURN

12.2 Babe Ruth. Here are Babe Ruth's home run counts for his 22 years in Major League Baseball, arranged in order from smallest to largest:

0	2	3	4	6	11	22	25	29	34	35
41	41	46	46	46	47	49	54	54	59	60

Draw a boxplot of this distribution. How does it compare with those of Barry Bonds and Hank Aaron in Figure 12.2?

Because boxplots show less detail than histograms or stemplots, they are best used for side-by-side comparison of more than one distribution, as in Figure 12.2. For such small numbers of observations, a back-to-back stemplot is better yet (see Exercise 11.22, page 265). It would make clear, as the boxplot cannot, that Bonds's record 73 home runs in 2001 is an outlier in his career. Let us look at an example where boxplots are more genuinely useful.

EXAMPLE 3 Income inequality

To investigate income inequality, we compare household incomes of Hispanics, blacks, and whites. The Census Bureau website provides information on income distribution by race. Figure 12.3 compares the income distributions for Hispanics, blacks, and whites in 2013. This figure is a variation on the boxplot idea. The largest income among several million people will surely be very large. Figure 12.3 uses the 95% points (the values representing where the top 5% of incomes start) in the distributions instead of the single largest incomes. So, for example, the line above the box for the Hispanic group extends only to $144,040 rather than to the highest income. Many statistical software packages allow you to produce boxplots that suppress extreme values, but the rules for what constitutes an extreme value usually do not use the 95% point in the distribution instead of the single largest value.

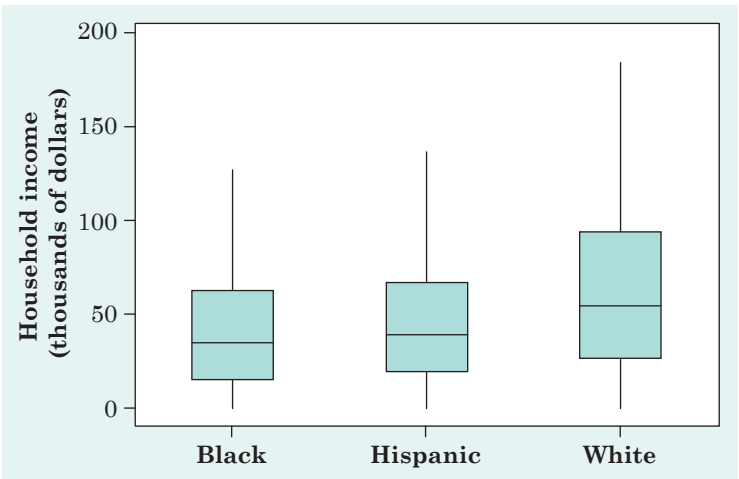


Figure 12.3 Boxplots comparing the distributions of income among Hispanics, blacks, and whites. The ends of each plot are at 0 and at the 95% point in the distribution.

Figure 12.3 gives us a clear and simple visual comparison. We see that the median and middle half are slightly greater for Hispanics than for blacks and that for whites the median and middle half are greater than for both blacks and Hispanics. The income of the bottom 5% stays small because there are some people in each group with no income or even negative income, perhaps due to illness or disability. The 95% point, marking off the top 5% of incomes, is greater for whites than for either blacks or Hispanics, and the 95% point of incomes for Hispanics is greater than for blacks. Overall, incomes for whites tend to be larger than those for Hispanics and blacks, highlighting racial inequities in income.

Figure 12.3 also illustrates how boxplots often indicate the symmetry or skewness of a distribution. In a symmetric distribution, the first and third quartiles are equally distant from the median. In most distributions that are skewed to the right, on the other hand, the third quartile will be farther above the median than the first quartile is below it. The extremes behave the same way. Even with the top 5% not present, we can see the right-skewness of incomes for all three races.

STATISTICAL CONTROVERSIES

Income Inequality

During the prosperous 1980s and 1990s, the incomes of all American households went up, but the gap between rich and poor grew. Figures 12.4 and 12.5 give two views of increasing inequality. Figure 12.4 is a line graph of household income, in dollars adjusted to have the same buying power every year. The lines show the 20th and 80th percentiles of income, which mark off the bottom fifth and the top fifth of households. The 80th percentile (up 47% between 1967 and 2013) is pulling away from the 20th percentile (up about 14%).

Figure 12.5 looks at the *share* of all income that goes to the top fifth and the bottom fifth. The bottom fifth's share has drifted down, to 3.2% of all income in 2013. The share of the top fifth grew to 51% (up 16.4% between 1967 and 2013). Although not displayed in the figures, the share of the top 5% grew even faster, from 17.5% in 1967 to 22.2% of the income of all households in the country in 2013. This is a 26.8% increase between 1967 and 2013. Income inequality in the United States is greater than in other developed nations and has been increasing.

Are these numbers cause for concern? And do they accurately reflect the disparity between the wealthy and the poor? For example, as people get older, their income increases. Perhaps these numbers reflect only the disparity between younger and older wage earners. What do you think?

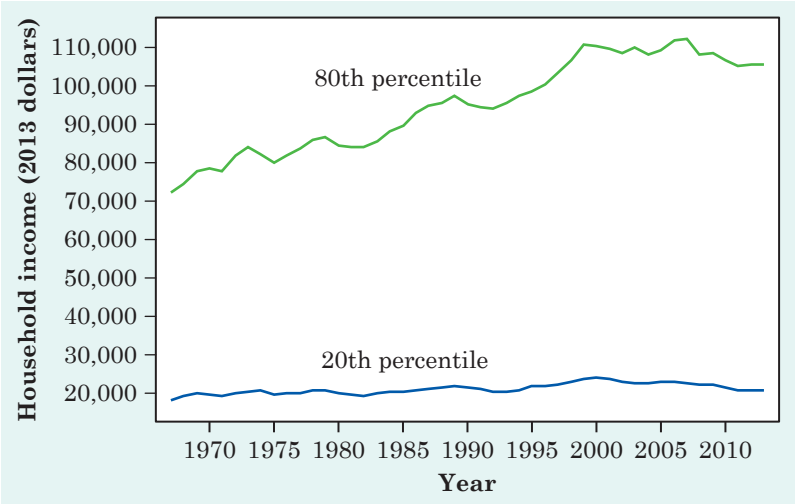
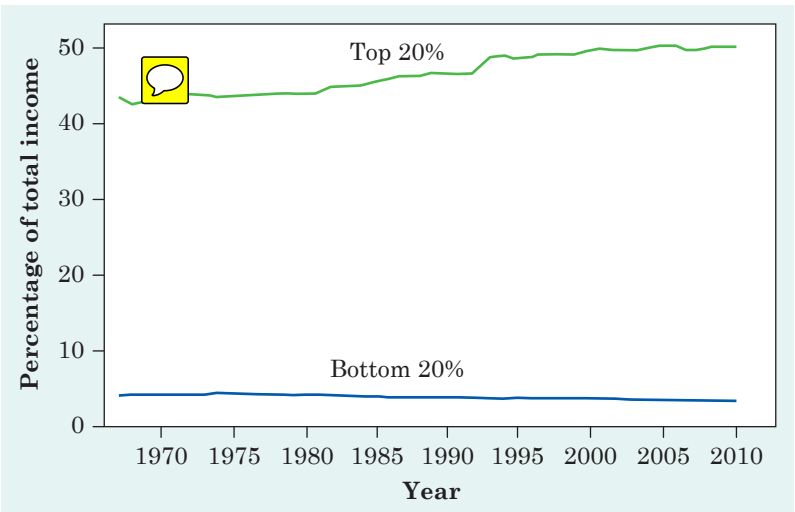


Figure 12.4 The change over time of two points in the distribution of incomes for American households. Eighty percent of households have incomes below the 80th percentile, and 20% have incomes below the 20th percentile. In 2013, the 20th percentile was $20,900, and the 80th percentile was $105,910.



AU: Please check art. Legend refers to data in 2013, but not shown in graph. Update needed?

Figure 12.5 The change over time of the shares of total household income that go to the highest-income 20% and to the lowest-income 20% of households. In 2013, the top 20% of households received 51% of all income.

Mean and 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 variability. The mean is familiar—it is the ordinary average of the observations. The idea of the standard deviation is to give the average distance of observations from the mean. The "average distance" in the standard deviation is found in a rather obscure way. We will give the details, but you may want to just think of the standard deviation as "average distance from the mean" and leave the details to your calculator.

Mean and standard deviation

The **mean** $\bar{x}$ (pronounced "x-bar") of a set of observations is their arithmetic average. To find the mean of n observations, add the values and divide by n:

$$\bar{x} = \frac{\text{sum of the observations}}{n}$$

The **standard deviation** s measures the average distance of the observations from their mean. It is calculated by finding an average of the squared distances and then taking the square root. To find the standard deviation of n observations:

1. Find the distance of each observation from the mean and square each of these distances.

2. Average the squared distances by dividing their sum by $n - 1$. This average squared distance is called the **variance**.

3. The standard deviation s is the square root of this average squared distance.

EXAMPLE 4 Finding the mean and standard deviation

The numbers of home runs Barry Bonds hit in his 22 major league seasons are

16 25 24 19 33 25 34 46 37 33 42
40 37 34 49 73 46 45 45 5 26 28

To find the mean of these observations,

$$\bar{x} = \frac{\text{sum of observations}}{n}$$
$$= \frac{16 + 25 + \cdots + 28}{22}$$
$$= \frac{762}{22} = 34.6$$

Figure 12.6 displays the data as points above the number line, with their mean marked by a vertical line. The arrow shows one of the distances from the mean. The idea behind the standard deviation s is to average the 22 distances. To find the standard deviation by hand, you can use a table layout:

Observation	Squared distance from mean
16	$(16 - 34.6)^2 = (-18.6)^2 = 345.96$
25	$(25 - 34.6)^2 = (-9.6)^2 = 92.16$
⋮	
28	$(28 - 34.6)^2 = (-6.6)^2 = 43.56$
	sum = 4139.12

The average is

$$\frac{4139.12}{21} = 197.1$$

Notice that we "average" by dividing by *one less* than the number of observations. Finally, the standard deviation is the square root of this number:

$$s = \sqrt{197.1} = 14.04$$

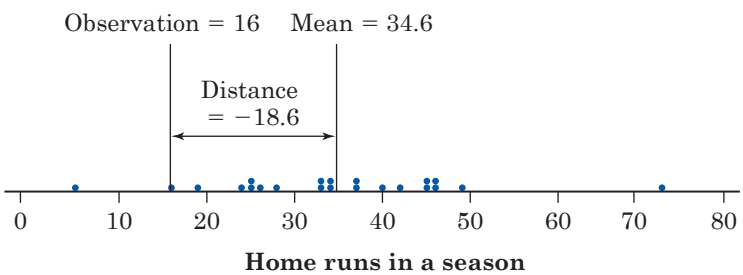


Figure 12.6 Barry Bonds's home run counts, Example 4, with their mean and the distance of one observation from the mean indicated. Think of the standard deviation as an average of these distances.

In practice, you can key the data into your calculator and hit the mean key and the standard deviation key. Or you can enter the data into a spreadsheet or other software to find $\bar{x}$ and s. It is usual, for good but somewhat technical reasons, to average the squared distances by dividing their total by $n - 1$ rather than by n. Many calculators have two standard deviation buttons, giving you a choice between dividing by n and dividing by $n - 1$. Be sure to choose $n - 1$.

NOW IT'S YOUR TURN

12.3 Hank Aaron. Here are Aaron's home run counts for his 23 years in baseball:

13	27	26	44	30	39	40	34	45	44	24	32
44	39	29	44	38	47	34	40	20	12	10	

Find the mean and standard deviation of the number of home runs Aaron hit in each season of his career. How do the mean and median compare?

More important than the details of the calculation are the properties that show how the standard deviation measures variability.

Properties of the standard deviation s

- s measures variability about the mean $\bar{x}$. Use s to describe the variability of a distribution only when you use $\bar{x}$ to describe the center.
- $s = 0$ only when there is *no variability*. This happens only when all observations have the same value. So standard deviation zero means no variability at all. Otherwise, $s > 0$. As the observations become more variable about their mean, s gets larger.

EXAMPLE 5 Investing 101

We have discussed examples about income. Here is an example about what to do with it once you've earned it. One of the first principles of investing is that taking more risk brings higher returns, at least on the average in the long run. People who work in finance define risk as the variability of returns from an investment (greater variability means higher risk) and measure risk by how unpredictable the return on an investment is. A bank account that is insured by the government and has a fixed rate of interest has no risk—its return is known exactly. Stock in

a new company may soar one week and plunge the next. It has high risk because you can't predict what it will be worth when you want to sell.

Investors should think statistically. You can assess an investment by thinking about the distribution of (say) yearly returns. That means asking about both the center and the variability of the pattern of returns. Only naive investors look for a high average return without asking about risk, that is, about how variable the returns are. Financial experts use the mean and standard deviation to describe returns on investments. The standard deviation was long considered too complicated to mention to the public, but now you will find standard deviations appearing regularly in mutual funds reports.

Here by way of illustration are the means and standard deviations of the yearly returns on three investments over the second half of the twentieth century (the 50 years from 1950 to 1999):

Investment	Mean return	Standard deviation
Treasury bills	5.34%	2.96%
Treasury bonds	6.12%	10.73%
Common stocks	14.62%	16.32%

You can see that risk (variability) goes up as the mean return goes up, just as financial theory claims. Treasury bills and bonds are ways of loaning money to the U.S. government. Treasury bills are paid back in one year, so their return changes from year to year depending on interest rates. Bonds are 30-year loans. They are riskier because the value of a bond you own will drop if interest rates go up. Stocks are even riskier. They give higher returns (on the average in the long run) but at the cost of lots of sharp ups and downs along the way. As the stemplot in Figure 12.7 shows, stocks went up by as much as 50% and down by as much as 26% in one year during the 50 years covered by our data.

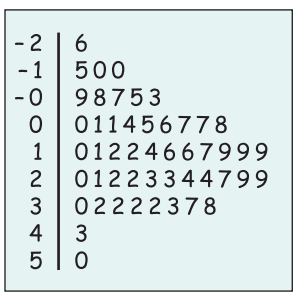


Figure 12.7 Stemplot of the yearly returns on common stocks for the 50 years 1950–1999, Example 5. The returns are rounded to the nearest whole percent. The stems are 10s of percents and the leaves are single percents.

TABLE 12.1 Salaries of the Cleveland Cavaliers, 2014–2015 season

Player	Salary ($)	Player	Salary ($)
LeBron James	20.6 million	Iman Shumpert	2.6 million
Kevin Love	15.7 million	Brendan Haywood	2.2 million
Anderson Varejao	9.7 million	James Jones	1.4 million
Kyrie Irving	7.1 million	Shawn Marion	1.4 million
J.R. Smith	6.0 million	Joe Harris	0.9 million
Tristan Thompson	5.1 million	Matthew Dellavedova	0.8 million
Timofey Mozgov	4.7 million	Kendrick Perkins	0.4 million
Mike Miller	2.7 million		

Source: The salaries are estimates from www.spotrac.com/nba/rankings/2014/base/cleveland-cavaliers/.

Choosing numerical descriptions

The five-number summary is easy to understand and is the best short description for most distributions. The mean and standard deviation are harder to understand but are more common. How can we decide which of these two descriptions of center and variability to use? Let's start by comparing the mean and the median. "Midpoint" and "arithmetic average" are both reasonable ideas for describing the center of a set of data, but they are different ideas with different uses. The most important distinction is that the mean (the average) is strongly influenced by a few extreme observations and the median (the midpoint) is not.

EXAMPLE 6 Mean versus median

Table 12.1 gives the approximate salaries (in millions of dollars) of the 15 members of the Cleveland Cavaliers basketball team for the 2014–2015 season. You can calculate that the mean is $\bar{x} = \$5.5$ million and that the median is $M = \$2.7$ million. No wonder professional basketball players have big houses.

Why is the mean so much higher than the median? Figure 12.8 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 LeBron James and Kevin Love pull up the sum of the salaries and so pull up the mean. If we drop the outliers, the mean for the other 13 players is only $3.5 million. The median doesn't change nearly as much: it drops from $2.7 million to $2.6 million.

stem	leaf
0	4 8 9
1	4 4
2	2 6 7
3	
4	7
5	1
6	0
7	1
8	
9	7
10	
11	
12	
13	
14	
15	7
16	
17	
18	
19	
20	6
21	
22	
23	
24	
25	

Figure 12.8 Stemplot of the salaries of Cleveland Cavaliers players, from Table 12.1.

We can make the mean as large as we like by just increasing LeBron James's salary. The mean will follow one outlier up and up. But to the median, LeBron's salary just counts as one observation at the upper end of the distribution. Moving it from \$20.6 million to \$206 million would not change the median at all.

Poor New York? Is New York a rich state? New York's mean income per person ranks seventh among the states, right up there with its rich neighbors Connecticut and New Jersey, which rank first and second. But while Connecticut and New Jersey rank third and second in median household income, New York stands 17th. What's going on? Just another example of mean versus median. New York has many very highly paid people, who pull up its mean income per person. But it also has a higher proportion of poor households than do Connecticut and New Jersey, and this brings the median down. New York is not a rich state—it's a state with extremes of wealth and poverty.

The mean and median of a symmetric distribution are close to each other. In fact, $\bar{x}$ and M are exactly equal if the distribution is exactly symmetric. In skewed distributions, however, the mean runs away from the median toward the long tail. Many distributions of monetary values—incomes, house prices, wealth—are strongly skewed to the right. The mean may be much larger than the median. For example, we saw in Example 3 that the distribution of incomes for blacks, Hispanics, and whites is skewed to the right. The Census Bureau website gives the mean incomes for 2013 as \$49,629 for blacks, \$54,644 for Hispanics, and \$75,839 for whites. Compare these with the corresponding medians of \$34,598, \$40,963, and \$55,257. Because monetary data often have a few extremely high observations, descriptions of these distributions usually employ the median.

You should think about more than symmetry versus skewness when choosing between the mean and the median. The distribution of selling prices for homes in Middletown is no doubt skewed to the right—but if the Middletown City Council wants to estimate the total market value of all houses in order to set tax rates, the mean and not the median helps them out because the mean will be larger. (The total market value is just the number of houses times the mean market value and has no connection with the median.)

The standard deviation is pulled up by outliers or the long tail of a skewed distribution even more strongly than the mean. The standard deviation of the Lakers' salaries is s = \$5.8 million for all 18 players and only s = \$3.2 million when the outlier is removed. The quartiles are much less sensitive to a few extreme observations. There is another reason to avoid the standard deviation in describing skewed distributions. Because the two sides of a strongly skewed distribution have different amounts of variability, no single number such as s describes the variability well. The five-number summary, with its two quartiles and two extremes, does a better job. In most situations, it is wise to use $\bar{x}$ and s only for distributions that are roughly symmetric.

Choosing a summary

The mean and standard deviation are strongly affected by outliers or by the long tail of a skewed distribution. The median and quartiles are less affected.

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

Why do we bother with the standard deviation at all? One answer appears in the next chapter: the mean and standard deviation are the natural measures of center and variability for an important kind of symmetric distribution, called the Normal distribution.

Do remember that a graph gives the best overall picture of a distribution. Numerical measures of center and variability 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 start with a graph of your data.*

STATISTICS IN SUMMARY

Chapter Specifics

- If we have data on a single quantitative variable, we start with a histogram or stemplot to display the distribution. Then we add numbers to describe the **center and variability** of the distribution.
- There are two common descriptions of center and variability: the **five-number summary** and the **mean and standard deviation**.
- The five-number summary consists of the **median** M, the midpoint of the observations, to measure center, the two **quartiles** Q_1 and Q_3, and the smallest and largest observations. The difference between Q_1 and Q_3 and the difference between the largest and smallest observations describes variability.
- A **boxplot** is a graph of the five-number summary.
- The **mean** $\bar{x}$ is the average of the observations.
- The **standard deviation** s measures variability as a kind of average distance from the mean, so use it only with the mean. The **variance** is the square of the standard deviation.

- The mean and standard deviation can be changed a lot by a few outliers. The mean and median are the same for symmetric distributions, but the mean moves further toward the long tail of a skewed distribution.
- In general, use the five-number summary to describe most distributions and the mean and standard deviation only for roughly symmetric distributions.

In Chapter 11, we discussed histograms and stemplots as graphical displays of the distribution of a single quantitative variable. We were interested in the shape, center, and variability of the distribution. In this chapter, we introduce numbers to describe the center and variability. For symmetric distributions, the mean and standard deviation are used to describe the center and variability. For distributions that are not roughly symmetric, we use the five-number summary to describe the center and variability.

In most of the examples, we used graphical displays and numbers to describe the distribution of data on a single quantitative variable. These data are typically a sample from some population. Thus, the numbers that describe features of the distribution are statistics, as discussed in Chapter 3. In the next chapter, we begin to think about distributions of populations. Thus, the numbers that describe features of these distributions are parameters. In later chapters, we will use statistics to draw conclusions, or make inferences, about parameters. Drawing conclusions about parameters that describe the center of a distribution of a single quantitative variable will be an important type of inference.

CASE STUDY EVALUATED Find the data on income by education at the Census Bureau website listed in the Notes and Data Sources section at the end of the book. Use what you have learned in this chapter to answer the following questions.

1. What are the median incomes for people 25 years old and over who are high school graduates only, have some college but no degree, have a bachelor's degree, have a master's degree, and have a doctoral degree? At the bottom of the table, you will find median earnings in dollars.
2. From the distribution given in the tables, can you find the (approximately) first and third quartiles?
3. Do people with more education earn more than people with less education? Discuss.

LaunchPad Online Resources
macmillan learning

Examples videos

- The StatClips ~~Videos~~ *Summaries of Quantitative Data Example A, Example B, and Example C* describe how to compute the mean, standard deviation, and median of data.

Examples video

- The StatClips ~~Video~~ *Exploratory Pictures for Quantitative Data Example C* describes how to construct boxplots.

video

- The Snapshots ~~Video~~ *Summarizing Quantitative Data* discusses the mean, standard deviation, and median of data, as well as boxplots, in the context of a real example.

CHECK THE BASICS

For Exercise 12.1, see page 271; for Exercise 12.2, see page 273; for Exercise 12.3, see page 279.

12.4 Mean. The mean of the three numbers 1, 2, and 9 is equal to

(a) 2.

(b) 4.

(c) 5.

(d) 6.

12.5 Standard deviation. Which of the following statements is true of the standard deviation?

(a) It measures the variability of a set of data.

(b) It can be strongly affected by outliers.

(c) It is best used as a measure of variability for roughly symmetric distributions.

(d) All of the above.

12.6 Median. Which of the following is true of the median of a set of observations?

(a) It must be larger than the mean of the observations.

(b) It must be smaller than the mean of the observations.

(c) It is the midpoint of the observations.

(d) It describes the variability of the observations.

12.7 The five-number summary. Which of the following is a graph of the five-number summary?

(a) a histogram

(b) a stemplot

(c) a boxplot

(d) a bar graph

12.8 Describing distributions. Which of the following should you use to describe a distribution that is skewed?

(a) the five-number summary

(b) the mean, the first quartile, and the third quartile

(c) the mean and standard deviation

(d) the median and standard deviation

CHAPTER 12 EXERCISES

12.9 Median income. You read that the median income of U.S. households in 2013 was $51,939. Explain in plain language what "the median income" is.

12.10 What's the average? The Census Bureau website gives several choices for "average income" in its historical income data. In 2013, the median income of American households was $51,939. The mean household income was $72,641. The median income of families was $63,815, and the mean family income was $84,687. The Census Bureau says, "Households consist of all people who occupy a housing unit. The term 'family' refers to a group of two or more people related by birth, marriage, or adoption who reside together." Explain carefully why mean incomes are higher than median incomes and why family incomes are higher than household incomes.

12.11 Rich magazine readers. Echo Media reports that the average income for readers of the magazine *WatchTime* (a magazine for people interested in fine watches) is $298,400. Is the median wealth of these readers greater or less than $298,400? Why?

12.12 College tuition. Figure 11.7 (page 255) is a stemplot of the tuition charged by 116 colleges in Illinois. The stems are thousands of dollars and the leaves are hundreds of dollars. For example, the highest tuition is $38,600 and appears as leaf 6 on stem 38.

(a) Find the five-number summary of Illinois college tuitions. You see that the stemplot already arranges the data in order.

(b) Would the mean tuition be clearly smaller than the median, about the same as the median, or clearly larger than the median? Why?

12.13 Where are the young more likely to live? Figure 11.11 (page 259) is a stemplot of the percentage of residents aged 18 to 34 in each of the 50 states. The stems are whole percents and the leaves are tenths of a percent.

(a) The shape of the distribution suggests that the mean will be larger than the median. Why?

(b) Find the mean and median of these data and verify that the mean is larger than the median.

12.14 Gas mileage. Table 11.2 (page 261) gives the highway gas mileages for some model year 2015 mid-sized cars.

(a) Make a stemplot of these data if you did not do so in Exercise 11.13.

(b) Find the five-number summary of gas mileages. Which cars are in the bottom quarter of gas mileages?

(c) The stemplot shows a fact about the overall shape of the distribution that the five-number summary cannot describe. What is it?

12.15 Yankee money. Table 11.4 (page 262) gives the salaries of the New York Yankees baseball team. What shape do you expect the distribution to have? Do you expect the mean salary to be close to the median, clearly higher, or clearly lower? Verify your choices by making a graph and calculating the mean and median.

12.16 The richest 5%. The distribution of individual incomes in the United States is strongly skewed to the right. In 2013, the mean and median incomes of the top 5% of Americans were $196,723 and $322,343. Which of these numbers is the mean and which is the median? Explain your reasoning.

12.17 How many calories does a hot dog have? *Consumer Reports* magazine presented the following data on the number of calories in a hot dog for each of 17 brands of meat hot dogs:

173	191	182	190	172	147
146	139	175	136	179	153
107	195	135	140	138	

Make a stemplot [if you did not already do so in Exercise 11.19 (page 263)], and find the five-number summary. The stemplot shows important facts about the distribution that the numerical summary does not tell us. What are these facts?

12.18 Returns on common stocks. Example 5 informs us that financial theory uses the mean and standard deviation to describe the returns on investments. Figure 11.13 (page 260) is a histogram of the returns of all New York Stock Exchange common stocks in one year. Are the mean and standard deviation suitable as a brief description of this distribution? Why?

12.19 Minority students in engineering. Figure 11.12 (page 259) is a histogram of the number of minority students (black, Hispanic, Native American) who earned doctoral degrees in engineering from each of 152 universities in the years 2000 through 2002. The classes for Figure 11.12 are 1–5, 6–10, and so on.

(a) What is the position of each number in the five-number summary in a list of 152 observations arranged from smallest to largest?

(b) Even without the actual data, you can use your answer to (a) and the histogram to give the five-number summary approximately. Do this. About how many minority engineering PhDs must a university graduate to be in the top quarter?

12.20 The statistics of writing style. Here are data on the percentages of words of 1 to 15 letters used in articles in *Popular Science* magazine. Exercise 11.16 (page 263) asked you to make a histogram of these data.

Length:	1	2	3	4	5
Percent:	3.6	14.8	18.7	16.0	12.5

Length:	6	7	8	9	10
Percent:	8.2	8.1	5.9	4.4	3.6

Length:	11	12	13	14	15
Percent:	2.1	0.9	0.6	0.4	0.2

Find the five-number summary of the distribution of word lengths from this table.

12.21 Immigrants in the eastern states. Here are the number of legal immigrants (in thousands) who settled in each state east of the Mississippi River in 2013:

Alabama	3.8	Connecticut	10.9	Delaware	2.3
Florida	102.9	Georgia	24.4	Illinois	36.0
Indiana	7.7	Kentucky	5.2	Maine	1.2
Maryland	25.4	Massachusetts	29.5	Michigan	17.0
Mississippi	1.7	New Hampshire	2.2	New Jersey	53.1
New York	133.6	North Carolina	16.8	Ohio	13.8
Pennsylvania	24.7	Rhode Island	3.3	South Carolina	4.3
Tennessee	8.4	Vermont	0.8	Virginia	27.9
West Virginia	0.8	Wisconsin	5.9		

Make a graph of the distribution. Describe its overall shape and any outliers. Then choose and calculate a suitable numerical summary.

12.22 Immigrants in the eastern states. New York and Florida are high outliers in the distribution of the previous exercise. Find the mean and the median for these data with and without New York and Florida. Which measure changes more when we omit the outliers?

12.23 State SAT scores. Figure 12.9 is a histogram of the average scores on the SAT Mathematics exam for college-bound senior students in the 50 states and the District of Columbia in 2014. The distinctive overall shape

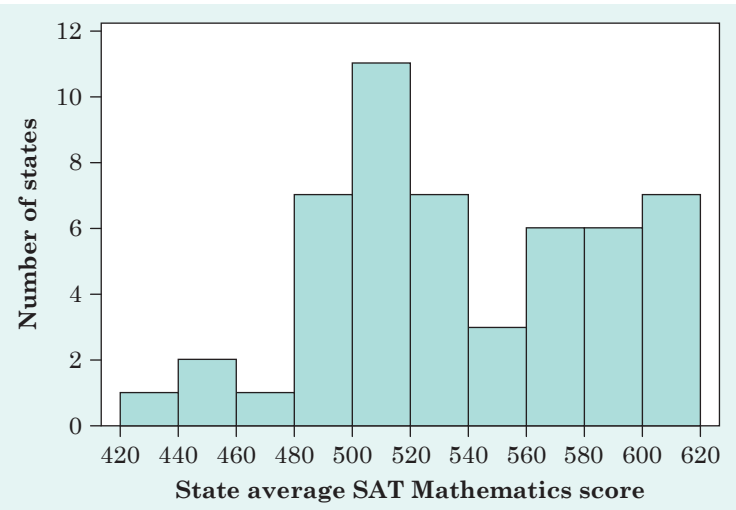


Figure 12.9 Histogram of the average scores on the SAT Mathematics exam for college-bound senior students in the 50 states and the District of Columbia in 2014, Exercise 12.23.

of this distribution implies that a single measure of center such as the mean or the median is of little value in describing the distribution. Explain why this is true.

12.24 Highly paid athletes. A news article reported that of the 411 players on National Basketball Association rosters in February 1998, only 139 "made more than the league average salary" of $2.36 million. Was $2.36 million the mean or median salary for NBA players? How do you know?

12.25 Mean or median? Which measure of center, the mean or the median, should you use in each of the following situations? Why?

(a) Middletown is considering imposing an income tax on citizens. The city government wants to know the average income of citizens so that it can estimate the total tax base.
(b) In a study of the standard of living of typical families in Middletown, a sociologist estimates the average family income in that city.

12.26 Mean or median? You are planning a party and want to know how many cans of soda to buy. A genie offers to tell you either the mean number of cans guests will drink or the median number of cans. Which measure of center should you ask for? Why? To make your answer concrete, suppose there will be 30 guests and the genie will tell you either $\bar{x} = 5$ cans or $M = 3$ cans. Which of these two measures would best help you determine how many cans you should have on hand?

12.27 State SAT scores. We want to compare the distributions of average SAT Math and Writing scores for the states and the District of Columbia. We enter these data into a computer with the names SATM for Math scores and SATW for Writing scores. At the bottom of the page is output from the statistical software package Minitab. (Other software produces similar output. Some software uses rules for finding the quartiles that differ slightly from ours. So software may not give exactly the answer you would get by hand.)

Use this output to make boxplots of SAT Math and Writing scores for the states. Briefly compare the two distributions in words.

12.28 Do SUVs waste gas? Table 11.2 (page 261) gives the highway fuel consumption (in miles per gallon) for 31 model year 2015 mid-sized cars. You found the five-number summary for these data in Exercise 12.14. Here are the highway gas mileages for 26 four-wheel-drive model year 2015 sport utility vehicles:

Minitab output for Exercise 12.27

Can we make this line the same font used for figure captions?

Variable	N	Mean	Median	StDev	Minimum	Maximum	Q1	Q3
SATM	51	537.82	525.00	48.26	502.00	620.00	502.00	585.00
SATW	51	517.86	508.00	45.89	431.00	587.00	477.00	566.00

Model	mpg	Model	mpg
BMW X5 xdrive35i	27	Lexus GX460	20
Chevrolet Tahoe K1500	22	Lexus LX570	17
Chevrolet Traverse	23	Lincoln Navigator	20
Dodge Durango	24	Lincoln MKT	23
Ford Expedition	20	Mercedes-Benz ML250 Bluetec 4matic	29
Ford Explorer	23	Mercedes-Benz G63 AMG	14
GMC Acadia	23	Nissan Armada	18
GMC Yukon	22	Nissan Pathfinder Hybrid	27
Infiniti QX80	19	Porsche Cayenne S	24
Jeep Grand Cherokee	20	Porsche Cayenne Turbo	21
Land Rover LR4	19	Toyota Highlander	24
Land Rover Range Rover	23	Toyota Land Cruiser Wagon	18
Land Rover Range Rover Sport	19	Toyota 4Runner	21

(a) Give a graphical and numerical description of highway fuel consumption for SUVs. What are the main features of the distribution?

(b) Make boxplots to compare the highway fuel consumption of the midsize cars in Table 11.2 and SUVs. What are the most important differences between the two distributions?

12.29 How many calories in a hot dog? Some people worry about how many calories they consume. *Consumer Reports* magazine, in a story on hot dogs, measured the calories in 20 brands of beef hot dogs, 17 brands of meat hot dogs, and 17 brands of poultry hot dogs. Here is computer output describing the beef hot dogs,

```
Mean = 156.8 Standard deviation = 22.64 Min = 111 Max = 190 N = 20 Median = 152.5 Quartiles = 140, 178.5
```

the meat hot dogs,

```
Mean = 158.7 Standard deviation = 25.24 Min = 107 Max = 195 N = 17 Median = 153 Quartiles = 139, 179
```

and the poultry hot dogs,

```
Mean = 122.5 Standard deviation = 25.48 Min = 87 Max = 170 N = 17 Median = 129 Quartiles = 102, 143
```

(Some software uses rules for finding the quartiles that differ slightly from ours. So software may not give exactly the answer you would get by hand.) Use this information to make boxplots of the calorie counts for the three types of hot dogs. Write a brief comparison of the distributions. Will eating poultry hot dogs usually lower your calorie consumption compared with eating beef or meat hot dogs? Explain.

12.30 Finding the standard deviation. 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 six consecutive visits to a clinic:

5.6 5.2 4.6 4.9 5.7 6.4

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

(a) Find the mean from its definition. That is, find the sum of the six observations and divide by 6.

(b) Find the standard deviation from its definition. That is, find the distance of each observation from the mean, square the distances, then calculate the standard deviation. Example 4 shows the method.

(c) Now enter the data into your calculator and use the mean and standard deviation keys to obtain $\bar{x}$ and s. Do the results agree with your hand calculations?

12.31 What *s* measures. Use a calculator to find the mean and standard deviation of these two sets of numbers:

(a) 4 0 1 4 3 6

(b) 5 3 1 3 4 2

Which data set is more variable?

12.32 What *s* measures. Add 2 to each of the numbers in data set (a) in the previous exercise. The data are now 6 2 3 6 5 8.

(a) Use a calculator to find the mean and standard deviation and compare your answers with those for data set part (a) in the previous exercise. How does adding 2 to each number change the mean? How does it change the standard deviation?

(b) Without doing the calculation, what would happen to $\bar{x}$ and s if we added 10 to each value in data set part (a) of the previous exercise? (This exercise demonstrates that the standard deviation measures only variability about the mean and ignores changes in where the data are centered.)

12.33 Cars and SUVs. Use the mean and standard deviation to compare the gas mileages of mid-size cars (Table 11.2, page 261) and SUVs (Exercise 12.28). Do these numbers catch the main points of your more detailed comparison in Exercise 12.28?

12.34 A contest. This is a standard deviation contest. You must choose four numbers from the whole numbers 0 to 9, with repeats allowed.

(a) Choose four numbers that have the smallest possible standard deviation.

(b) Choose four numbers that have the largest possible standard deviation.

(c) Is more than one choice correct in either (a) or (b)? Explain.

12.35 $\bar{x}$ and s are not enough. The mean $\bar{x}$ and standard deviation s measure center and variability but are not a complete description of a distribution. Data sets with different shapes can have the same mean and standard deviation. To demonstrate this fact, use your calculator to find $\bar{x}$ and s for these two small data sets. Then make a stemplot of each and comment on the shape of each distribution.

Data A:	9.14	8.14	8.74	8.77
Data B:	6.58	5.76	7.71	8.84

Data A:	9.26	8.10	6.13	3.10
Data B:	8.47	7.04	5.25	5.56

Data A:	9.13	7.26	4.74
Data B:	7.91	6.89	12.50

12.36 Raising pay. A school system employs teachers at salaries between \$40,000 and \$70,000. 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 \$3000 raise.

(a) How much will the mean salary increase? The median salary?

(b) Will a flat \$3000 raise increase the variability as measured by the distance between the quartiles? Explain.

(c) Will a flat \$3000 raise increase the variability as measured by the standard deviation of the salaries? Explain.

12.37 Raising pay. Suppose that the teachers in the previous exercise each receive a 5% raise. The amount of the raise will vary from \$2000 to \$3500, depending on present salary. Will a 5% across-the-board raise increase the variability of the distribution as measured by the distance between the quartiles? Do you think it will increase the standard deviation? Explain your reasoning.

12.38 Making colleges look good. Colleges announce an "average" SAT score for their entering freshmen. Usually the college would like this "average" to be as high as possible. A *New York Times* article noted, "Private colleges that buy lots of top students with merit scholarships prefer the mean, while open-enrollment public institutions like medians." Use what you know about the behavior of means and medians to explain these preferences.

12.39 What graph to draw? We now understand three kinds of graphs to display distributions of quantitative variables: histograms, stemplots, and boxplots. Give an example (just words, no data) of a situation in which you would prefer that graphing method.

EXPLORING THE WEB

Follow the QR code to access exercises.