

- (c) Make a time plot of Newcomb's values. They are listed in order from left to right, starting with the top row.
- (d) What does the time plot tell you that the display you made in part (a) does not?

Lesson: Sometimes you need to make more than one graphical display to uncover all of the important features of a distribution.

1.2 DESCRIBING DISTRIBUTIONS WITH NUMBERS

Who is baseball's greatest home run hitter? In the summer of 1998, Mark McGwire and Sammy Sosa captured the public's imagination with their pursuit of baseball's single-season home run record (held by Roger Maris). McGwire eventually set a new standard with 70 home runs. Barry Bonds broke Mark McGwire's record when he hit 73 home runs in the 2001 season. How does this accomplishment fit Bonds's career? Here are Bonds's home run counts for the years 1986 (his rookie year) to 2001 (the year he broke McGwire's record):

1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001
16	25	24	19	33	25	34	46	37	33	42	40	37	34	49	73

The stemplot in Figure 1.16 shows us the *shape*, *center*, and *spread* of these data. The distribution is roughly symmetric with a single peak and a possible high outlier. The center is about 34 home runs, and the spread runs from 16 to the record 73. Shape, center, and spread provide a good description of the overall pattern of any distribution for a quantitative variable. Now we will learn specific ways to use numbers to measure the center and spread of a distribution.

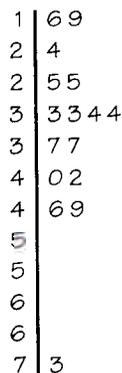


FIGURE 1.16 Number of home runs hit by Barry Bonds in each of his 16 major league seasons.

Measuring center: the mean

A description of a distribution almost always includes a measure of its center or average. The most common measure of center is the ordinary arithmetic average, or *mean*.

THE MEAN \bar{x}

To find the **mean** of a set of observations, add their values and divide by the number of observations. If the n observations are x_1, x_2, \dots, x_n , their mean is

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

or in more compact notation,

$$\bar{x} = \frac{1}{n} \sum x_i$$

The Σ (capital Greek sigma) in the formula for the mean is short for “add them all up.” The subscripts on the observations x_i are just a way of keeping the n observations distinct. They do not necessarily indicate order or any other special facts about the data. The bar over the x indicates the mean of all the x -values. Pronounce the mean \bar{x} as “x-bar.” This notation is very common. When writers who are discussing data use \bar{x} or \bar{y} , they are talking about a mean.

EXAMPLE 1.10 BARRY BONDS VERSUS HANK AARON

The mean number of home runs Barry Bonds hit in his first 16 major league seasons is

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{16 + 25 + \dots + 73}{16} = \frac{567}{16} = 35.4375$$

We might compare Bonds to Hank Aaron, the all-time home run leader. Here are the numbers of home runs hit by Hank Aaron through his last year with Atlanta:

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

Aaron's mean number of home runs hit in a year is

$$\bar{x} = \frac{1}{21}(13 + 27 + \dots + 20) = \frac{733}{21} = 34.9$$

Barry Bonds's exceptional performance in 2001 stands out from his home run production in the previous 15 seasons. Use your calculator to check that his mean home run production in his first 15 seasons is $\bar{x} = 32.93$. One outstanding season increased Bonds's mean home run count by 2.5 home runs per year.

Example 1.10 illustrates an important fact about the mean as a measure of center: it is sensitive to the influence of a few extreme observations. These may be outliers, but a skewed distribution that has no outliers will also pull the mean toward its long tail. Because the mean cannot resist the influence of extreme observations, we say that it is not a *resistant measure* of center.

resistant measure

Measuring center: the median

In Section 1.1, we used the midpoint of a distribution as an informal measure of center. The *median* is the formal version of the midpoint, with a specific rule for calculation.

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.
3. If the number of observations n is even, the median M is the mean of the two center observations in the ordered list.

Medians require little arithmetic, so they are easy to find by hand for small sets of data. Arranging even a moderate number of observations in order is very tedious, however, so that finding the median by hand for larger sets of data is unpleasant. You will need computer software or a graphing calculator to automate finding the median.

EXAMPLE 1.11 FINDING MEDIANS

To find the median number of home runs Barry Bonds hit in his first 16 seasons, first arrange the data in increasing order:

16 19 24 25 25 33 33 **34 34** 37 37 40 42 46 49 73

The count of observations $n = 16$ is even. There is no center observation, but there is a center pair. These are the two bold 34s in the list, which have 7 observations to their left in the list and 7 to their right. The median is midway between these two observations. **Because both of the middle pair are 34, $M = 34$.**

How much does the apparent outlier affect the median? Drop the 73 from the list and find the median for the remaining $n = 15$ years. It is the 8th observation in the edited list, $M = 34$.

How does Bonds's median compare with Hank Aaron's? Here, arranged in increasing order, are Aaron's home run counts:

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

The number of observations is odd, so there is one center observation. This is the median. It is the bold 38, which has 10 observations to its left in the list and 10 observations to its right. Bonds now holds the single-season record, but he has hit fewer home runs in a typical season than Aaron. Barry Bonds also has a long way to go to catch Aaron's career total of 733 home runs.

Comparing the mean and the median

Examples 1.10 and 1.11 illustrate an important difference between the mean and the median. The one high value pulls Bonds's mean home run count up from 32.93 to 35.4375. The median is not affected at all. The median, unlike the mean, is *resistant*. If Bonds's record 73 had been 703, his median would not change at all. The 703 just counts as one observation above the center, no matter how far above the center it lies. The mean uses the actual value of each observation and so will chase a single large observation upward.

The mean and median of a symmetric distribution are close together. If the distribution is exactly symmetric, the mean and median are exactly the same. In a skewed distribution, the mean is farther out in the long tail than is the median. For example, the distribution of house prices is strongly skewed to the right. There are many moderately priced houses and a few very expensive mansions. The few expensive houses pull the mean up but do not affect the median. The mean price of new houses sold in 1997 was \$176,000, but the median price for these same houses was only \$146,000. Reports about house prices, incomes, and other strongly skewed distributions usually give the median ("midpoint") rather than the mean ("arithmetic average"). However, if you are a tax assessor interested in the total value of houses in your area, use the mean. The total value is the mean times the number of houses; it has no connection with the median. The mean and median measure center in different ways, and both are useful.

EXERCISES

1.31 Joey's first 14 quiz grades in a marking period were

86	84	91	75	78	80	74	87	76	96	82	90	98	93
----	----	----	----	----	----	----	----	----	----	----	----	----	----

(a) Use the formula to calculate the mean. Check using "one-variable statistics" on your calculator.

(b) Suppose Joey has an unexcused absence for the fifteenth quiz and he receives a score of zero. Determine his final quiz average. What property of the mean does this situation illustrate? Write a sentence about the effect of the zero on Joey's quiz average that mentions this property.

(c) What kind of plot would best show Joey's distribution of grades? Assume an 8-point grading scale (A: 93 to 100, B: 85 to 92, etc.). Make an appropriate plot, and be prepared to justify your choice.

1.32 SSHA SCORES The Survey of Study Habits and Attitudes (SSHA) is a psychological test that evaluates college students' motivation, study habits, and attitudes toward school. A private college gives the SSHA to a sample of 18 of its incoming first-year women students. Their scores are

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

(a) Make a stemplot of these data. The overall shape of the distribution is irregular, as often happens when only a few observations are available. Are there any potential outliers? About where is the center of the distribution (the score with half the scores above it and half below)? What is the spread of the scores (ignoring any outliers)?

(b) Find the mean score from the formula for the mean. Then enter the data into your calculator. You can find the mean from the home screen as follows:

TI-83	TI-89
• Press $\boxed{2nd}$ \boxed{STAT} (LIST) $\boxed{\blacktriangleright}$ $\boxed{\blacktriangleright}$ (MATH).	• Press $\boxed{CATALOG}$ then $\boxed{5}$ (M).
• Choose 3:mean(, enter list name, press \boxed{ENTER} .	• Choose mean(, type list name, press \boxed{ENTER} .

(c) Find the median of these scores. Which is larger: the median or the mean? Explain why.

1.33 Suppose a major league baseball team's mean yearly salary for a player is \$1.2 million, and that the team has 25 players on its active roster. What is the team's annual payroll for players? If you knew only the median salary, would you be able to answer the question? Why or why not?

1.34 Last year a small accounting firm paid each of its five clerks \$22,000, two junior accountants \$50,000 each, and the firm's owner \$270,000. What is the mean salary paid at this firm? How many of the employees earn less than the mean? What is the median salary? Write a sentence to describe how an unethical recruiter could use statistics to mislead prospective employees.

1.35 U.S. INCOMES The distribution of individual incomes in the United States is strongly skewed to the right. In 1997, the mean and median incomes of the top 1% of Americans were \$330,000 and \$675,000. Which of these numbers is the mean and which is the median? Explain your reasoning.

Measuring spread: the quartiles

The mean and median provide two different measures of the center of a distribution. But a measure of center alone can be misleading. The Census Bureau reports that in 2000 the median income of American households was \$41,345. Half of all households had incomes below \$41,345, and half had higher incomes. But these figures do not tell the whole story. Two nations with the same median household income are very different if one has extremes of wealth and poverty and the other has little variation among households. A drug with the correct mean concentration of active ingredient is dangerous if some batches are much too high and others much too low. We are interested in the *spread* or *variability* of incomes and drug potencies as well as their centers. The simplest useful numerical description of a distribution consists of both a measure of center and a measure of spread.

range

One way to measure spread is to calculate the **range**, which is the difference between the largest and smallest observations. For example, the number of home runs Barry Bonds has hit in a season has a *range* of $73 - 16 = 57$. The range shows the full spread of the data. But it depends on only the smallest observation and the largest observation, which may be outliers. We can improve our description of spread by also looking at the spread of the middle half of the data. The *quartiles* mark out the middle half. Count up the ordered list of observations, starting from the smallest. The *first quartile* lies one-quarter of the way up the list. The *third quartile* lies three-quarters of the way up the list. In other words, the first quartile is larger than 25% of the observations and the third quartile is larger than 75% of the observations. The second quartile is the median, which is larger than 50% of the observations. That's the idea of quartiles. We need a rule to make the idea exact. 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.
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.

Here is an example that shows how the rules for the quartiles work for both odd and even numbers of observations.

EXAMPLE 1.12 FINDING QUARTILES

Barry Bonds's home run counts (arranged in order) are

16 19 24 25 25 33 33 34 34 37 37 40 42 46 49 73

\uparrow q_1 \uparrow M \uparrow q_3

There is an even number of observations, so the median lies midway between the middle pair, the 8th and 9th in the list. The first quartile is the median of the 8 observations to the left of $M = 34$. So $Q_1 = 25$. The third quartile is the median of the 8 observations to the right of M . $Q_3 = 41$. Note that we don't include M when we're computing the quartiles.

The quartiles are *resistant*. For example, Q_3 would have the same value if Bonds's record 73 were 703.

Hank Aaron's data, again arranged in increasing order, are

13 20 24 26 27 29 30 32 34 34 **38** 39 39

\downarrow q_1 \downarrow M

40 40 44 44 44 44 45 47

\uparrow q_3

In Example 1.11, we determined that the median is the bold 38 in the list. The first quartile is the median of the 10 observations to the left of $M = 38$. This is the mean of the 5th and 6th of these 10 observations, so $Q_1 = 28$. $Q_3 = 44$. The overall median is left out of the calculation of the quartiles.

Be careful when, as in these examples, several observations take the same numerical value. Write down all of the observations and apply the rules just as if they all had distinct values. Some software packages use a slightly different rule to find the quartiles, so computer results may be a bit different from your own work. Don't worry about this. The differences will always be too small to be important.

The distance between the first and third quartiles is a simple measure of spread that gives the range covered by the middle half of the data. This distance is called the *interquartile range*.

THE INTERQUARTILE RANGE (IQR)

The *interquartile range (IQR)* is the distance between the first and third quartiles,

$$IQR = Q_3 - Q_1$$

If an observation falls between Q_1 and Q_3 , then you know it's neither unusually high (upper 25%) or unusually low (lower 25%). The IQR is the basis of a rule of thumb for identifying suspected outliers.

OUTLIERS: THE $1.5 \times IQR$ CRITERION

Call an observation an outlier if it falls more than $1.5 \times IQR$ above the third quartile or below the first quartile.

EXAMPLE 1.13 DETERMINING OUTLIERS

We suspect that Barry Bonds's 73 home run season is an outlier. Let's test.

$$IQR = Q_3 - Q_1 = 41 - 25 = 16$$

$$Q_3 + 1.5 \times IQR = 41 + (1.5 \times 16) = 65 \text{ (upper cutoff)}$$

$$Q_1 - 1.5 \times IQR = 25 - (1.5 \times 16) = 1 \text{ (lower cutoff)}$$

Since 73 is above the upper cutoff, Bonds's record-setting year was an outlier.

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 Q_1 , M , and Q_3 . To get a quick summary of both center and spread, combine all five numbers.

THE FIVE-NUMBER SUMMARY

The **five-number summary** of a data set 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

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

These five numbers offer a reasonably complete description of center and spread. The five-number summaries from Example 1.12 are

$$16 \quad 25 \quad 34 \quad 41 \quad 73$$

for Bonds and

$$13 \quad 28 \quad 38 \quad 44 \quad 47$$

for Aaron. The five-number summary of a distribution leads to a new graph the **boxplot**. Figure 1.17 shows boxplots for the home run comparison.

boxplot

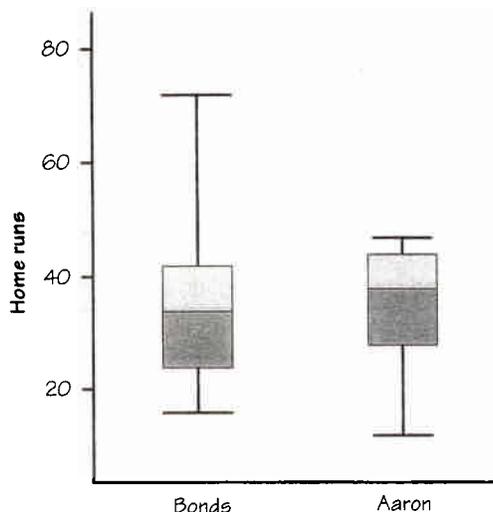


FIGURE 1.17 Side-by-side boxplots comparing the numbers of home runs per year by Barry Bonds and Hank Aaron.

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 1.17. 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 spread. The quartiles show the spread of the middle half of the data, and the extremes (the smallest and largest observations) show the spread of the entire data set. We see from Figure 1.17 that Aaron and Bonds are about equally consistent when we look at the middle 50% of their home run distributions.

A boxplot also gives an indication of 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, however, the third quartile will be farther above the median than the first quartile is below it. The extremes behave the same way, but remember that they are just single observations and may say little about the distribution as a whole. In Figure 1.17, we can see that Aaron's home run distribution is skewed to the left. Barry Bonds's distribution is more difficult to describe.

Outliers usually deserve special attention. Because the regular boxplot conceals outliers, we will adopt the *modified boxplot*, which plots outliers as isolated points. Figures 1.18(a) and (b) show regular and modified boxplots for the home runs hit by Bonds and Aaron. The regular boxplot suggests a very large spread in the upper 25% of Bonds's distribution. The modified boxplot shows that if not for the outlier, the distribution would show much less variability. Because the modified boxplot shows more detail, when we say "boxplot" from now on, we will mean "modified boxplot." Both the TI-83 and the TI-89 give you a choice of regular or modified boxplot. When you construct a (modified) boxplot by hand, extend the "whiskers"

modified boxplot

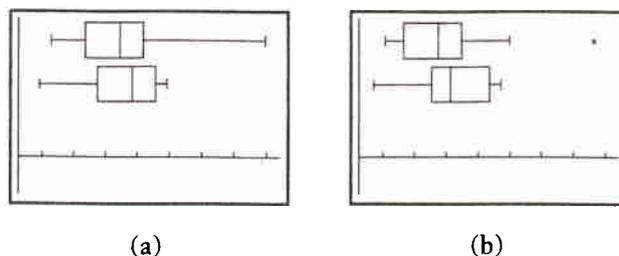


FIGURE 1.18 Regular (a) and modified (b) boxplots comparing the home run production of Barry Bonds and Hank Aaron.

out to the largest and the smallest data points that are not outliers. Then plot outliers as isolated points.

BOXPLOT (MODIFIED)

A **modified boxplot** is a graph of the five-number summary, with outliers plotted individually.

- A central box spans the quartiles.
- A line in the box marks the median.
- Observations more than $1.5 \times IQR$ outside the central box are plotted individually.
- Lines extend from the box out to the smallest and largest observations are not outliers.

TECHNOLOGY TOOLBOX *Calculator boxplots and numerical summaries*

The TI-83 and TI-89 can plot up to three boxplots in the same viewing window. Both calculators can also calculate the mean, median, quartiles, and other one-variable statistics for data stored in lists. In this example, we compare Barry Bonds to Babe Ruth, the “Sultan of Swat.” Here are the numbers of home runs hit by Ruth in each of his seasons as a New York Yankee (1920 to 1934):

54	59	35	41	46	25	47	60	54	46	49	46	41	34	22
----	----	----	----	----	----	----	----	----	----	----	----	----	----	----

1. Enter Bonds’s home run data in L_1 /list1 and Ruth’s in L_2 /list2.
2. Set up two statistics plots: Plot 1 to show a modified boxplot of Bonds’s data and Plot 2 to show a modified boxplot of Ruth’s data.