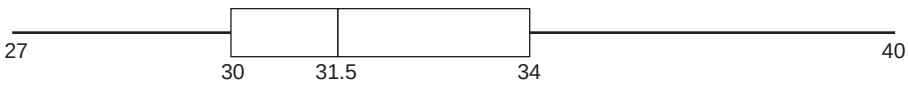

FIGURE 2.7 *A box plot.*

SOLUTION A stem and leaf plot of the data is as follows:

6	0, 5, 5, 8, 9
7	2, 4, 4, 5, 7, 8
8	2, 3, 3, 5, 7, 8, 9
9	0, 0, 1, 4, 4, 5, 7
10	0, 2, 7, 8
11	0, 2, 4, 5
12	2, 4, 5

The first quartile is 74.5, the average of the 9th and 10th smallest data values; the second quartile is 89.5, the average of the 18th and 19th smallest values; the third quartile is 104.5, the average of the 27th and 28th smallest values. ■

A *box plot* is often used to plot some of the summarizing statistics of a data set. A straight line segment stretching from the smallest to the largest data value is drawn on a horizontal axis; imposed on the line is a "box," which starts at the first and continues to the third quartile, with the value of the second quartile indicated by a vertical line. For instance, the 42 data values presented in Table 2.1 go from a low value of 27 to a high value of 40. The value of the first quartile (equal to the value of the 11th smallest on the list) is 30; the value of the second quartile (equal to the average of the 21st and 22nd smallest values) is 31.5; and the value of the third quartile (equal to the value of the 32nd smallest on the list) is 34. The box plot for this data set is shown in Figure 2.7.

The length of the line segment on the box plot, equal to the largest minus the smallest data value, is called the *range* of the data. Also, the length of the box itself, equal to the third quartile minus the first quartile, is called the *interquartile range*.

2.4 CHEBYSHEV'S INEQUALITY

Let $\bar{x}$ and s be the sample mean and sample standard deviation of a data set. Assuming that $s > 0$, Chebyshev's inequality states that for any value of $k \geq 1$, greater than $100(1 - 1/k^2)$ percent of the data lie within the interval from $\bar{x} - ks$ to $\bar{x} + ks$. Thus, by letting $k = 3/2$, we obtain from Chebyshev's inequality that greater than $100(5/9) = 55.56$ percent of the data from any data set lies within a distance $1.5s$ of the sample mean $\bar{x}$; letting $k = 2$ shows that greater than 75 percent of the data lies within $2s$ of the sample mean; and letting $k = 3$ shows that greater than $800/9 \approx 88.9$ percent of the data lies within 3 sample standard deviations of $\bar{x}$.

When the size of the data set is specified, Chebyshev's inequality can be sharpened, as indicated in the following formal statement and proof.

Chebyshev's Inequality

Let $\bar{x}$ and s be the sample mean and sample standard deviation of the data set consisting of the data $x_1, \ldots, x_n$, where $s > 0$. Let

$$S_k = \{i, 1 \le i \le n : |x_i - \bar{x}| < ks\}$$

and let $N(S_k)$ be the number of elements in the set S_k. Then, for any $k \ge 1$,

$$\frac{N(S_k)}{n} \ge 1 - \frac{n-1}{nk^2} > 1 - \frac{1}{k_2}$$

Proof

$$\begin{aligned}
(n-1)s^2 &= \sum_{i=1}^{n} (x_i - \bar{x})^2 \\
&= \sum_{i \in S_k} (x_i - \bar{x})^2 + \sum_{i \notin S_k} (x_i - \bar{x})^2 \\
&\ge \sum_{i \notin S_k} (x_i - \bar{x})^2 \\
&\ge \sum_{i \notin S_k} k^2 s^2 \\
&= k^2 s^2 (n - N(S_k))
\end{aligned}$$

where the first inequality follows because all terms being summed are nonnegative, and the second follows since $(x_1 - \bar{x})^2 \ge k^2 s^2$ when $i \notin S_k$. Dividing both sides of the preceding inequality by $nk^2 s^2$ yields that

$$\frac{n-1}{nk^2} \ge 1 - \frac{N(S_k)}{n}$$

and the result is proven. □

Because Chebyshev's inequality holds universally, it might be expected for given data that the actual percentage of the data values that lie within the interval from $\bar{x} - ks$ to $\bar{x} + ks$ might be quite a bit larger than the bound given by the inequality.

EXAMPLE 2.4a Table 2.7 lists the 10 top-selling passenger cars in the United States in 1999. A simple calculation gives that the sample mean and sample standard deviation of

TABLE 2.7 *Top 10 Selling Cars for 1999*

	1999	
1.	Toyota Camry	448,162
2.	Honda Accord	404,192
3.	Ford Taurus	368,327
4.	Honda Civic	318,308
5.	Chevy Cavalier	272,122
6.	Ford Escort	260,486
7.	Toyota Corolla	249,128
8.	Pontiac Grand Am	234,936
9.	Chevy Malibu	218,540
10.	Saturn S series	207,977

these data are

$$\bar{x} = 298{,}217.8, \qquad s = 124{,}542.9$$

Thus Chebyshev's inequality yields that at least $100(5/9) = 55.55$ percent of the data lies in the interval

$$\left(\bar{x} - \frac{3}{2}s,\ \bar{x} + \frac{3}{2}s\right) = (173{,}674.9,\ 422,760.67)$$

whereas, in actuality, 90 percent of the data falls within those limits. ■

Suppose now that we are interested in the fraction of data values that exceed the sample mean by at least k sample standard deviations, where k is positive. That is, suppose that $\bar{x}$ and s are the sample mean and the sample standard deviation of the data set $x_1, x_2, \ldots, x_n$. Then, with

$$N(k) = \text{number of } i : x_i - \bar{x} \geq ks$$

what can we say about $N(k)/n$? Clearly,

$$\begin{aligned} \frac{N(k)}{n} &\leq \frac{\text{number of } i : |x_i - \bar{x}| \geq ks}{n} \\ &\leq \frac{1}{k^2} \quad \text{by Chebyshev's inequality} \end{aligned}$$

However, we can make a stronger statement, as is shown in the following one-sided version of Chebyshev's inequality.

The One-Sided Chebyshev Inequality

For $k > 0$,

$$\frac{N(k)}{n} \leq \frac{1}{1 + k^2}$$

Proof

Let $y_i = x_i - \bar{x},\ i = 1, \ldots, n$. For any $b > 0$, we have that

$$\begin{aligned}\sum_{i=1}^{n} (y_i + b)^2 &\geq \sum_{i:y_i \geq ks} (y_i + b)^2 \\ &\geq \sum_{i:y_i \geq ks} (ks + b)^2 \\ &= N(k)(ks + b)^2 \end{aligned} \tag{2.4.1}$$

where the first inequality follows because $(y_i + b)^2 \geq 0$, and the second because both ks and b are positive. However,

$$\begin{aligned}\sum_{i=1}^{n} (y_i + b)^2 &= \sum_{i=1}^{n} (y_i^2 + 2by_i + b^2) \\ &= \sum_{i=1}^{n} y_i^2 + 2b \sum_{i=1}^{n} y_i + nb^2 \\ &= (n-1)s^2 + nb^2 \end{aligned}$$

where the final equation used that $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n\bar{x} = 0$. Therefore, we obtain from Equation (2.4.1) that

$$N(k) \leq \frac{(n-1)s^2 + nb^2}{(ks + b)^2}$$

implying that

$$\frac{N(k)}{n} \leq \frac{s^2 + b^2}{(ks + b)^2}$$

Because the preceding is valid for all $b > 0$, we can set $b = s/k$ (which is the value of b that minimizes the right-hand side of the preceding) to obtain that

$$\frac{N(k)}{n} \leq \frac{s^2 + s^2/k^2}{(ks + s/k)^2}$$

Multiplying the numerator and the denominator of the right side of the preceding by k^2/s^2 gives

$$\frac{N(k)}{n} \leq \frac{k^2 + 1}{(k^2 + 1)^2} = \frac{1}{k^2 + 1}$$

and the result is proven. Thus, for instance, where the usual Chebyshev inequality shows that at most 25 percent of data values are at least 2 standard deviations greater than the sample mean, the one-sided Chebyshev inequality lowers the bound to "at most 20 percent." ■