Notice that the "average square deviation" in the sample variance s^2 actually divides the sum by 1 less than the number of observations—that is, by $n-1$ rather than by n. The reason is that the deviations $x_i - \bar{x}$ always sum to exactly 0, so knowing $n-1$ of them determines the last one. Only $n-1$ of the squared deviations in a sample can vary freely once we know $\bar{x}$, and we find their average by dividing the total by $n-1$. This adjustment is necessary to give the sample variance important mathematical properties that will be addressed in later chapters on statistical inference. The number $n-1$ is called the **degrees of freedom** of the sample variance or sample standard deviation. Some calculators and software offer a choice between dividing by n and dividing by $n-1$, so be sure to use $n-1$ when calculating the sample standard deviation. For instance, the TI-83 calculator output in Figure 2.1 gives both the sample standard deviation, "$\mathrm{S_X} = 1.588\ldots$," and the value you would get using n in the denominator, "σx = 1.534..." (We will see in Chapter 13 that the notation for the population standard deviation is σ.)

More important than the details of hand calculation are the properties that determine the usefulness of the sample standard deviation:

- s measures *spread about the mean*[7] and should be used only when the mean is chosen as the measure of center. Although this is a gross simplification, you might find it helpful to think of the standard deviations as representing roughly the average dispersion, in either direction, of the n data points relative to their mean.
- s is *always zero or greater than zero.* $s = 0$ only when there is no spread (the values in the sample are all identical). Otherwise, $s > 0$. As the observations become more spread out about their mean, s gets larger.
- s has the *same units of measurement as the original observations*. For example, if you measure metabolic rates in kilocalories, both the mean $\bar{x}$ and the standard deviation s are also in kilocalories. This is one reason to prefer s to the variance s^2, which in this case is measured in squared kilocalories (cal^2). The variance, however, has more interesting mathematical properties that are beyond the scope of this chapter.
- Like the mean $\bar{x}$, s is *not resistant*. Outliers and skew increase the spread of a distribution and, therefore, also increase s.

The use of squared deviations renders s even more sensitive than $\bar{x}$ to a few extreme observations. In Example 2.4, we saw that the mean number of flood events between 2009 and 2013 for the 52 East Coast communities was $\bar{x} = 52.3$, but that changing the largest observation from 250 to 2500 would increase the mean to 98.6. This hypothetical typo would also result in a much larger standard deviation of 344.6 instead of the true standard deviation $s = 65.0$.

APPLY YOUR KNOWLEDGE

2.4 Mercury levels in pregnant women. A study recorded the blood mercury levels (in micrograms per liter, μg/l) of 4134 pregnant British women enrolled in the Avon Longitudinal Study of Parents and Children.[8] The published findings include the following statement: "Blood mercury levels ranged from 0.17 to 12.8 μg/l. The 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles were 0.81, 0.99, 1.35, 1.86, 2.52, 3.33, and 4.02 μg/l, respectively."

a. Find the five-number summary of the distribution of blood mercury levels among the pregnant women enrolled in this study.

b. What are the range and the interquartile range for this distribution?

c. Do you have enough information to compute the standard deviation of this sample of blood mercury levels? Explain your reasoning.

2.5 Spider silk, continued. In Exercise 2.1 you plotted the silk yield stress for 21 female golden orb weaver spiders.

a. Obtain the five-number summary of the distribution of yield stresses.

b. Obtain the mean and the standard deviation of the sample of yield stresses.

c. Which summary gives more information about the distribution of silk yield stresses? How do they reflect what you see in your dotplot? Remember that a summary for a quantitative variable, no matter how detailed, will never be as informative as a graph of the raw data.

Jonathan Heger/istockphoto

APPLY YOUR KNOWLEDGE

2.4 Mercury levels in pregnant women. A study recorded the blood mercury levels (in micrograms per liter, μg/l) of 4134 pregnant British women enrolled in the Avon Longitudinal Study of Parents and Children.[8] The published findings include the following statement: "Blood mercury levels ranged from 0.17 to 12.8 μg/l. The 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles were 0.81, 0.99, 1.35, 1.86, 2.52, 3.33, and 4.02 μg/l, respectively."

a. Find the five-number summary of the distribution of blood mercury levels among the pregnant women enrolled in this study.

b. What are the range and the interquartile range for this distribution?

c. Do you have enough information to compute the standard deviation of this sample of blood mercury levels? Explain your reasoning.

2.5 Spider silk, continued. In Exercise 2.1 you plotted the silk yield stress for 21 female golden orb weaver spiders.

a. Obtain the five-number summary of the distribution of yield stresses.

b. Obtain the mean and the standard deviation of the sample of yield stresses.

c. Which summary gives more information about the distribution of silk yield stresses? How do they reflect what you see in your dotplot? Remember that a summary for a quantitative variable, no matter how detailed, will never be as informative as a graph of the raw data.

Jonathan Heger/istockphoto

GRAPHICAL DISPLAYS OF NUMERICAL SUMMARIES

Numerical values are always more challenging to interpret and communicate than graphical displays. For this reason, we also like to display numerical summaries in graphical format.

The five-number summary of a distribution can be displayed in a *boxplot*. Figure 2.4 shows boxplots comparing needle lengths for the Aleppo pine and Torrey pine species of Examples 2.1 and 2.2.

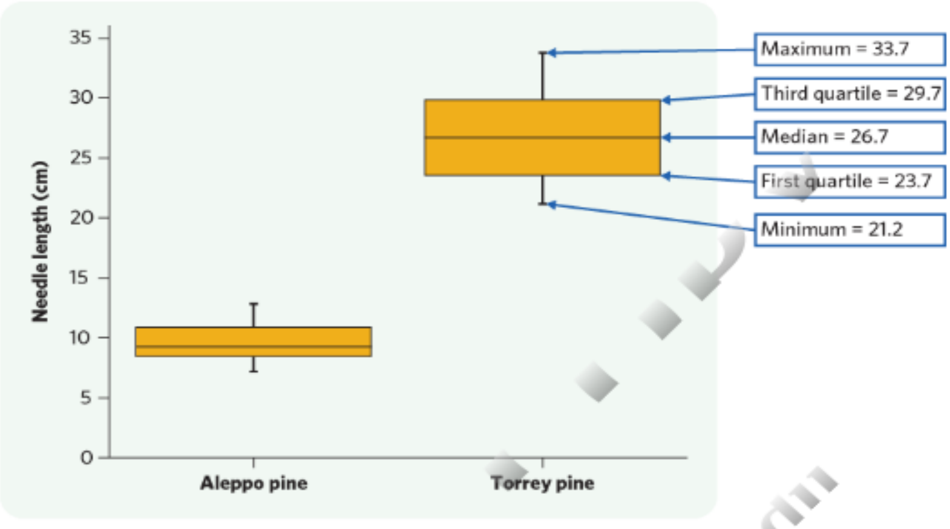


Figure 2.4
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 2.4 Boxplots comparing the lengths of 15 Aleppo pine needles and 18 Torrey pine needles.

BOXPLOT

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

- A central box spans the quartiles Q_1 and Q_3.
- A line in the box marks the median M.
- Lines extend from the box out to the smallest and largest observations.

Because boxplots show less detail than histograms and especially dotplots, they are best used for side-by-side comparison of more than one distribution, as in Figure 2.4. 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 box shows 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.

In a symmetric distribution, the first and third quartiles are equally distant from the median. In contrast, in most distributions that are skewed to the right, the third quartile will be farther above the median than the first quartile is below it. In most distributions that are skewed to the left, the first quartile will be farther below the median than the third quartile is above 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.

EXAMPLE 2.8 Needle lengths of two pine tree species

In Examples 2.1 and 2.2 we examined sample sets of needles from two distinct species of pine trees, Aleppo and Torrey. In Figure 2.4 we can see that the needles in the Torrey pine set are all longer than the needles in the Aleppo pine set: The minimum length of the Torrey pine set is larger than the maximum length for the Aleppo pine set. Torrey pine needle lengths also have greater variability, as shown by the spread of the box and the spread between the extremes. Finally, the data for the Torrey pine are symmetrical, whereas the data for the Aleppo pine are mildly right-skewed.

■

Scientific publications sometimes provide summary statistics for two or more groups in the form of a graph displaying each mean with **error bars** extending on either side to show the standard deviation in each group. Another increasingly common display combines the raw data in the form of a dotplot with the group means shown as a symbol or the height of a bar and error bars scaled to equal one standard deviation on either side of the mean. Figure 2.5 shows such a graph for the needle lengths of the two species of pine trees. *Always check the legend of such graphs, because error bars may represent some other measure of variability* (such as the standard error of the mean or the margin of error, two concepts we will discuss in later chapters).

CAUTION

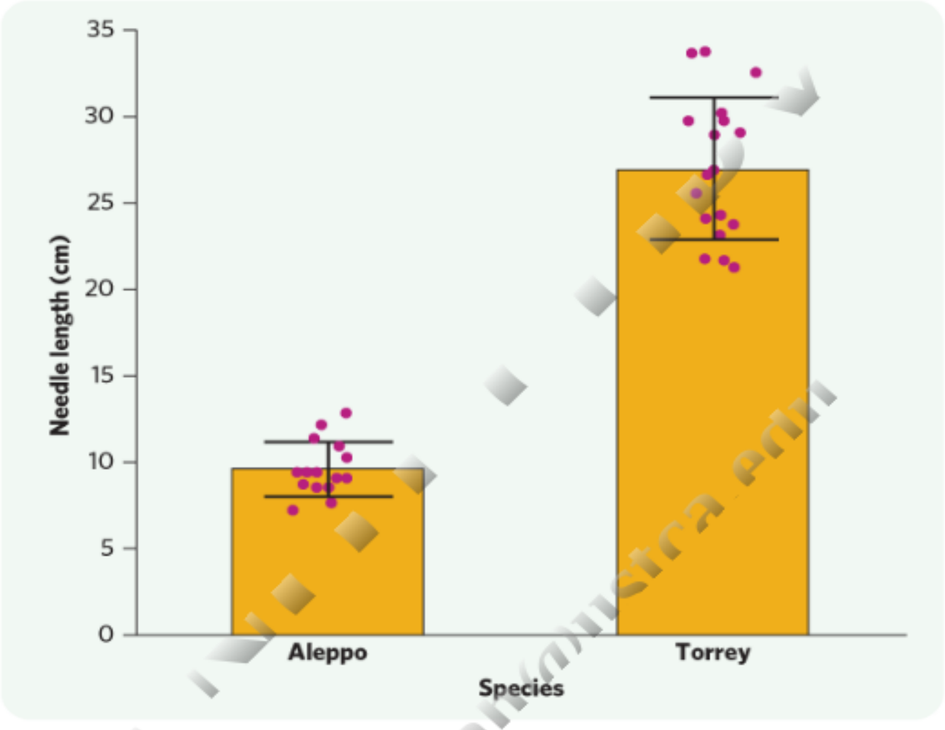


Figure 2.5

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

FIGURE 2.5 The same data as in Figure 2.4 now displayed as the mean (columns) plus and minus one standard deviation (error bars) overlaid on top of a dotplot of the raw data for both species.

APPLY YOUR KNOWLEDGE

2.6 Glucose levels. People with diabetes must monitor and control their blood glucose level. The goal is to maintain a "fasting plasma glucose" between approximately 90 and 130 milligrams per deciliter (mg/dl). Exercise 1.10 (page 21) gave the fasting plasma glucose levels for two groups of diabetics five months after they received either group instruction or individual instruction on glucose control.

a. Calculate the five-number summary for each of the two data sets.

b. Make side-by-side boxplots comparing the two groups (as in Figure 2.4). What can you say from this graph about the differences between the two diabetes control instruction methods?

c. Obtain the mean and standard deviation for each sample. Does this information give any clue about the shape of the two distributions?

the shape of the two distributions?

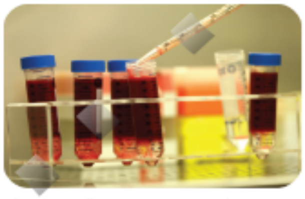
d. Add to the dotplots you created in Exercise 1.10 a symbol representing the mean of each group and error bars representing one standard deviation above and below the mean. You can do this by hand or with statistical software, if your software has this capability. Compare this graphical summary with the boxplot display you also created.

2.7 The cost of common blood tests. A 2014 study examined the prices (in dollars) billed by hospitals all over California for common blood tests such as a blood lipid panel (to check cholesterol level, first row) and a blood metabolic panel (including fasting plasma glucose level, second row). Here is how the findings were reported:[9]

N	Average	Standard deviation	Min	5th percentile	25th percentile	Median	75th percentile	95th percentile	Max
178	299	759	10	76	134	220	303	602	10,169
189	371	814	35	62	111	214	389	716	7,303

a. Make side-by-side boxplots comparing the distributions of prices for the two procedures (as in Figure 2.4). Describe the distribution of prices for each blood test. Are there substantial differences between the two?

b. The report also provided the mean and standard deviation for the two distributions. Explain why in this case the mean and standard deviation would be poor choices of summary statistics to cite in a news report.



Claudio Gallone/AgeFotostock

2.8 $\bar{x}$ and s are not enough. The mean $\bar{x}$ and standard deviation s measure center and spread 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 dotplot of each and comment on the shape of each distribution.

Data A	9.14	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74
Data B	6.58	5.76	7.71	8.84	8.47	7.04	5.25	5.56	7.91	6.89	12.50

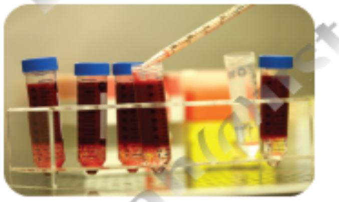
Make side-by-side boxplots comparing the two groups (as in Figure 2.4). What can you say from this graph about the differences between the two diabetes control instruction methods?

- Obtain the mean and standard deviation for each sample. Does this information give any clue about the shape of the two distributions?
- Add to the dotplots you created in Exercise 1.10 a symbol representing the mean of each group and error bars representing one standard deviation above and below the mean. You can do this by hand or with statistical software, if your software has this capability. Compare this graphical summary with the boxplot display you also created.
- **2.7 The cost of common blood tests.** A 2014 study examined the prices (in dollars) billed by hospitals all over California for common blood tests such as a blood lipid panel (to check cholesterol level, first row) and a blood metabolic panel (including fasting plasma glucose level, second row). Here is how the findings were reported:[9]

N	Average	Standard deviation	Min	5th percentile	25th percentile	Median	75th percentile	95th percentile	Max
178	299	759	10	76	134	220	303	602	10,169
189	371	814	35	62	111	214	389	716	7,303

a. Make side-by-side boxplots comparing the distributions of prices for the two procedures (as in Figure 2.4). Describe the distribution of prices for each blood test. Are there substantial differences between the two?

b. The report also provided the mean and standard deviation for the two distributions. Explain why in this case the mean and standard deviation would be poor choices of summary statistics to cite in a news report.



Claudio Gallone/AgeFotostock

- **2.8 $\bar{x}$ and s are not enough.** The mean $\bar{x}$ and standard deviation s measure center and spread 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 dotplot of each and comment on the shape of each distribution.

Data A	9.14	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74
Data B	6.58	5.76	7.71	8.84	8.47	7.04	5.25	5.56	7.91	6.89	12.50

SPOTTING SUSPECTED OUTLIERS[i]

Figure 2.6 shows the dotplot of the sizes in cubic centimeters (cm^3) of 11 acorns from oak trees found on the Pacific coast.[10] The five-number summary for this distribution is

0.4 1.6 4.1 6.0 17.1

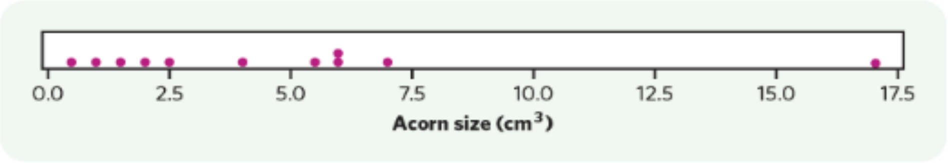


Figure 2.6
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 2.6 Dotplot of the sizes of 11 acorns.

How shall we describe the spread of this distribution? The dotplot shows that the largest observation is clearly an extreme that doesn't represent most of the data very well. The interquartile range (the range of the center half of the data) is a more resistant measure of spread. In this example, $IQR = 6.0 - 1.6 = 4.4\,\text{cm}^3$. However, the two sides of a skewed distribution have different spreads around the center of the distribution, so one number can't represent them both. The interquartile range is mainly used as the basis for a rule of thumb for identifying *suspected* outliers.

THE $1.5 \times IQR$ RULE FOR SUSPECTED OUTLIERS

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

EXAMPLE 2.9 Spotting suspected outliers: acorns

For the acorn sizes data, the five-number summary is 0.4, 1.6, 4.1, 6.0, 17.1, so

$$1.5 \times IQR = 1.5 \times 4.4 = 6.6.$$

Any values not falling between

$$\begin{aligned} Q_1 - (1.5 \times IQR) &= 1.6 - 6.6 = -5.0 \text{ and} \\ Q_3 + (1.5 \times IQR) &= 6.0 + 6.6 = 12.6 \end{aligned}$$

are flagged as *suspected* outliers. In this case, the $1.5 \times IQR$ rule flags only one value in the data set, the largest value of 17.1, and suggests that it may be an outlier. Looking at the dotplot of acorn sizes in Figure 2.6, we can confirm that the largest acorn (17.1 cm^3) is indeed an outlier.

The $1.5 \times IQR$ rule is not a replacement for looking at the data. It is most useful when large volumes of data are scanned automatically. Some software programs create **modified boxplots,** which display any value outside of the $1.5 \times IQR$ interval around either quartile as individual data points (rather than including them within the long tail going to the minimum or maximum). Examine the data carefully and decide for yourself if a particular data point fits the definition of an outlier, "an individual value that falls outside the overall pattern."

overall pattern.

EXAMPLE 2.10 Spotting suspected outliers: flood events

In Example 2.4 we looked at the number of flood events recorded between 2009 and 2013 in each of 52 Atlantic coastal communities. The five-number summary for these data is 0, 5, 37.5, 73.5, 250 and

$$1.5 \times IQR = 1.5 \times 68.5 = 102.75$$

so that any values not falling between

$$\begin{aligned} Q_1 - (1.5 \times IQR) &= 5 - 102.75 = -97.75 \text{ and} \\ Q_3 + (1.5 \times IQR) &= 73.5 + 102.75 = 176.25 \end{aligned}$$

are flagged as *suspected* outliers. We can see from the dotplot in Figure 2.2 that the four largest values are flagged by this rule. Now we must decide whether they actually are outliers.

Figure 2.7 shows a combination histogram–modified boxplot created by the statistical software JMP. The four data points flagged by the $1.5 \times IQR$ rule are displayed as individual dots rather than as part of the boxplot's high whisker. A gap appears between these four points and the rest of the high whisker, but it is not very large. In fact, the histogram underneath has a very pronounced right skew without any gap. So are the *suspected* outliers flagged by the $1.5 \times IQR$ rule *actual* outliers? They are certainly somewhat more extreme than the rest of the distribution, yet they fit well within the overall right-skew pattern.

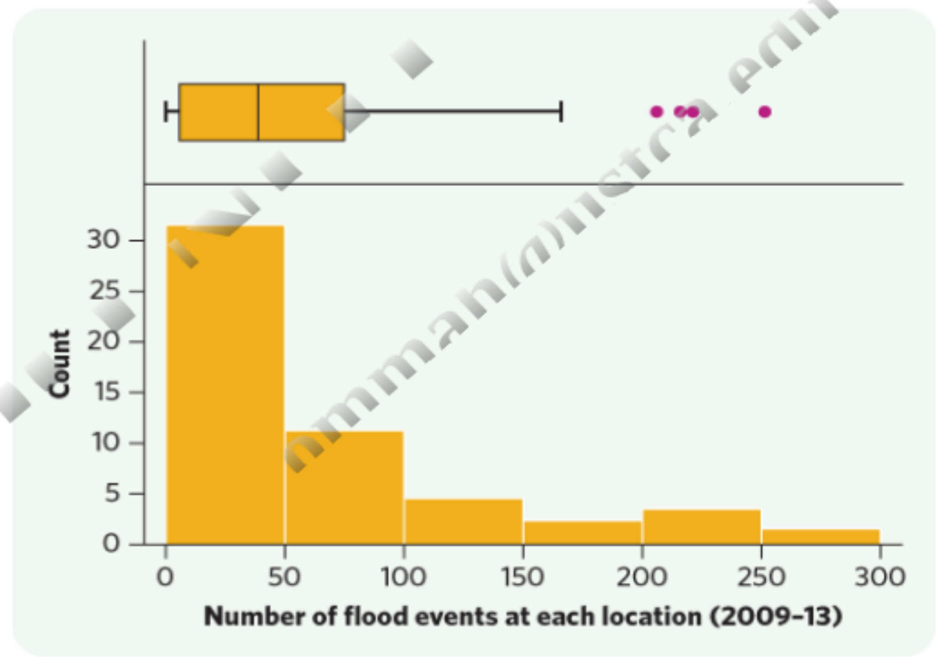


Figure 2.7
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018
W. H. Freeman and Company

FIGURE 2.7 Histogram and modified boxplot (using the same horizontal scale) of the number of flood events recorded over the 2009-2013 period for each of 52 coastal communities.

Sometimes outliers are really obvious, as in Example 2.9, and sometimes they are more ambiguous, as in Example 2.10. It can be helpful to use words such as "mild" or "moderate" versus "extreme" to describe them. The four largest numbers of flood events may be described, for instance, as possibly mild outliers within a pronounced right skew. Simply be honest and describe your findings as they are.

APPLY YOUR KNOWLEDGE

2.9 Spider silk, continued. In Exercise 2.1 you plotted the silk yield stress for 21 female golden orb weaver spiders. Use the $1.5 \times IQR$ rule to identify suspected outliers.

2.10 Glucose levels, continued. In Exercise 2.6 you made side-by-side boxplots comparing the group and individual training methods for diabetes control. Use the $1.5 \times IQR$ rule to identify any suspected outliers. Then look at the raw data to determine if unusually high or low values in either data set actually are outliers.

CDC/Amanda Mills

APPLY YOUR KNOWLEDGE

2.9 Spider silk, continued. In Exercise 2.1 you plotted the silk yield stress for 21 female golden orb weaver spiders. Use the $1.5 \times IQR$ rule to identify suspected outliers.

2.10 Glucose levels, continued. In Exercise 2.6 you made side-by-side boxplots comparing the group and individual training methods for diabetes control. Use the $1.5 \times IQR$ rule to identify any suspected outliers. Then look at the raw data to determine if unusually high or low values in either data set actually are outliers.

CDC/Amanda Mills

DISCUSSION Dealing with outliers: recognition and treatment

When collecting data, we sometimes come across wild observations that clearly fall outside the overall pattern of data distribution. In this chapter, we have presented a way to identify suspected outliers and discussed how outliers can affect numerical summaries. But what are these wild observations and how do we deal with them? Entire books have been written on the topic. Here we describe three major types of wild observations and offer some general guidance on how to handle each type.

Human error in recording information

Errors in data recording are not that uncommon. Typos are an obvious concern. In addition, the data themselves are not always clean. Surveys of individuals are particularly prone to errors. People may forget, lie, or simply misunderstand a question. In an online survey, undergraduate students enrolled in a biostatistics course were asked to record their heights in inches. Of the 149 numerical values submitted, two wild observations appeared as 5.3 and 6. These are obvious errors. Maybe the students meant 5 feet 3 inches and 6 feet, respectively, or maybe the 5.3 value was a typo for 53 inches.

What should you do with wild observations that you have clearly identified as being errors in data entry? The obvious answer is that these values do not belong and should not be included with the whole data set. You might be able to correct the mistakes by checking your original records (notes, data tables, photos). Good scientific practice always includes keeping clear and extensive records of data and the way they were obtained.

Human error in experimentation or data collection

Sloppy experimentation methods can lead to unexpected results. If you forget to add bacteria to one of your Petri dishes and find that nothing grew in it, that's just a silly mistake. But some experimental blunders lead to more interesting results.

Fleming, for instance, had less-than-ideal lab techniques. A few of his Petri dishes ended up being contaminated with a mold. Instead of simply throwing them away, Fleming noticed a halo around the mold where no bacteria grew. He went on to cultivate the mold and discover its antibiotic properties, revolutionizing medicine and later earning a Nobel Prize. When Pasteur was studying chicken cholera, his assistant left some bacterial cultures out while he went on vacation. The dried-out cultures failed to kill inoculated chickens, as other cultures usually did. The assistant's first impulse was to discard the data, but Pasteur decided to take a closer look and explore the reason for this unexpected result. This work led him to understand the workings of the immune system and develop a vaccine.

Not all technical errors lead to fame or a Nobel Prize. Some mistakes are truly not worth a second look. Sometimes, however, the wild observations can be very interesting. You should always keep detailed notes of everything you do when gathering data, as this might help you identify mistakes and understand how they arose. Either way, these kinds of wild observations do not belong with the rest of your data (including them would be like comparing apples and oranges). If you suspect an experimental error occurred, the wild observation should be either ignored or studied separately.

Unexplainable but apparently legitimate wild observations

Many studies in the life sciences are conducted by collecting data about a small sample taken from the whole population of interest. In such cases, it can be difficult to determine whether a suspected outlier in a sample is truly a wild observation or just the consequence of studying only a small subset of the population. When you find a suspected outlier in a sample and have ruled out human error, you are faced with the challenging task of deciding what to do with it.

This is an important step because, for many statistical procedures, outliers are influential and can distort conclusions. Running the analysis first with the outlier and then again without the outlier, can help you determine whether the outlier affects your conclusions substantially or not. We will see in future chapters that some statistical methods are robust against mild outliers. That is, the method will be valid despite the presence of a mild outlier. Extreme outliers, however, are always a cause of concern. Many complex statistical approaches have been devised to deal with outliers, though most are well beyond the scope of this introductory textbook. Nonparametric tests, described in optional Chapter 27, are just a few examples. For simple studies, deciding what to do with a suspected outlier typically boils down to deciding whether to include the wild observation with the rest of the data.

ORGANIZING A STATISTICAL PROBLEM

Most of our examples and exercises have been aimed at helping you learn basic tools (graphs and calculations) for describing and comparing distributions. You have also learned principles that guide the use of these tools, such as “always start with a graph” and “look for the overall pattern and striking deviations from the pattern.” The data you work with are not just numbers: They describe specific settings, such as water salinity in the Everglades or needle length for two species of pine trees. Because data come from a specific setting, the final step in examining data is drawing a conclusion for that setting. For example, water depth in the Everglades has a yearly cycle that reflects Florida’s wet and dry seasons. Needle lengths are longer in Torrey pine trees than in Aleppo pine trees.

As you learn more statistical tools and principles, you will face more complex statistical problems. Although no framework can possibly accommodate all the varied issues arising in applying statistics to real settings, the following four-step thought process gives useful guidance. In particular, the first and last steps emphasize that statistical problems are tied to specific real-world settings and, therefore, involve more than doing calculations and making graphs.

ORGANIZING A STATISTICAL PROBLEM: THE FOUR-STEP PROCESS

STATE: What is the practical question, in the context of the real-world setting?

PLAN: What specific statistical operations does this problem call for?

SOLVE: Analyze the data with graphs and computations suitable for this problem.

CONCLUDE: Give your practical conclusion in the setting of the real-world problem.