CHAPTER 4

# Summary Statistics

The prior chapter used stemplots, frequency tables, and frequency charts to help describe the shape, location, and spread of a distribution. This chapter introduces numerical summaries that are used for similar purposes. Numerical summaries of location and spread are covered. Numerical summaries of shape are *not* covered because they are seldom used in practice.

## 4.1 Central Location: Mean

When used without specification, **mean** refers to the arithmetic average of a data set. This is the most common measure of central location.

To calculate the mean, add up all the values in the data set and then divide by the number of observations. Although this is a simple procedure, it will help to establish **notation** for future use. Consider the following 10 age values:

21 42 5 11 30 50 28 27 24 52

Let:

- $n$ represent the sample size ($n = 10$)
- $x_i$ denotes the value of the $i^{th}$ observation in the data set (e.g., $x_1 = 21$, $x_2 = 42$)
- $\sum_{i=1}^{n} x_i$ tells you to add all the values from $x_1$ to $x_n$. We often drop the subscripts, so $\sum x$ tells you the same thing. For this data, $\sum x = 21 + 42 + 5 + 11 + 30 + 50 + 28 + 27 + 24 + 52 = 290$.

The symbol $\bar{x}$ ("x-bar") represents the sample mean:

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

We can write this in succinct form as:

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

For the data listed, $\bar{x} = \frac{1}{10} \cdot 290 = 29.0$.

**Notes**

1. **Gravitational center.** The mean $\bar{x}$ is the gravitational center of the distribution; it is where the data would balance if placed on a seesaw. **Figure 4.1** depicts this graphically.
2. **Susceptibility to skews.** Because the mean is a balancing point, a small weight far from the center will counterbalance a large weight near the center. This makes the mean highly susceptible to the influence of outliers and skews. **Figure 4.2** depicts this graphically.
3. **Three functions of the mean.** The sample mean tells you three things you might want to know: (1) It can be used to predict an individual value drawn at random from the sample; (2) It can be used to predict a value drawn at random from the population; (3) and it can be used to predict the population mean. These three different uses often get confused.
4. **The mean from a frequency table.** The mean can be derived from a frequency table by calculating the weighted average of values with weights provided by the relative frequency (proportions) of each value using this formula:

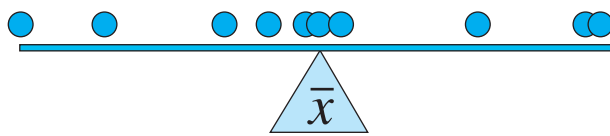


**FIGURE 4.1** The mean is the balancing point of a distribution.

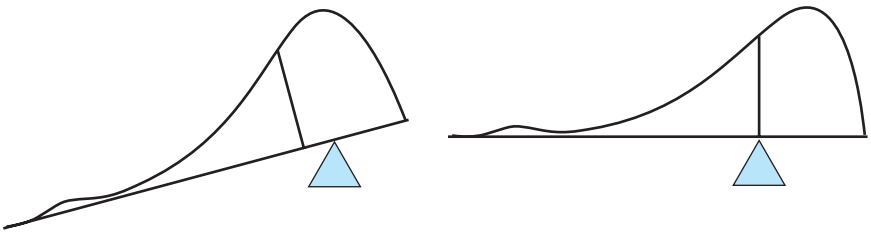

**FIGURE 4.2** A skew tips the distribution causing the mean to shift.

$$\bar{x} = \sum p_i x_i$$

where $p_i$ is the proportion of points with value $x_i$. **Table 4.1** illustrates how this can be done. When data are in uniform class intervals, you use the midpoint of the class interval in this formula as the value of $x_i$ to get an *approximate* mean.

## Exercises

**4.1** ***Gravitational center.*** This exercise will demonstrate how the mean is the balancing point of a set of numbers.

(a) **Distribution *A*.** The values 1 and 5 are marked as "Xs" on the following number line. Calculate the mean and mark its location on the number line.

```
X                   X
1----2----3----4----5
```

**Table 4.1** The mean calculated from frequencies. Data are frequencies of fatal horse kicks in the Prussian Army; a classical example from Bortkiewicz (1898).

| Interval $i$ | No. of fatalities $x_i$ | Frequency $f_i$ | Proportion $p_i$ |
|---|---|---|---|
| 1 | 0 | 109 | 0.545 |
| 2 | 1 | 65 | 0.325 |
| 3 | 2 | 22 | 0.110 |
| 4 | 3 | 3 | 0.015 |
| 5 | 4 | 1 | 0.005 |
| 6 | 5 | 0 | 0.000 |

$\bar{x} = \sum p_i x_i = (0.545 \cdot 0)+(0.325 \cdot 1)+(0.110 \cdot 2)+(0.015 \cdot 3)+(0.005 \cdot 4) = 0.61$

Data source: Bortkiewicz, L. V. (1898). *Das Gesetz Der Kleinen Zahlen*. Liepzig: Tuebner.

(b) **Distribution *B*.** The values 1, 5, and 5 are shown on the number line. Calculate the mean and show its location on the number line. Notice how the extra 5 pulls the mean to the right.

```
                    X
X                   X
1----2----3----4----5
```

(c) **Distribution *C*.** Calculate and show the location of the mean for the data points 2.75, 3.00, and 3.25 on the following number line:

```
         XXX
1----2----3----4----5
```

(d) **Distribution *D*.** Calculate and show the mean of these three points:

```
          X
          X         X
1----2----3----4----5
```

Exercise 3.1 should convince you that the mean tells you nothing about the spread or shape of a distribution. All four distributions are different, yet distributions *A* and *C* have the same means ($\bar{x} = 3$), as do distributions *B* and *D* ($\bar{x} = 3.67$). Reliance solely on the mean would have missed the full picture. Consider:

- Describing the central location of a pendulum tells you little of its motion.
- If you have your head in the freezer and your feet in the oven, your average body temperature can still be normal.
- You can drown in a deep area of a lake that has an average depth of just a few inches.

**Sole reliance on a mean often misses the true picture. When you report a distribution's mean, you should always do so in the context of its shape and spread.**

**4.2** ***Visualizing the mean.*** Consider these eight data points:

1.47 2.06 2.36 3.43 3.74 3.78 3.94 4.42

A stemplot of the data looks like this:

```
1|4
2|03
3|4779
4|4
×1
```

Visually estimate ("eyeball") the balancing point of the distribution; then calculate the distribution's mean. How well did you do with your eyeball estimate?

**4.3** ***More visualization.*** **Figure 4.3** contains three stemplots. Stemplot *A* represents ages (years) of participants in a childhood health survey ($n = 654$). Stemplot *B* represents body weights of students in a class ($n = 53$). Stemplot *C* represents coliform levels in water samples ($n = 25$). Look at each of these plots and visually estimate each mean. It may help to tip your head to the right to help imagine where the stemplots would balance. The calculated means are listed in the answer key in the back of the book. *After* you have completed your visual estimates, compare them to the actual means.

## 4.2 Central Location: Median

The **median** is a different kind of average. It is the midpoint of a distribution—the point that is greater than or equal to half of the values in the data set. To find the median**,** arrange the data in rank order (i.e., create an ordered array of the data) and then count in from either end of the array to a **depth** of $\frac{n+1}{2}$. When $n$ is odd, you will land on the median. When $n$ is even, you will land between two values. Average these values to determine the actual median. For example, the median of this small data set with 10 observations has a depth $\frac{10+1}{2} = 5.5$, placing it between 27 and 28.

```
5   11   21   24   27   28   30   42   50   52
                   ↑
                 Median
```

Average these values to find the median $= \frac{27+28}{2} = 27.5$.

**Stemplot *A***

```
 3 | 0
 4 | 0000
 5 | 00000000000000
 6 | 000000000000000000
 7 | 00000000000000000000000000
 8 | 0000000000000000000000000000000000000000
 9 | 000000000000000000000000000000000000000000000000
10 | 00000000000000000000000000000000000000
11 | 000000000000000000000000000000000000000000000
12 | 0000000000000000000000000000
13 | 00000000000000000000
14 | 000000000000
15 | 000000000
16 | 000000
17 | 0000
18 | 000
19 | 0
x1
```

**Stemplot *B***

```
10 | 0166
11 | 009
12 | 0034578
13 | 00359
14 | 08
15 | 00257
16 | 555
17 | 000255
18 | 000055567
19 | 245
20 | 3
21 | 025
22 | 0
23 |
24 |
25 |
26 | 0
x10
```

**Stemplot *C***

```
1 | 4
1 | 789
2 | 2234
2 | 66789
3 | 00012344
3 | 5678
x1
```

**FIGURE 4.3** Stemplots for Exercise 4.3; visualize the locations of the means.

The median is more **resistant** to outliers than the mean. Consider theses five values:

```
4  7  8  9  12
      ↑
```

The median is 8. Now *suppose* there was a data entry error, so that the value of 12 was mistakenly recorded as 120:

```
4  7  8  9  120
      ↑
```

The median is still 8, but the mean goes from 8 to 29.6. The median stayed put, while the mean more than tripled.

**The median is relatively resistant to outliers and skews.**

**Illustrative Example:** ***Comparison of mean, median, and mode.*** The duration of 15 complex surgical procedures are shown in this stemplot:

```
1|8
2|1345568
3|01158
4|
5|
6|5
7|0
×1
```

The median of this distribution is 2.8 (underlined). The mean is about 3.3.[a] The mean has been "pulled up" by the two high outliers (6.5 and 7.0). ■

[a] $\bar{x} = \frac{1}{15}(1.8 + 2.1 + \cdots + 7.0) = \frac{1}{15}(49.0) > 3.3$

## 4.3 Central Location: Mode

The **mode** is the most frequently occurring value in the data set. For example, the following data set has a mode of 7:

4 7 7 7 8 8 9

With small data sets, it is often preferable to report a class interval as the mode (as opposed to using a single value). This will be where the stemplot shows a peak. As examples,

- Stemplot *A* in **Figure 4.3** has a mode at 9.
- Stemplot *B* in **Figure 4.3** seems to be **bimodal**, with peaks in the 120s and 180s, probably reflecting different weight distributions for female and male students.
- Stemplot *C* in **Figure 4.3** has a mode in the interval 3.0 to 3.4.

## 4.4 Comparison of the Mean, Median, and Mode

The mean, median, and mode will coincide in **unimodal** distributions that are symmetrical. With asymmetry, with the mean will be pulled toward the skew more so than the median. **Figure 4.4** depicts this fact. Because of this, you can predict the shape of a distribution by comparing its mean and median.

When mean = median → distribution is symmetrical
When mean > median → there is a positive skew
When mean < median → there is a negative skew

### *Exercise*

**4.4** ***Seizures following meningitis.*** Exercise 3.9 considered induction times (months) for seizures in 13 cases. Data were {0.10, 0.25, 0.50, 4, 12, 12, 24, 24, 31, 36, 42, 55, 96}.

(a) Calculate the mean and median of this data set.

(b) Compare the mean and median. What does this tell you about the distribution's shape?

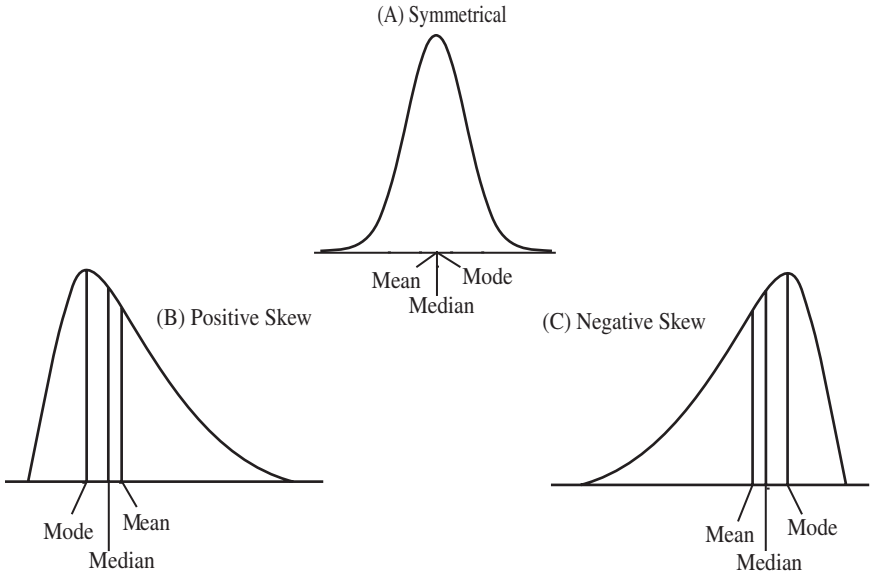


**FIGURE 4.4** Effect of a skew on the mean, median, and mode.

(c) Which measure of central location would you use to describe the distribution's center? Explain your preference.

## 4.5 Spread: Quartiles

As noted earlier, sole reliance on a measure of central location to describe a distribution can often miss the point. Accompanying the measure of central location with a measure of spread (**variability**) will help fill out the picture. Here is an example to illustrate this fact:

**Illustrative Example:** ***Importance of spread (Air samples).*** Air samples were collected on eight successive days at two different sites. Particulate matter was measured in these samples ($\mu g/m^3$) as an index of environmental quality. **Table 4.2** lists the data for the two sites.

*continues*

*continues*

**Table 4.2** Data for "Air samples" illustrative example. Particulates in air samples on successive days at two sites ($\mu g/m^3$).

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Site 1: | 68 | 22 | 36 | 32 | 42 | 24 | 28 | 38 |
| Site 2: | 36 | 38 | 39 | 40 | 36 | 34 | 33 | 32 |

Data are fictitious but realistic. The data file is stored online in the file AIRSAMPLES.*.

To the nearest unit, both sites have means of 36 $\mu g/m^3$. However, side-by-side stemplots reveal different pictures:

```
Site 1| |Site 2
--------------
    42|2|
     8|2|
     2|3|234
    86|3|6689
     2|4|0
      |4|
      |5|
      |5|
      |6|
     8|6|
    ×10
```

Notice how there is much greater variability of readings at site 1. One particular reading at site 1 was very high (68 $\mu g/m^3$), indicating a very "dirty" day. It is this type of occurrence that can be hazardous, especially for individuals with compromised cardiorespiratory function. ■

## *Range*

There are several ways to measure the spread of a distribution. The simplest measure of spread is the **range**, which is merely:

$$\text{Range} = \text{Maximum} - \text{Minimum}$$

The range of the particulate matter in air samples for site 1 in the prior illustrative example is equal to $68 - 22 = 46$. The range at site 2 is equal to $40 - 32 = 8$. Variability is much greater at site 1.

Although suited for some purposes, the range is *not* a very good general measure of spread. It accounts only for two values in the data set (the maximum and minimum), making it a relatively *inefficient* statistic. In addition, it is unlikely to capture both the maximum and minimum in the population, so it has a tendency to underestimate the population's range, making it a *biased* statistic. Therefore, when used, the range should be supplemented with at least one of the other measures of spread described in this section.

## Quartiles, 5-Point Summary, Interquartile Range

**Quartiles** provide an inefficient and intuitive way to describe variability. The idea is to divide the data set into four equal segments.[b] **Quartile 1 (Q1)** cuts off the bottom quarter of data points. **Quartile 3 (Q3)** cuts off the top quarter. We then describe the distance between the quartiles as a measure of spread.

Because data do not always divide neatly into quarters, we need rules for interpolating quartiles. Several interpolation rule systems exist. The rule system we will use derives from Tukey's hinges.[c] **Hinges** are where the ordered array "folds" upon itself. To determine quartiles by the hinge method:

1. Put the data in rank order; then locate the median.
2. Divide the data set into a "low group" and a "high group" where they split at the median. When $n$ is odd, put the median in both groups.
3. Find the middle value ("median") of the low group. This is **Q1**.
4. Find the middle value ("median") of the high group. This is **Q3**.

**Illustrative Example: *Quartiles.*** Consider this small ordered array ($n = 10$):

```
5    11    21    24    27    28    30    42    50    52
      ↑           ↑           ↑
      Q1          M           Q3
```

The low group is {5, 11, 21, 24, 27}, so Q1 = 21. The high group is {28, 30, 42, 50, 52}, so Q3 = 42. ■

[b]The idea of dividing a data set into groups of equal size was initially described by Francis Galton in the 19th century. He was also first to use the term "quartile." See: Galton, F. (1879). On a proposed statistical scale. *Nature*, *9*, 342–343.

[c]Beware that there are different rule systems to determine quartiles. If you are using a software program or calculator to determine quartiles, results may differ from Tukey's hinges.

We can summarize the data set with the points that define its quartiles. This is known as the **five-point summary** for the data set, consisting of:

- **Q0** (the minimum)
- **Q1** (the lower hinge)
- **Q2** (the median)
- **Q3** (the upper hinge)
- **Q4** (the maximum)

The five-point summary for the data set in the previous illustrative example is 5, 21, 27.5, 42, 52.

The **interquartile range (IQR)** is a summary measure of spread consisting of:

$$\text{IQR} = \text{Q3} - \text{Q1}$$

This range captures the middle 50% of data points in a set. The IQR for the current illustrative data $= 42 - 21 = 21$.

**Illustrative Example: *IQR* ($n = 7$).** What is the five-point summary and IQR of this small data set?

```
1.47   2.06   2.36   3.43   3.74   3.78   3.94
```

- The median is 3.43.
- The low group consisting of {1.47, 2.06, 2.36, 3.43} has a middle value between 2.06 and 2.36. We average these to calculate

  $$\text{Q1} = \frac{2.06 + 2.36}{2} = 2.21.$$

- The high group consisting of {3.43, 3.74, 3.78, 3.94} has a middle value between 3.74 and 3.78, so $\text{Q3} = \frac{3.74 + 3.78}{2} = 3.76.$
- The five-point summary is 1.47, 2.21, 3.43, 3.76, 3.94.
- The IQR $= 3.76 - 2.21 = 1.55$.

**Illustrative Example: *IQRs (Air samples data).*** Recall the "air samples" illustrative data presented earlier in this chapter. **Table 4.2** lists data for particulate matter at two sampling sites.

**For site 1**, the ordered array or data is:

```
22   24   28   32   36   38   42   68
       ↑         ↑         ↑
      Q1         M         Q3
```

Thus, Q1 = the average of 24 and 28 (which is 26) and Q3 = the average of 38 and 42 (which is 40). The IQR = 40 − 26 = 14.

**For site 2**, the ordered array is:

```
32   33   34   36   36   38   39   40
       ↑         ↑         ↑
      Q1         M         Q3
```

Q1 = 33.5 and Q3 = 38.5. Therefore, IQR = 38.5 − 33.5 = 5.

Data are less variable at site 2 (IQR: 14 vs. 5). ■

## 4.6 Boxplots

**Box-and-whiskers plots (boxplots)** display five-point summaries and "potential outliers" in graphical form. The box of the boxplot spans the IQR of a data set. The median is indicated as a line within the box. Extreme values (potential outliers) are plotted as separate points beyond data set "whiskers." To construct a boxplot:

1. Determine the five-point summary for the data. Draw a **box** extending from hinge to hinge (i.e., from Q1 to Q3).
2. Calculate the IQR and use this to determine **fences** as follows:

$$\text{Fence}_{\text{Lower}} = \text{Q1} - (1.5)(\text{IQR})$$
$$\text{Fence}_{\text{Upper}} = \text{Q3} + (1.5)(\text{IQR})$$

   Do *not* plot the fences.
3. Determine if the data set contains **outside values**. Values below the lower fence are **lower outside values**. Values above the upper fence are **upper outside values**. Plot these as separate points on the graph.
4. The smallest value inside the lower fence is the **lower inside value**. The largest value inside the upper fence is the **upper inside value**. Draw **whiskers** from respective quartiles (hinges) to inside values.

Boxplots provide insight into the distribution's location, spread, and shape:

- **Location:** The location of the entire distribution is visible. The box cradles the middle 50% of values. The line in the box locates the median.
- **Spread:** The IQR (**hinge spread**) is visible as the height of the box. The **whisker spread** (from whisker to whisker) and range also provide visual clues of spread.
- **Shape:** Shape is difficult to judge when the sample is small. With large data sets, you can judge symmetry and get a sense of whether the distribution has long or short tails.

**Illustrative Example:** ***Boxplots (Air samples data).*** We propose to draw boxplots for the air sample data presented in prior illustrations (see **Table 4.2**).

Data for **site 1** is: 22 24 28 32 36 38 42 68

1. The five-point summary (determined in a prior illustration) is: 22, 26, 34, 40, 68. Therefore, the box extends from 26 to 40. A line for the median is drawn in the box at 34.
2. IQR = 40 − 26 = 14. Therefore, $\text{Fence}_{\text{Lower}} = 26 - (1.5)(14) = 5$ and $\text{Fence}_{\text{Upper}} = 40 + (1.5)(14) = 61$.
3. There are no values below the lower fence. There is one value (68) above the upper fence. The *upper outside value* of 68 is plotted as a separate point above the box.
4. The smallest value still inside the fence (lower inside value) is 22. A whisker is drawn from Q1 to this inside value. The upper inside value is 42. A whisker is drawn from Q3 to this inside value.

**Figure 4.5** shows the boxplot for site 1 on the left.

Data for site 2 is: 32 33 34 36 36 38 39 40

1. The five-point summary is 32, 33.5, 36, 38.5, 40. The box extends from 33.5 to 38.5. A line for the median is drawn inside the box at 36.
2. IQR = 38.5 − 33.5 = 5. $\text{Fence}_{\text{Lower}} = 33.5 - (1.5)(5) = 26$. $\text{Fence}_{\text{Upper}} = 38.5 + (1.5)(5) = 46$.
3. There are no outside values in this group.

*continues*

*continues*

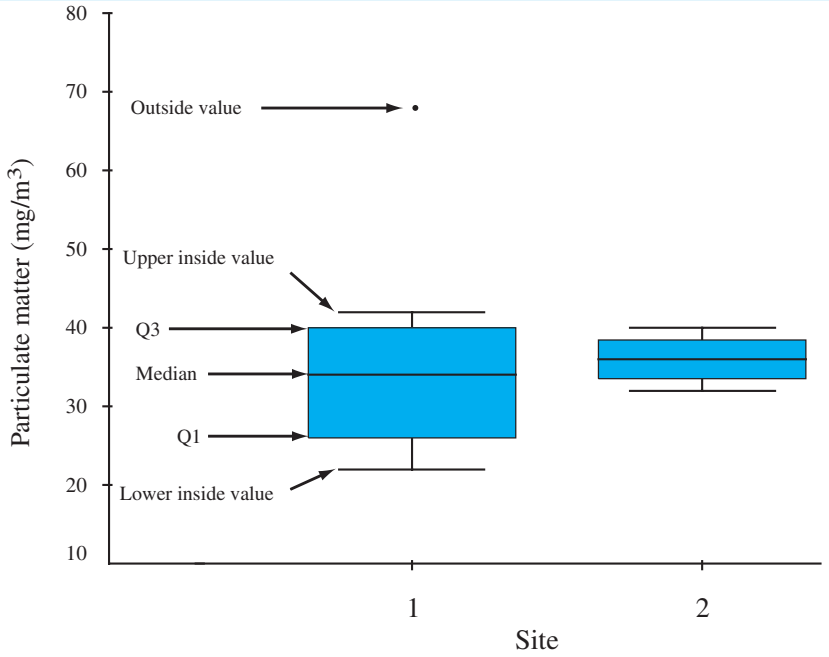


**FIGURE 4.5** Side-by-side boxplots, Air samples illustrative data.

4. The lower inside value is 32. A whisker is drawn from Q1 to this lower inside value. The upper inside value is 42. A whisker is drawn from Q3 to this upper inside value.

**Figure 4.5** shows this boxplot on the right. The side-by-side boxplots in this figure show site 1 and site 2 have similar central locations but different spreads. ■

## *Exercises*

**4.5** ***Outside?*** The following stemplot has 18 observations. Prove that the value 152 in this data set is an outside value.

```
08|8
09|59
10|0147
11|444679
12|0145
13|
14|
15|2
×10
```

**4.6** ***Seizures following bacterial meningitis.*** In Exercise 4.4, you calculated the mean and median of induction times in 13 seizures cases. Data were {0.10, 0.25, 0.50, 4, 12, 12, 24, 24, 31, 36, 42, 55, 96} months. Now construct a boxplot for these data. Are there any outside values in this data set? Does the boxplot show evidence of asymmetry?

## 4.7 Spread: Variance and Standard Deviation

The **variance** and **standard deviation** are the most common measures of spread. These statistics are based on the average squared distances of values around the data set's mean. Here is how they are calculated:

- Determine the **deviation** of each data point. A deviation is the data point minus its mean: $x_i - \bar{x}$. **Figure 4.6** shows deviations for two of the values in the air samples data (**Table 4.2**) for site 2.
- **Square each deviation**. This is equal to $(x_i - \bar{x})^2$.

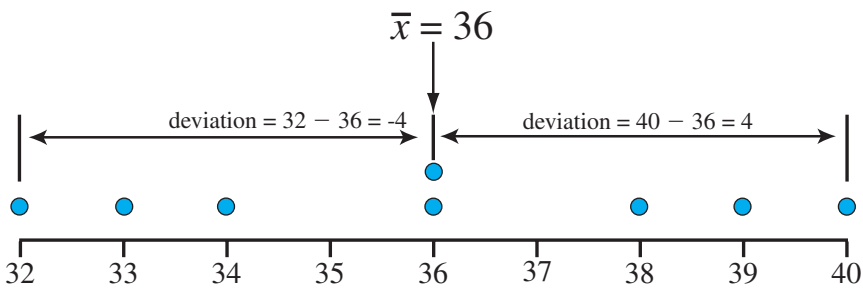


**FIGURE 4.6** Deviations of two observations, site 2, air samples illustrative data, Table 4.2.

- **Sum the squared deviations.** This statistic is the **sum of squares (*SS*)**:

$$SS = (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2$$

- Divide the sum of squares by $n$ minus 1. This is the **variance**, denoted $s^2$:

$$s^2 = \frac{SS}{n - 1}$$

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

- Take the square root of the variance. This is the **standard deviation** ($s$):

$$s = \sqrt{s^2}$$

Equivalently,

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

In practice, we often use software to compute the variance and standard deviation. However, it is instructive to complete these calculations by hand, so let's practice a few.

**Illustrative Example: *Standard deviation (Air samples from site 2).*** We propose to calculate the variance and standard deviation for the particulate matter in the air samples from site 2 for the air samples data (**Table 4.2** and **Figure 4.6**). At site 2, $n = 8$ and

$$\bar{x} = \frac{36 + 38 + 39 + 40 + 36 + 34 + 33 + 32}{8} = 36 \ (\mu g/m^3).$$

Deviations and squared deviations for each point are :

| Data $x_i$ | Deviations $x_i - \bar{x}$ | Squared deviations $(x_i - \bar{x})^2$ |
|---|---|---|
| 36 | $36 - 36 = 0$ | $0^2 = 0$ |
| 38 | $38 - 36 = 2$ | $2^2 = 4$ |
| 39 | $39 - 36 = 3$ | $3^2 = 9$ |
| 40 | $40 - 36 = 4$ | $4^2 = 16$ |
| 36 | $36 - 36 = 0$ | $0^2 = 0$ |
| 34 | $34 - 36 = -2$ | $-2^2 = 4$ |
| 33 | $33 - 36 = -3$ | $-3^2 = 9$ |
| 32 | $32 - 36 = -4$ | $-4^2 = 16$ |
| $\sum x_i = 288$ | $\sum (x_i - \bar{x}) = 0$ | $\sum (x_i - \bar{x})^2 = 58$ |

*continues*

*continues*

The sum of squares (last column) $SS = \Sigma(x_i - \bar{x})^2 = 0 + 4 + 9 + 16 + 0 + 4 + 9 + 16 = 58$.

The variance is $s^2 = \frac{SS}{n - 1} = \frac{58}{8 - 1} = 8.286\ (\text{mg/m}^3)^2$.

The standard deviation is the square root of the variance: $s = \sqrt{8.286\ (\text{mg/m}^3)^2} = 2.88\ \mu\text{g/m}^3$. ■

## *Facts About the Standard Deviation*

1. **Deviations.** The deviations of a data set always sum to 0. (Data points balance perfectly around the mean in positive and negative directions.) Squaring deviations makes their signs unimportant, so the sum of the squared deviations will be always be a positive value.
2. **Units.** The variance carries "units squared." The standard deviation carries the same units as the data, making it a better choice for descriptive purposes.
3. **No variability → standard deviation = 0.** When all values in the data set are the same, there is no spread and the variance and standard deviation equal 0. In all other instances, these statistics are positive values.
4. **Degrees of freedom.** The variance is the average of the sum of squares, with the sum of squares divided by $n - 1$ instead of $n$. The number $n - 1$ is the **degrees of freedom** of the variance. You lose one degree of freedom because knowing $n - 1$ of the deviations determines the last deviation.
5. **The standard deviation is sensitive to outliers and skews.** Like the mean, the standard deviation is not resistant to outliers or strong skews. Consider the data set {4, 7, 8, 11, 12}. These data have mean $\bar{x} = 8.4$ and standard deviation $s = 3.2$. Had we made a data entry problem and entered data as {4, 7, 8, 11, 120}, $\bar{x}$ would go to 30.0 and $s$ would go to 50.4. In contrast, the median and IQR would be unaffected.
6. **Standard deviations are useful when making comparisons.** The greater the variability within a group, the larger its standard deviation. For example, if the standard deviation of an age variable in one group is 15 years and that of another group is 7 years, the first group has much greater age variability than the second.
7. **Percentage of data points falling in a range.** The standard deviation can be used to describe the percentage of the data points that will fall in

certain ranges. Two rules are applied: One rule is **Chebychev's rule**; the other rule is the **Normal rule**.

**Chebychev's rule** applies to all data set, regardless of their shape. It says that *at least* three-fourths of the data points will lie within two standard deviations of the mean. These boundaries are $\bar{x} \pm 2s$. For example, if a data set has a mean age of 30 years and a standard deviation of 10 years, then *at least* three-quarters of the values lie in the range $30 \pm (2)(10) = 10$ to 50. The key phrase here is *at least*; it is possible that more than three-fourths of the individuals fall in the range (maybe even 100%).

The **Normal rule** (also known as the **68-95-99.7 rule**) applies only to distributions with a particular **Normal** shape. We will cover the Normal distribution in Chapter 7, but for now it is important not to confuse the term ***N***ormal (meaning a particular bell shape) with the common term ***n***ormal (meaning "typical"). Many natural distributions are *not* Normal.[d] However, when a distribution is Normal

- 68% of the data points will lie within one standard deviation of the mean ($\bar{x} \pm s$).
- 95% of data points lie within two standard deviations of the mean ($\bar{x} \pm 2s$).
- 99.7% of the data points lie within three standard deviations of the mean ($\bar{x} \pm 3s$).

For example, a Normal distribution with a mean of 30 and standard deviation of 10 has 68% of its values in the range $30 \pm (1)(10) = 20$ to 40, 95% of its values in the range $30 \pm (2)(10) = 10$ to 50, and 99.7% of its values in the range $30 \pm (3)(10) = 0$ to 60.

By putting Chebychev and the Normal rule together, we can say that between 75% and 95% of the data points usually fall within two standard deviations of the mean.

## *Exercises*

**4.7** ***Spread.*** Each of the following batches of numbers has a mean of 100. Which has the most variability? (Arithmetic *not* required.)

[d]Elveback, L. R., Guillier, C. L., & Keating, F. R., Jr. (1970). Health, normality, and the ghost of Gauss. *JAMA*, *211*(1), 69–75.

| | | | | | |
|---|---|---|---|---|---|
| Batch A: | 0 | 50 | 100 | 150 | 200 |
| Batch B: | 50 | 75 | 100 | 125 | 150 |
| Batch C: | 75 | 88 | 100 | 113 | 125 |

**4.8** ***Standard deviation for site 1.*** Use a step-by-step approach to calculate the standard deviation of the data for site 1 listed in **Table 4.2**. Compare this standard deviation to that of the data from site 2 ($s_2 = 2.88\ \mu g/m^3$, pp. 79–80). How does this numerical comparison relate to what you see in **Figure 4.5**?

**4.9** ***Standard deviation via technology.*** In practice, we normally use statistical calculators or computers to calculate standard deviations. Using your statistical calculator or computer, calculate the standard deviations and variances for the air samples data originally presented in **Table 4.2**. Make certain results agree with prior hand-calculations.

**Notes**

1. **Sample standard deviation and population standard deviation.** Some calculators and programs have two different standard deviations formulas. One is the same as ours; this is called the sample standard deviation ($s$). The other is for the population standard deviation ($\sigma$).[e] In Excel (Microsoft Corp.), for instance, the sample standard deviation corresponds to the =STDEV function and the population standard deviation formula corresponds to =STDEVP. Hand calculators often have two different keys, one labeled $s$ and one labeled $\sigma$. In practice, you should always use the sample standard deviation formula ($s$) to calculate the standard deviation of a data set.
2. **Calculators and applets.** If you do not have a statistical calculator or software package, consider using one of the many free statistical applets on the Web. The Web site http://statpages .org/ lists about a dozen free online applets that calculate summary statistics.[f] The program *WinPepi* > Describe.exe can also be used for this purpose.[g] There is a link to download *WinPepi* on the Web site for this book.

[e]Population standard deviation. $\sigma = \sqrt{\frac{1}{n}\sum (x_i - \mu)^2}$ There is no loss of one degree of freedom when calculating the population standard deviation because the mean is not derived from the data. When $n$ is large, $(n - 1) \approx n$ and $s \approx \sigma$.

[f]Pezzullo, J. C. (2006). *StatPages.net: Web Pages that Perform Statistical Calculations.* Available: http://statpages.org/.

[g]Abramson, J. H. (2004). *WINPEPI* (PEPI-for-Windows): Computer programs for epidemiologists. *Epidemiologic Perspectives & Innovations*, *1*(1), 6.

**4.10** ***Heart rate.*** An individual with an irregular heartbeat is given a medication to stabilize his condition. Heart rates (beats per minute) before and after treatment are shown here.[h] Determine the means and standard deviations before and after treatment. Did the drug work?

| | | | | | | |
|---|---|---|---|---|---|---|
| Before: | 65 | 85 | 90 | 65 | 55 | 60 |
| After: | 68 | 70 | 69 | 70 | 71 | 72 |

**4.11** ***Units of measure changes numeric values of a standard deviation.*** Calculate the standard deviations of the batches of numbers here. Which batch has the greatest variability?

| | | | |
|---|---|---|---|
| Batch A: | 0 years | 1 year | 2 years |
| Batch B: | 0 months | 12 months | 24 months |
| Batch C: | 0 days | 365 days | 730 days |

This exercise demonstrates a potential problem when comparing standard deviations for variables using different units of measurement. The three batches of numbers describe identical information, but their standard deviations differ. Reporting units of measure mitigates the problem: Batch A has a standard deviation of 1 year, batch 2 has a standard deviation of 12 months, and batch 3 has a standard deviation of 365 days.

When it is necessary to compare standard deviations of measurements on different scales, convert each standard deviation to a **coefficient of variation (*CV*)**, defined as

$$CV = \frac{s}{\bar{x}}$$

The *CV* removes the units of measure and expresses the standard deviation in relation to the size of the mean. This allows for direct comparison of numerical results. In Exercise 4.11, batch A has *CV* = 1 year/1 year = 1, batch 2 has *CV* = 12 months/12 months = 1, and batch 3 has *CV* = 365 days/365 days = 1. All three *CVs* are equal to 1.

**4.12** ***Test scores with mean 100 and standard deviation 10.*** Test scores have a Normal distribution with a mean of 100 and standard deviation of 10. What percentage of scores fall in the range 80 to 120? Explain.

[h]Data are fictitious and are stored online in the file *HEARTRATE.**.

## 4.8 Selecting Summary Statistics

How do you choose which summary statistics to use in a given situation? Each situation differs, and there is no "one size fits all" solution, but certain general rules apply.

1. Always report the sample size ($n$), a measure of central location, and a measure of spread. The mean should be accompanied by the standard deviation. The median should be accompanied by the IQR or quartiles.
2. Use the mean/standard deviation when the distribution is symmetrical. You should use the median/quartiles when the distribution is asymmetrical or has outliers.
3. Accompany summary statistics with plots, when possible. Graphs can provide information that escapes numerical summaries: "You can observe a lot by watching."[i]

### *Exercises*

**4.13** ***Which statistics?*** Which measures of central location and spread would you use to describe each of the datasets depicted in **Figure 4.3**?

**4.14** ***Effect of removing an outlier.***[j] Exercise 4.6 looked at months between bacterial meningitis and the onset of seizures in 13 cases. Data were {0.10, 0.25, 0.50, 4, 12, 12, 24, 24, 31, 36, 42, 55, 96}.

(a) Calculate the mean and standard deviation for these data.

(b) Calculate the median and IQR.

(c) Remove the outlier (96) and recalculate the mean and standard deviation. Also, recalculate the median and IQR. What effect did removing the outlier have on the mean and standard deviation? What effect did it have on the median and IQR?

### *Vocabulary*

Bimodal
Box-and-whiskers plot (Boxplots)
Chebychev's rule
Coefficient of variation ($CV$)
Degrees of freedom
Depth

[i] Lawrence Peter (Yogi) Berra.
[j] You may wish to use an applet or statistical calculator to complete calculations for this problem.

Deviation
Fence
Five-point summary
Hinges
Hinge spread
Inside value
Interquartile range (IQR)
Mean
Median
Mode
Normal rule (68-95-99.7 rule)
Outside values
Q0 (quartile 0; minimum)
Q1 (quartile 1; 25th percentile)
Q2 (quartile 2; median)
Q3 (quartile 3; 75th percentile)
Q4 (quartile 4; maximum)
Quartiles
Range
Resistant
Standard deviation
Sum of squares
Unimodal
Variability
Variance
Whisker spread

## *Exercises*

**4.15** ***Leaves on stems.*** Calculate the mean and standard deviation of each group depicted in each of the side-by-side stemplots in this problem. Discuss how these statistics relate to what you see.

(a) Comparison A:

```
Group 1|  |Group 2
-------------
       0|1|
        |2|
       0|3|0
        |4|0
       0|5|0
        |6|0
       0|7|0
        |8|
       0|9|
         ×10
```

(b) Comparison B:

```
Group 1| |Group 2
-------------
        |1|
        |2|
        |3|0
        |4|0
       0|5|0
       0|6|0
       0|7|0
       0|8|
       0|9|
         ×10
```

(c) Comparison C:

```
Group 1| |Group 2
-------------
       0|1|
        |2|
       0|3|
        |4|
       0|5|0
        |6|0
       0|7|0
        |8|0
       0|9|0
         ×10
```

**4.16** ***Irish health care Web sites.*** **Table 3.1** in the prior chapter considered the reading levels of Irish health care Web sites. Here's a reissue of the stemplot for the data:

```
08|0
09|
10|0
11|00
12|0
13|0000
14|000
15|0000000
16|0
17|00000000000000000000000
×1
```

(a) Which measure of location and spread would you use to describe this distribution?
(b) Calculate the five-point summary for the distribution.
(c) Does the data set have any outside values? (Show all your work.)

**4.17** ***Health insurance by state.*** **Table 4.3** lists the percent of people without health insurance by state.
(a) Calculate the mean and median of these data. Compare these statistics. What does this tell you about the shape of the distribution?
(b) Determine the five-point summary for the data.
(c) Are there any outside values in this dataset?

**4.18** ***Skinfold thickness.*** Skinfold thickness over the triceps muscle in the arm is an anthropometric measure that varies with states of health. **Table 4.4** lists skinfold measurements at the midpoint of the triceps in five men with chronic lung disease and six comparably aged controls. Compare the groups with side-by-side stemplots. Then calculate group means and standard deviations.

**4.19** ***What would you report?*** A small data set ($n = 9$) has the following values {3.5, 8.1, 7.4, 4.0, 0.7, 4.9, 8.4, 7.0, 5.5}. Plot the data as a stemplot and then report an appropriate measures of central location and spread for the data.

**Table 4.3** Percent of residents without health insurance by state, U.S., 2004, $n = 51$, presented in rank order.

| State | % w/o insurance | depth | State | % w/o insurance | depth |
|---|---|---|---|---|---|
| Minnesota | 8.5 | 1 | Kentucky | 13.9 | 27 |
| Hawaii | 9.9 | 2 | Maryland | 14.0 | 28 |
| Iowa | 10.1 | 3 | Illinois | 14.2 | 29 |
| Wisconsin | 10.4 | 4 | Washington | 14.2 | 30 |
| Rhode Island | 10.5 | 5 | New Jersey | 14.4 | 31 |
| Vermont | 10.5 | 6 | New York | 15.0 | 32 |
| Maine | 10.6 | 7 | West Virginia | 15.9 | 33 |
| New Hampshire | 10.6 | 8 | Wyoming | 15.9 | 34 |
| Kansas | 10.8 | 9 | Oregon | 16.1 | 35 |
| Massachusetts | 10.8 | 10 | Georgia | 16.6 | 36 |
| Connecticut | 10.9 | 11 | North Carolina | 16.6 | 37 |
| Nebraska | 11.0 | 12 | Arkansas | 16.7 | 38 |
| North Dakota | 11.0 | 13 | Colorado | 16.8 | 39 |
| Michigan | 11.4 | 14 | Arizona | 17.0 | 40 |
| Pennsylvania | 11.5 | 15 | Mississippi | 17.2 | 41 |
| Missouri | 11.7 | 16 | Idaho | 17.3 | 42 |
| Delaware | 11.8 | 17 | Montana | 17.9 | 43 |
| Ohio | 11.8 | 18 | Alaska | 18.2 | 44 |
| South Dakota | 11.9 | 19 | California | 18.4 | 45 |
| Tennessee | 12.7 | 20 | Florida | 18.5 | 46 |
| Utah | 13.4 | 21 | Louisiana | 18.8 | 47 |
| Alabama | 13.5 | 22 | Nevada | 19.1 | 48 |
| Dist. of Columbia | 13.5 | 23 | Oklahoma | 19.2 | 49 |
| Virginia | 13.6 | 24 | New Mexico | 21.4 | 50 |
| Indiana | 13.7 | 25 | Texas | 25.1 | 51 |
| South Carolina | 13.8 | 26 | | | |

Source: DeNavas-Walt, C., Proctor, B. D., & Lee, C. H. (2005). *Income, Poverty, and Health Insurance Coverage in the United States: 2004* (No. P60-229). Washington, DC: U.S. Government Printing Office.

Data are stored in INC-POV-HLTHINS.SAV as the variable NOINS (no insurance).

**Table 4.4** Data for exercise 4.18; skinfold thickness over the triceps (millimeters).

| | | | | | | |
|---|---|---|---|---|---|---|
| Chronic lung disease: | 9.1 | 10.9 | 11.4 | 15.3 | 18.4 | |
| Controls: | 10.4 | 19.6 | 20.6 | 23.8 | 24.7 | 32.8 |

Data are fictitious but realistic and are stored online in the file SKINFOLD.*.