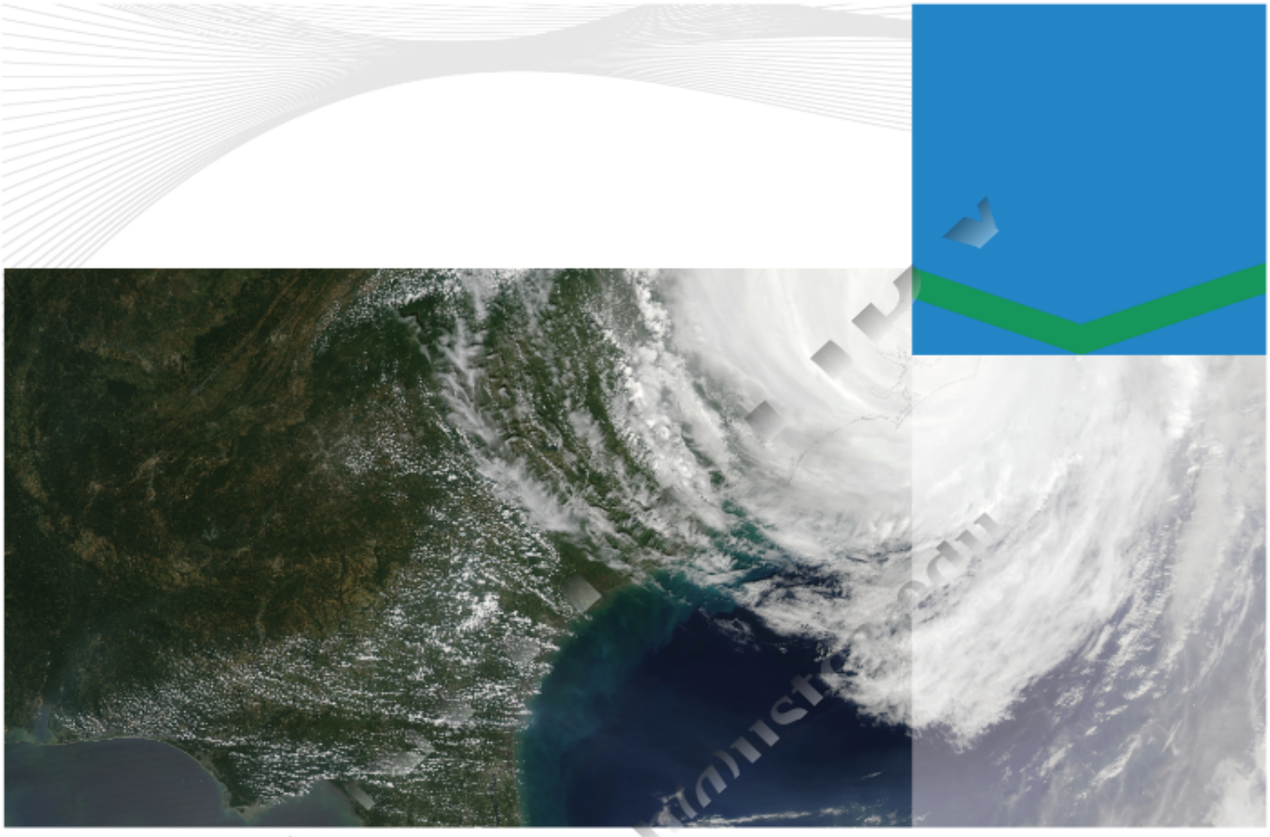
CHAPTER 2 Describing Quantitative Distributions with Numbers



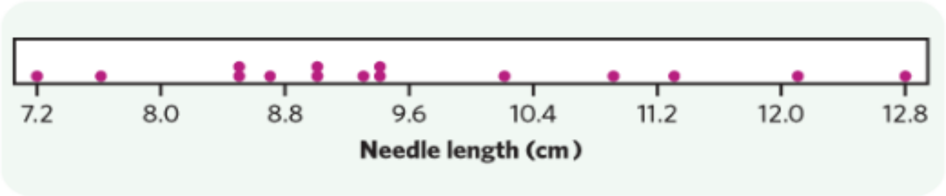
NASA

IN THIS CHAPTER WE COVER...

- Measures of center: median, mean
- Measures of spread: percentiles, standard deviation
- Graphical displays of numerical summaries
- Spotting suspected outliers
- *Discussion: Dealing with outliers*
- Organizing a statistical problem

In Chapter 1 we discussed how the techniques of exploratory data analysis apply equally to data sets obtained from a given *population* (the entire group of individuals about which we want information) or from only those individuals in a smaller *sample*. The distinction between population and sample is an important one when making statistical inferences, which is why different terms and notations are typically used to distinguish them. Numerical summaries are called **parameters** when they describe an entire population and **statistics** when they describe only a sample. In this chapter, we introduce the specific notations and computations for summary statistics relating to samples. We will discuss parameters in more detail starting in Part II of this book.

Examining nature closely reveals surprising variability. The needles in a given pine tree species are not all the same size, for instance. It is the distribution of needle lengths, then, that characterizes the pine species. The following is a dotplot of the lengths (in centimeters, cm) of a sample of 15 needles taken at random from different parts of several Aleppo pine trees located in Southern California:[1]



Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

The distribution is single-peaked and slightly skewed to the right without outliers. Our goal in this chapter is to use numbers to describe the center and the spread of this and other distributions of *quantitative variables*. Categorical variables are simply summarized by the count or proportion (percent) of the data falling into each category, as we saw in Chapter 1.

of this and other distributions of *quantitative variables*. Categorical variables are simply summarized by the count or proportion (percent) of the data falling into each category, as we saw in Chapter 1.

MEASURES OF CENTER: MEDIAN, MEAN

In Chapter 1, we used the approximate midpoint of a distribution (displayed in a dotplot or a histogram) as an informal measure of center. The *median* is the formal version of the midpoint, with a specific rule for calculation. It is the simplest measure of center.

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 smallest observation in the 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. The location of the median is again $(n+1)/2$, counting from the smallest observation in the list.

Note that the formula $(n+1)/2$ does *not* give the median, but rather the *location* of the median in the ordered list. Finding the median requires little arithmetic, so it is easy to do by hand for small sets of data. You can even find the median when the data are summarized in a frequency table, as illustrated in Exercise 2.3. In practice, most people rely on a calculator or statistical software to obtain the median and other numerical summaries.

EXAMPLE 2.1 Finding the median: odd *n*

Craig Tuttle/Corbis Documentary/Getty Images

What is the median length for our 15 Aleppo pine needles? Here are the data, arranged in increasing order:

7.2 7.6 8.5 8.5 8.7 9.0 9.0 **9.3** 9.4 9.4
10.2 10.9 11.3 12.1 12.8

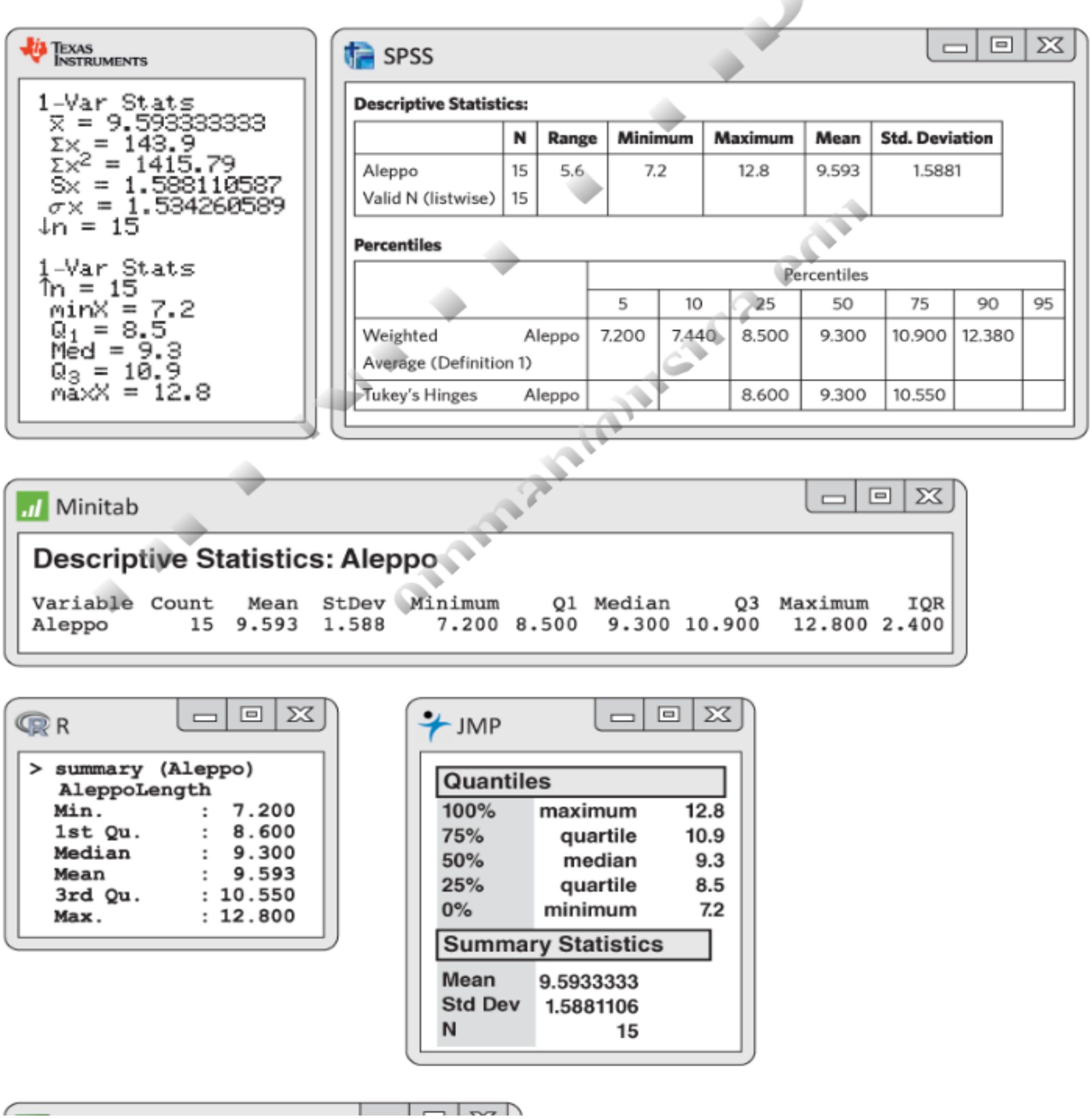
The number of observations $n=15$ is odd. The bold **9.3** is the center observation in the ordered list, with 7 observations to its left and 7 observations to its right. This value is the median, $M=9.3$ cm.

Because $n=15$, our rule for the location of the median gives

$$\text{location of } M = \frac{n+1}{2} = \frac{16}{2} = 8$$

That is, the median is the 8th observation in the ordered list. Use this rule to locate the center in an ordered list or even in a dotplot.

Figure 2.1 displays output describing the 15 Aleppo pine needle lengths using a graphing calculator, a spreadsheet program, and four statistical software packages. Most refer to the median as "Median" (or "Med"), although SPSS labels it "Percentile 50." The median is, in fact, the 50th **percentile** (JMP calls it a "quantile") of a data set because it is the value such that half (50%) of the observations are smaller and the other half are larger. Differences in output from different statistical programs can be frustrating at first, but they are typically superficial. Once you know what to look for, you will be able to correctly interpret output from any technological tool.



```
1-Var Stats
 x̄ = 9.593333333
 Σx = 143.9
 Σx² = 1415.79
 Sx = 1.588110587
 σx = 1.534260589
↓n = 15

1-Var Stats
↑n = 15
 minX = 7.2
 Q1 = 8.5
 Med = 9.3
 Q3 = 10.9
 maxX = 12.8
```

Descriptive Statistics:

	N	Range	Minimum	Maximum	Mean	Std. Deviation
Aleppo	15	5.6	7.2	12.8	9.593	1.5881
Valid N (listwise)	15					

Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average (Definition 1)	Aleppo	7.200	7.440	8.500	9.300	10.900	12.380	
Tukey's Hinges	Aleppo			8.600	9.300	10.550		

Descriptive Statistics: Aleppo

```
Variable  Count   Mean  StDev  Minimum     Q1  Median      Q3  Maximum    IQR
Aleppo       15  9.593  1.588    7.200  8.500   9.300  10.900   12.800  2.400
```

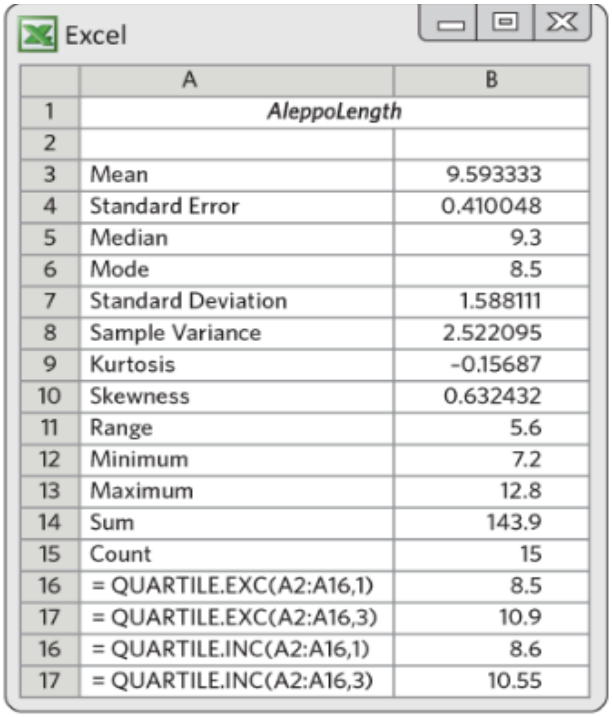
```
> summary (Aleppo)
  AleppoLength
 Min.     :  7.200
 1st Qu.  :  8.600
 Median   :  9.300
 Mean     :  9.593
 3rd Qu.  : 10.550
 Max.     : 12.800
```

Quantiles

100%	maximum	12.8
75%	quartile	10.9
50%	median	9.3
25%	quartile	8.5
0%	minimum	7.2

Summary Statistics

Mean	9.5933333
Std Dev	1.5881106
N	15



	A	B
1	*AleppoLength*	
2		
3	Mean	9.593333
4	Standard Error	0.410048
5	Median	9.3
6	Mode	8.5
7	Standard Deviation	1.588111
8	Sample Variance	2.522095
9	Kurtosis	-0.15687
10	Skewness	0.632432
11	Range	5.6
12	Minimum	7.2
13	Maximum	12.8
14	Sum	143.9
15	Count	15
16	= QUARTILE.EXC(A2:A16,1)	8.5
17	= QUARTILE.EXC(A2:A16,3)	10.9
16	= QUARTILE.INC(A2:A16,1)	8.6
17	= QUARTILE.INC(A2:A16,3)	10.55

Figure 2.1
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 2.1 Output from a graphing calculator, a spreadsheet program, and four software packages describing the Aleppo pine needle length data.

EXAMPLE 2.2 Finding the median: even n

We also have the lengths (in cm) of 18 needles from trees of the noticeably different Torrey pine species. What is the median length for these 18 pine needles? The ordered data are:

21.2	21.6	21.7	23.1	23.7	24.2	24.2	25.5	**26.6**	**26.8**
28.9	29.0	29.7	29.7	30.2	32.5	33.7	33.7		

There is no unique center observation, but there is a center pair—the bold values **26.6** and **26.8**, which have 8 observations before them in the ordered list and 8 observations after them. The median is midway between these two observations:

$$M = \frac{26.6+26.8}{2} = 26.7\ \text{cm}$$

With $n = 18$, the rule for locating the median in the list gives

$$\text{location of } M = \frac{n+1}{2} = \frac{19}{2} = 9.5$$

The location 9.5 means "halfway between the 9th and 10th observations in the ordered list."

■

Another important measure of center is the ordinary arithmetic average, or *mean*. The mean is the most commonly reported measure of center.

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 $x_1, x_2, \ldots, x_n$, their mean is

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

or, in more compact notation,

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

MEASURES OF SPREAD: PERCENTILES, STANDARD DEVIATION

In Chapter 1, we described the distribution of a quantitative variable in a histogram or a dotplot by both its center and its spread. Similarly, *the numerical summaries we provide for a quantitative variable should offer both a measure of center and a measure of spread.*

One way to measure spread is to give the smallest and largest observations in the distribution. For example, the number of coastal flood events in Example 2.4 ranged from 0 to 250. The distance between the minimum and the maximum values in a distribution is called the **range;** in this case, it is 250 (250 − 0).

The minimum and the maximum show the full spread of the data, but they may actually be outliers. Another, more resistant way to describe the spread of a quantitative variable is to look at the spread of the middle half of the data. The first and third *quartiles* mark out the central half of the distribution. The *interquartile range* is the distance between the first and third quartiles.

When the list of observations is sorted in increasing order, the *first quartile* lies one-quarter of the way up the list, and 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. Note that the 50th percentile (the median) is the second quartile, which is larger than 50% of the observations. That is the key idea for quartiles and any other percentiles.

You would typically use technology to obtain summary statistics, but here is how to find the quartiles of a data set by hand.

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.

The **interquartile range *IQR*** is the distance between the first and third quartiles:

$$IQR = Q_3 - Q_1$$

The next two examples show how the rules for the quartiles work for both odd and even numbers of observations.

EXAMPLE 2.5 Finding the quartiles: odd *n*

Our sample of 15 Aleppo pine needles (Example 2.1), arranged in increasing order, is:

7.2 7.6 8.5 8.5 8.7 9.0 9.0 **9.3** 9.4 9.4
10.2 10.9 11.3 12.1 12.8

We have an odd number of observations, so the median is the middle one, the bold **9.3** in the list. The first quartile is the median of the 7 observations to the left of the median. This is the 4th of these 7 observations, so $Q_1 = 8.5\,\text{cm}$. If you want, you can use the formulation for locating the median with $n = 7$:

$$\text{location of } Q_1 = \frac{n+1}{2} = \frac{7+1}{2} = 4$$

The third quartile is the median of the 7 observations to the right of the median, $Q_3 = 10.9\,\text{cm}$.

The quartiles are resistant to outliers. For example, Q_3 would still be 10.9 if the largest needle length were 50 cm rather than 12.8 cm.

CAUTION

Look at the output displays for these 15 Aleppo pine needle lengths in Figure 2.1. The TI-83, JMP, and Minitab all agree with our work. The R output, however, says that $Q_1 = 8.6$ (“1st Qu.”) and $Q_3 = 10.55$ (“3rd Qu.”). How can this be? *There are several rules for finding the quartiles. SPSS provides the results from both methods (“Percentiles 25 and 75”). Some software packages use rules that give results different from ours for some sets of data.* Our rule is the simplest way to perform the computations by hand. Results from applying the various rules are always close to each other, however, so *to describe data you should just use the answer your technology gives you.* Excel’s “Descriptive Statistics” menu item gives more information than we typically need (such as the “mode,” the most common numerical value in the data set), but it does not provide the quartiles. They can be obtained from a separate quartile function, which offers a choice of two computation approaches (as does SPSS).

EXAMPLE 2.6 Finding the quartiles: even *n*

Here are the lengths of the 18 Torrey pine needles (Example 2.3), arranged in increasing order:

21.2	21.6	21.7	23.1	23.7	24.2	24.2	25.5	26.6	*	26.8
28.9	29.0	29.7	29.7	30.2	32.5	33.7	33.7			

We have an even number of observations, so the median lies midway between the middle pair, the 9th and 10th values in the list. The median value is $M = 26.7\,\text{cm}$ and its location is marked by a star. The first quartile is the median of the first 9 observations, because these are the observations to the left of the location of the median. Confirm that $Q_1 = 23.7\,\text{cm}$ and $Q_3 = 29.7\,\text{cm}$.

Be careful when several observations take the same numerical value, as in these examples and in Exercise 2.3. Include all the observations and apply the rules just as if they all had distinct values.

The minimum, first quartile, median, third quartile, and maximum are commonly reported together as the *five-number summary.* These five numbers offer a reasonably complete description of a data set’s center and spread.

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

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

The five-number summaries from Examples 2.5 and 2.6 are

Aleppo pine	7.2	8.5	9.3	10.9	12.8
Torrey pine	21.2	23.7	26.7	29.7	33.7

The five-number summary is not the most common numerical description of a distribution. Rather, that distinction belongs to the combination of the mean (a measure of center) and the *standard deviation* (a measure of spread). The standard deviation and its square value, the *variance*, measure spread by looking at how far the observations are from their mean.

THE SAMPLE STANDARD DEVIATION s

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

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

or, more compactly,

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

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

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

In practice, you will always use software or your calculator to obtain the standard deviation from keyed-in data. However, doing the calculations step by step once helps understand how the variance and standard deviation measure spread.

EXAMPLE 2.7 Calculating the standard deviation

PJF Military Collection/Alamy Stock Photo

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 kilocalories (Cal) for a 24-hour period, where kilocalories are the same calories used to describe the energy content of foods.

1792 1666 1362 1614 1460 1867 1439

The researchers reported $\bar{x} = 1600\,\text{Cal}$ and $s = 189.2\,\text{Cal}$ for this sample of 7 men. Here is how they were computed, step by step. First find the mean:

$$\begin{aligned}\bar{x} &= \frac{1792+1666+1362+1614+1460+1867+1439}{7} \\ &= \frac{11{,}200}{7} = 1600\,\text{Cal}\end{aligned}$$

Figure 2.3 displays the data as points above the number line, with the sample mean $\bar{x}$ marked by an asterisk (*). The arrows mark two of the deviations from the mean. These deviations show how spread out the data are about their mean. They are the starting point for calculating the variance and the standard deviation.

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

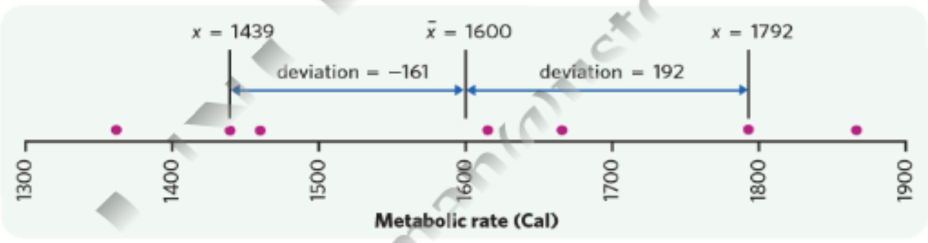


Figure 2.3
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 2.3 Metabolic rates for 7 men, with their mean (*) and the deviations of two observations from the mean indicated.

The sample variance is the sum of the squared deviations divided by 1 less than the number of observations:

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

The sample standard deviation is the square root of the sample variance:

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

were computed, step by step. First find the mean.

$$\begin{aligned} x &= \frac{1792+1666+1362+1614+1460+1867+1439}{7} \\ &= \frac{11{,}200}{7} = 1600\,\text{Cal} \end{aligned}$$

Figure 2.3 displays the data as points above the number line, with the sample mean x marked by an asterisk (*). The arrows mark two of the deviations from the mean. These deviations show how spread out the data are about their mean. They are the starting point for calculating the variance and the standard deviation.

Observations x_i	Deviations $x_i - \overline{x}$			Squared deviations $(x_i - \overline{x})^2$		
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

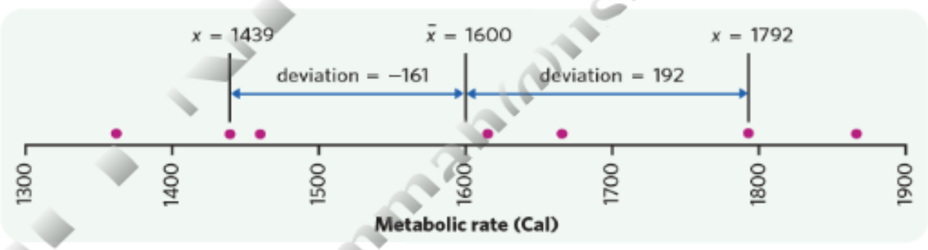


Figure 2.3
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 2.3 Metabolic rates for 7 men, with their mean (*) and the deviations of two observations from the mean indicated.

The sample variance is the sum of the squared deviations divided by 1 less than the number of observations:

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

The sample standard deviation is the square root of the sample variance:

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