



For our data on service call lengths, $IQR = 200 - 54.5 = 145.5$. The quartiles and the IQR are not affected by changes in either tail of the distribution. They are therefore resistant, because changes in a few data points have no further effect once these points move outside the quartiles. However, *no single numerical measure of spread, such as IQR , is very useful for describing skewed distributions.* The two sides of a skewed distribution have different spreads, so one number can't summarize them. We can often detect skewness from the five-number summary by comparing how far the first quartile and the minimum are from the median (left tail) with how far the third quartile and the maximum are from the median (right tail). 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 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

1.18 Outliers for call length data. For the call length data in Table 1.1,

$$1.5 \times IQR = 1.5 \times 145.5 = 218.25$$

Any values below $54.5 - 218.25 = -163.75$ or above $200 + 218.25 = 418.25$ are flagged as possible outliers. There are no low outliers, but the 8 longest calls are flagged as possible high outliers. Their lengths are

438 465 479 700 700 951 1148 2631

modified boxplot

Statistical software often uses the $1.5 \times IQR$ rule. For example, the stemplot in Figure 1.6 lists these 8 observations separately. Boxplots drawn by software are often **modified boxplots** that plot suspected outliers individually. Figure 1.20 is a modified boxplot of the call length data. The lines extend out from the central box only to the smallest and largest observations that are not flagged by the $1.5 \times IQR$ rule. The 8 largest call lengths are plotted as individual points, though 2 of them are identical and so do not appear separately.

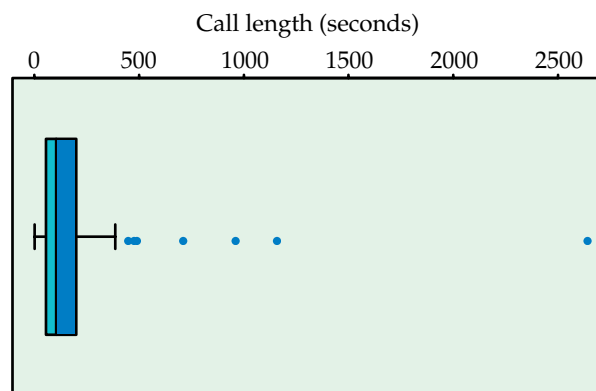


FIGURE 1.20 Modified boxplot of the call lengths in Table 1.1, for Example 1.18.

The distribution of call lengths is very strongly skewed. We may well decide that only the longest call is truly an outlier in the sense of deviating from the overall pattern of the distribution. The other 7 calls are just part of the long right tail. The $1.5 \times IQR$ rule does not remove the need to look at the distribution and use judgment. It is useful mainly to call our attention to unusual observations.

USE YOUR KNOWLEDGE

1.52 Find the IQR. Here are the scores on the first exam in an introductory statistics course for 10 students:

80 73 92 85 75 98 93 55 80 90

Find the interquartile range and use the $1.5 \times IQR$ rule to check for outliers. How low would the lowest score need to be for it to be an outlier according to this rule?



The stemplot in Figure 1.6 and the modified boxplot in Figure 1.20 tell us much more about the distribution of call lengths than the five-number summary or other numerical measures. The routine methods of statistics compute numerical measures and draw conclusions based on their values. These methods are very useful, and we will study them carefully in later chapters. But they cannot be applied blindly, by feeding data to a computer program, because *statistical measures and methods based on them are generally meaningful only for distributions of sufficiently regular shape*. This principle will become clearer as we progress, but it is good to be aware at the beginning that quickly resorting to fancy calculations is the mark of a statistical amateur. Look, think, and choose your calculations selectively.

Measuring spread: the standard deviation

The five-number summary is not the most common numerical description of a distribution. That distinction belongs to the combination of the mean to measure center and the *standard deviation* to measure spread. The standard deviation measures spread by looking at how far the observations are from their mean.

THE STANDARD DEVIATION s

The **variance** s^2 of a set of observations is the average of the squares of the deviations of the observations from their mean. In symbols, the variance of n observations x_1, x_2, \dots, 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, in more compact notation,

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

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

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

The idea behind the variance and the standard deviation as measures of spread is as follows: The deviations $x_i - \bar{x}$ display the spread of the values x_i about their mean \bar{x} . Some of these deviations will be positive and some negative because some of the observations fall on each side of the mean. In fact, *the sum of the deviations of the observations from their mean will always be zero*. Squaring the deviations makes them all positive, so that observations far from the mean in either direction have large positive squared deviations. The variance is the average squared deviation. Therefore, s^2 and s will be large if the observations are widely spread about their mean, and small if the observations are all close to the mean.

EXAMPLE

1.19 Metabolic rate. 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 calories per 24 hours. These are the same calories used to describe the energy content of foods.)

1792 1666 1362 1614 1460 1867 1439

Enter these data into your calculator or software and verify that

$\bar{x} = 1600$ calories $s = 189.24$ calories

Figure 1.21 plots these data as dots on the calorie scale, with their mean marked by an asterisk (*). The arrows mark two of the deviations from the mean. If you were calculating s by hand, you would find the first deviation as

$$x_1 - \bar{x} = 1792 - 1600 = 192$$

Exercise 1.70 asks you to calculate the seven deviations, square them, and find s^2 and s directly from the deviations. Working one or two short examples by hand helps you understand how the standard deviation is obtained. In practice you will use either software or a calculator that will find s from keyed-in data. The two software outputs in Figure 1.18 both give the variance and standard deviation for the highway mileage data.

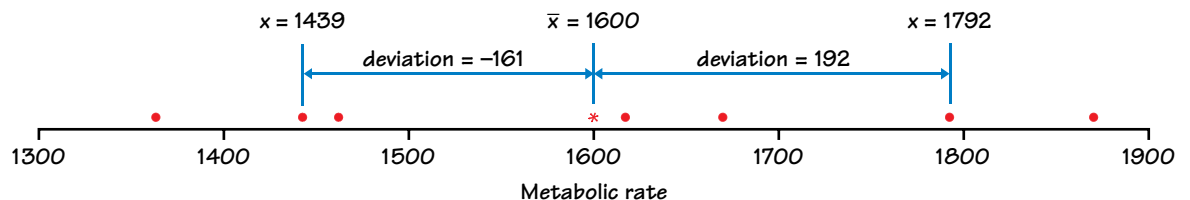


FIGURE 1.21 Metabolic rates for seven men, with the mean (*) and the deviations of two observations from the mean, for Example 1.19.

USE YOUR KNOWLEDGE

1.53 Find the variance and the standard deviation. Here are the scores on the first exam in an introductory statistics course for 10 students:

80 73 92 85 75 98 93 55 80 90

Find the variance and the standard deviation for these first-exam scores.

The idea of the variance is straightforward: it is the average of the squares of the deviations of the observations from their mean. The details we have just presented, however, raise some questions.

Why do we square the deviations?

- First, the sum of the squared deviations of any set of observations from their mean is the smallest that the sum of squared deviations from any number can possibly be. This is not true of the unsquared distances. So squared deviations point to the mean as center in a way that distances do not.
- Second, the standard deviation turns out to be the natural measure of spread for a particularly important class of symmetric unimodal distributions, the *Normal distributions*. We will meet the Normal distributions in the next section. We commented earlier that the usefulness of many statistical procedures is tied to distributions of particular shapes. This is distinctly true of the standard deviation.

Why do we emphasize the standard deviation rather than the variance?

- One reason why is that s , not s^2 , is the natural measure of spread for Normal distributions.
- There is also a more general reason to prefer s to s^2 . Because the variance involves squaring the deviations, it does not have the same unit of measurement as the original observations. The variance of the metabolic rates, for example, is measured in squared calories. Taking the square root remedies this. The standard deviation s measures spread about the mean in the original scale.

Why do we average by dividing by $n - 1$ rather than n in calculating the variance?

degrees of freedom

- Because the sum of the deviations is always zero, the last deviation can be found once we know the other $n - 1$. So we are not averaging n unrelated numbers. Only $n - 1$ of the squared deviations can vary freely, and we average by dividing the total by $n - 1$.
- The number $n - 1$ is called the **degrees of freedom** of the variance or standard deviation. Many calculators offer a choice between dividing by n and dividing by $n - 1$, so be sure to use $n - 1$.

Properties of the standard deviation

Here are the basic properties of the standard deviation s as a measure of spread.

PROPERTIES OF THE STANDARD DEVIATION

- s measures spread about the mean and should be used only when the mean is chosen as the measure of center.
- $s = 0$ only when there is *no spread*. This happens only when all observations have the same value. Otherwise, $s > 0$. As the observations become more spread out about their mean, s gets larger.
- s , like the mean \bar{x} , is not resistant. A few outliers can make s very large.

USE YOUR KNOWLEDGE

1.54 A standard deviation of zero. Construct a data set with 5 cases that has a variable with $s = 0$.



The use of squared deviations renders s even more sensitive than \bar{x} to a few extreme observations. For example, dropping the Honda Insight from our list of two-seater cars reduces the mean highway mileage from 24.7 mpg to 22.6 mpg. It cuts the standard deviation more than half, from 10.8 mpg with the Insight to 5.3 mpg without it. Distributions with outliers and strongly skewed distributions have large standard deviations. The number s does not give much helpful information about such distributions.

Choosing measures of center and spread

How do we choose between the five-number summary and \bar{x} and s to describe the center and spread of a distribution? Because the two sides of a strongly skewed distribution have different spreads, no single number such as s describes the spread well. The five-number summary, with its two quartiles and two extremes, does a better job.

CHOOSING A SUMMARY

The five-number summary is usually better than the mean and standard deviation for describing a skewed distribution or a distribution with strong outliers. Use \bar{x} and s only for reasonably symmetric distributions that are free of outliers.



EXAMPLE

1.20 Standard deviation as a measure of risk. A central principle in the study of investments is that taking bigger risks is rewarded by higher returns, at least on the average over long periods of time. It is usual in finance to measure risk by the standard deviation of returns, on the grounds that investments whose returns vary a lot from year to year are less predictable and therefore more risky than those whose returns don't vary much. Compare, for example, the approximate mean and standard deviation of the annual percent

returns on American common stocks and U.S. Treasury bills over the period from 1950 to 2003:

Investment	Mean return	Standard deviation
Common stocks	13.2%	17.6%
Treasury bills	5.0%	2.9%

Stocks are risky. They went up more than 13% per year on the average during this period, but they dropped almost 28% in the worst year. The large standard deviation reflects the fact that stocks have produced both large gains and large losses. When you buy a Treasury bill, on the other hand, you are lending money to the government for one year. You know that the government will pay you back with interest. That is much less risky than buying stocks, so (on the average) you get a smaller return.

Are \bar{x} and s good summaries for distributions of investment returns? Figure 1.22 displays stemplots of the annual returns for both investments. (Because stock returns are so much more spread out, a back-to-back stemplot does not work well. The stems in the stock stemplot are tens of percents; the stems for bills are percents. The lowest returns are -28% for stocks and 0.9% for bills.) You see that returns on Treasury bills have a right-skewed distribution. Convention in the financial world calls for \bar{x} and s because some parts of investment theory use them. For describing this right-skewed distribution, however, the five-number summary would be more informative.

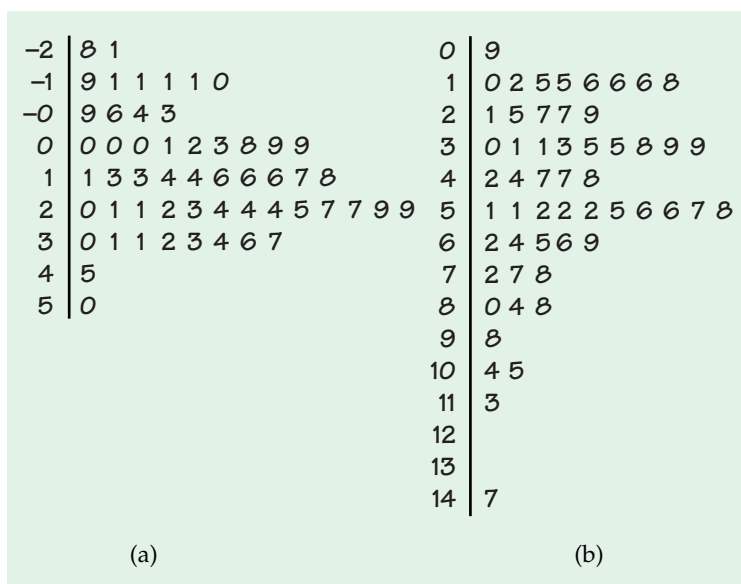


FIGURE 1.22 Stemplots of annual returns for stocks and Treasury bills, 1950 to 2003, for Example 1.20. (a) Stock returns, in whole percents. (b) Treasury bill returns, in percents and tenths of a percent.



*Remember that a graph gives the best overall picture of a distribution. Numerical measures of center and spread report specific facts about a distribution, but they do not describe its entire shape. Numerical summaries do not disclose the presence of multiple modes or gaps, for example. **Always plot your data.***

Changing the unit of measurement

The same variable can be recorded in different units of measurement. Americans commonly record distances in miles and temperatures in degrees Fahrenheit, while the rest of the world measures distances in kilometers and temperatures in degrees Celsius. Fortunately, it is easy to convert numerical descriptions of a distribution from one unit of measurement to another. This is true because a change in the measurement unit is a *linear transformation* of the measurements.

LINEAR TRANSFORMATIONS

A **linear transformation** changes the original variable x into the new variable x_{new} given by an equation of the form

$$x_{\text{new}} = a + bx$$

Adding the constant a shifts all values of x upward or downward by the same amount. In particular, such a shift changes the origin (zero point) of the variable. Multiplying by the positive constant b changes the size of the unit of measurement.

EXAMPLE

1.21 Change the units.

- (a) If a distance x is measured in kilometers, the same distance in miles is

$$x_{\text{new}} = 0.62x$$

For example, a 10-kilometer race covers 6.2 miles. This transformation changes the units without changing the origin—a distance of 0 kilometers is the same as a distance of 0 miles.

- (b) A temperature x measured in degrees Fahrenheit must be reexpressed in degrees Celsius to be easily understood by the rest of the world. The transformation is

$$x_{\text{new}} = \frac{5}{9}(x - 32) = -\frac{160}{9} + \frac{5}{9}x$$

Thus, the high of 95°F on a hot American summer day translates into 35°C. In this case

$$a = -\frac{160}{9} \quad \text{and} \quad b = \frac{5}{9}$$

This linear transformation changes both the unit size and the origin of the measurements. The origin in the Celsius scale (0°C, the temperature at which water freezes) is 32° in the Fahrenheit scale.

Linear transformations do not change the shape of a distribution. If measurements on a variable x have a right-skewed distribution, any new variable x_{new}

obtained by a linear transformation $x_{\text{new}} = a + bx$ (for $b > 0$) will also have a right-skewed distribution. If the distribution of x is symmetric and unimodal, the distribution of x_{new} remains symmetric and unimodal.

Although a linear transformation preserves the basic shape of a distribution, the center and spread will change. Because linear changes of measurement scale are common, we must be aware of their effect on numerical descriptive measures of center and spread. Fortunately, the changes follow a simple pattern.

EXAMPLE

1.22 Use scores to find the points. In an introductory statistics course, homework counts for 300 points out of a total of 1000 possible points for all course requirements. During the semester there were 12 homework assignments and each was given a grade on a scale of 0 to 100. The maximum total score for the 12 homework assignments is therefore 1200. To convert the homework scores to final grade points, we need to convert the scale of 0 to 1200 to a scale of 0 to 300. We do this by multiplying the homework scores by $300/1200$. In other words, we divide the homework scores by 4. Here are the homework scores and the corresponding final grade points for 5 students:

Student	1	2	3	4	5
Score	1056	1080	900	1164	1020
Points	264	270	225	291	255

These two sets of numbers measure the same performance on homework for the course. Since we obtained the points by dividing the scores by 4, the mean of the points will be the mean of the scores divided by 4. Similarly, the standard deviation of points will be the standard deviation of the scores divided by 4.

USE YOUR KNOWLEDGE

1.55 Calculate the points for a student. Use the setting of Example 1.22 to find the points for a student whose score is 950.

Here is a summary of the rules for linear transformations:

EFFECT OF A LINEAR TRANSFORMATION

To see the effect of a linear transformation on measures of center and spread, apply these rules:

- Multiplying each observation by a positive number b multiplies both measures of center (mean and median) and measures of spread (interquartile range and standard deviation) by b .

- Adding the same number a (either positive or negative) to each observation adds a to measures of center and to quartiles and other percentiles but does not change measures of spread.

In Example 1.22, when we converted from score to points, we described the transformation as dividing by 4. The multiplication part of the summary of the effect of a linear transformation applies to this case because division by 4 is the same as multiplication by 0.25. Similarly, the second part of the summary applies to subtraction as well as addition because subtraction is simply the addition of a negative number.

The measures of spread IQR and s do not change when we add the same number a to all of the observations because adding a constant changes the location of the distribution but leaves the spread unaltered. You can find the effect of a linear transformation $x_{\text{new}} = a + bx$ by combining these rules. For example, if x has mean \bar{x} , the transformed variable x_{new} has mean $a + b\bar{x}$.

SECTION 1.2 Summary

A numerical summary of a distribution should report its **center** and its **spread** or **variability**.

The **mean** \bar{x} and the **median** M describe the center of a distribution in different ways. The mean is the arithmetic average of the observations, and the median is their midpoint.

When you use the median to describe the center of the distribution, describe its spread by giving the **quartiles**. The **first quartile** Q_1 has one-fourth of the observations below it, and the **third quartile** Q_3 has three-fourths of the observations below it.

The **interquartile range** is the difference between the quartiles. It is the spread of the center half of the data. The **$1.5 \times IQR$ rule** flags observations more than $1.5 \times IQR$ beyond the quartiles as possible outliers.

The **five-number summary** consisting of the median, the quartiles, and the smallest and largest individual observations provides a quick overall description of a distribution. The median describes the center, and the quartiles and extremes show the spread.

Boxplots based on the five-number summary are useful for comparing several distributions. The box spans the quartiles and shows the spread of the central half of the distribution. The median is marked within the box. Lines extend from the box to the extremes and show the full spread of the data. In a **modified boxplot**, points identified by the $1.5 \times IQR$ rule are plotted individually.

The **variance** s^2 and especially its square root, the **standard deviation** s , are common measures of spread about the mean as center. The standard deviation s is zero when there is no spread and gets larger as the spread increases.

A **resistant measure** of any aspect of a distribution is relatively unaffected by changes in the numerical value of a small proportion of the total number of observations, no matter how large these changes are. The median and quartiles are resistant, but the mean and the standard deviation are not.

The mean and standard deviation are good descriptions for symmetric distributions without outliers. They are most useful for the Normal distributions introduced in the next section. The five-number summary is a better exploratory summary for skewed distributions.

Linear transformations have the form $x_{\text{new}} = a + bx$. A linear transformation changes the origin if $a \neq 0$ and changes the size of the unit of measurement if $b > 0$. Linear transformations do not change the overall shape of a distribution. A linear transformation multiplies a measure of spread by b and changes a percentile or measure of center m into $a + bm$.

Numerical measures of particular aspects of a distribution, such as center and spread, do not report the entire shape of most distributions. In some cases, particularly distributions with multiple peaks and gaps, these measures may not be very informative.

SECTION 1.2 Exercises

For Exercise 1.47, see page 32; for Exercise 1.48, see page 33; for Exercise 1.49, see page 35; for Exercises 1.50, see page 37; for Exercise 1.51, see page 38; for Exercise 1.52, see page 40; for Exercise 1.53, see page 42; for Exercise 1.54, see page 43; and for Exercise 1.55, see page 46.

1.56 Longleaf pine trees. The Wade Tract in Thomas County, Georgia, is an old-growth forest of longleaf pine trees (*Pinus palustris*) that has survived in a relatively undisturbed state since before the settlement of the area by Europeans. A study collected data about 584 of these trees.²⁷ One of the variables measured was the diameter at breast height (DBH). This is the diameter of the tree at 4.5 feet and the units are centimeters (cm). Only trees with DBH greater than 1.5 cm were sampled. Here are the diameters of a random sample of 40 of these trees:

10.5	13.3	26.0	18.3	52.2	9.2	26.1	17.6	40.5	31.8
47.2	11.4	2.7	69.3	44.4	16.9	35.7	5.4	44.2	2.2
4.3	7.8	38.1	2.2	11.4	51.5	4.9	39.7	32.6	51.8
43.6	2.3	44.6	31.5	40.3	22.3	43.3	37.5	29.1	27.9

- Find the five-number summary for these data.
- Make a boxplot.
- Make a histogram.
- Write a short summary of the major features of this distribution. Do you prefer the boxplot or the histogram for these data?

1.57 Blood proteins in children from Papua New Guinea. C-reactive protein (CRP) is a substance that can be measured in the blood. Values increase substantially within 6 hours of an infection and

reach a peak within 24 to 48 hours after. In adults, chronically high values have been linked to an increased risk of cardiovascular disease. In a study of apparently healthy children aged 6 to 60 months in Papua New Guinea, CRP was measured in 90 children.²⁸ The units are milligrams per liter (mg/l). Here are the data from a random sample of 40 of these children:

0.00	3.90	5.64	8.22	0.00	5.62	3.92	6.81	30.61	0.00
73.20	0.00	46.70	0.00	0.00	26.41	22.82	0.00	0.00	3.49
0.00	0.00	4.81	9.57	5.36	0.00	5.66	0.00	59.76	12.38
15.74	0.00	0.00	0.00	0.00	9.37	20.78	7.10	7.89	5.53

- Find the five-number summary for these data.
- Make a boxplot.
- Make a histogram.
- Write a short summary of the major features of this distribution. Do you prefer the boxplot or the histogram for these data?

1.58




Transform the blood proteins values.

Refer to the previous exercise. With strongly skewed distributions such as this, we frequently reduce the skewness by taking a log transformation. We have a bit of a problem here, however, because some of the data are recorded as 0.00 and the logarithm of zero is not defined. For this variable, the value 0.00 is recorded whenever the amount of CRP in the blood is below the level that the measuring instrument is capable of detecting. The usual procedure in this circumstance is to add a small number to each observation before taking the logs. Transform these data by adding 1 to each observation and then taking the logarithm. Use the questions in the previous exercise as a guide to your

analysis and prepare a summary contrasting this analysis with the one that you performed in the previous exercise.

- 1.59



Vitamin A deficiency in children from Papua New Guinea. In the Papua New Guinea study that provided the data for the previous two exercises, the researchers also measured serum retinol. A low value of this variable can be an indicator of vitamin A deficiency. Below are the data on the same sample of 40 children from this study. The units are micromoles per liter ($\mu\text{mol/l}$).

1.15	1.36	0.38	0.34	0.35	0.37	1.17	0.97	0.97	0.67
0.31	0.99	0.52	0.70	0.88	0.36	0.24	1.00	1.13	0.31
1.44	0.35	0.34	1.90	1.19	0.94	0.34	0.35	0.33	0.69
0.69	1.04	0.83	1.11	1.02	0.56	0.82	1.20	0.87	0.41

Analyze these data. Use the questions in the previous two exercises as a guide.

- 1.60

Luck and puzzle solving. Children in a psychology study were asked to solve some puzzles and were then given feedback on their performance. Then they were asked to rate how luck played a role in determining their scores.²⁹ This variable was recorded on a 1 to 10 scale with 1 corresponding to very lucky and 10 corresponding to very unlucky. Here are the scores for 60 children:

1	10	1	10	1	1	10	5	1	1	8	1	10	2	1
9	5	2	1	8	10	5	9	10	10	9	6	10	1	5
1	9	2	1	7	10	9	5	10	10	10	1	8	1	6
10	1	6	10	10	8	10	3	10	8	1	8	10	4	2

Use numerical and graphical methods to describe these data. Write a short report summarizing your work.

- 1.61

College tuition and fees. Figure 1.16 (page 25) is a histogram of the tuition and fees charged by the 56 four-year colleges in the state of Massachusetts. Here are those charges (in dollars), arranged in increasing order:

4,123	4,186	4,324	4,342	4,557	4,884	5,397	6,129
6,963	6,972	8,232	13,584	13,612	15,500	15,934	16,230
16,696	16,700	17,044	17,500	18,550	18,750	19,145	19,300
19,410	19,700	19,700	19,910	20,234	20,400	20,640	20,875
21,165	21,302	22,663	23,550	24,324	25,840	26,965	27,522
27,544	27,904	28,011	28,090	28,420	28,420	28,900	28,906
28,950	29,060	29,338	29,392	29,600	29,624	29,630	29,875

Find the five-number summary and make a boxplot. What distinctive feature of the histogram do

these summaries miss? Remember that numerical summaries are not a substitute for looking at the data.

- 1.62

Outliers in percent of older residents. The stemplot in Exercise 1.21 (page 24) displays the distribution of the percents of residents aged 65 and over in the 50 states. Stemplots help you find the five-number summary because they arrange the observations in increasing order.

- (a)

Give the five-number summary of this distribution.
- (b)

Does the $1.5 \times IQR$ rule identify Alaska and Florida as suspected outliers? Does it also flag any other states?

- 1.63

Tornados and property damage. Table 1.5 (page 25) shows the average property damage caused by tornadoes over a 50-year period in each of the states. The distribution is strongly skewed to the right.

- (a)

Give the five-number summary. Explain why you can see from these five numbers that the distribution is right-skewed.
- (b)

A histogram or stemplot suggests that a few states are outliers. Show that there are *no* suspected outliers according to the $1.5 \times IQR$ rule. You see once again that a rule is not a substitute for plotting your data.
- (c)

Find the mean property damage. Explain why the mean and median differ so greatly for this distribution.

- 1.64

Carbon dioxide emissions. Table 1.6 (page 26) gives carbon dioxide (CO_2) emissions per person for countries with population at least 20 million. The distribution is strongly skewed to the right. The United States and several other countries appear to be high outliers.

- (a)

Give the five-number summary. Explain why this summary suggests that the distribution is right-skewed.
- (b)

Which countries are outliers according to the $1.5 \times IQR$ rule? Make a stemplot or histogram of the data. Do you agree with the rule's suggestions about which countries are and are not outliers?

- 1.65

Median versus mean for net worth. A report on the assets of American households says that the median net worth of households headed by someone aged less than 35 years is \$11,600. The mean net worth of these same young households is \$90,700.³⁰ What explains the difference between these two measures of center?

1.66 Mean versus median for oil wells. Exercise 1.39 (page 28) gives data on the total oil recovered from 64 wells. Your graph in that exercise shows that the distribution is clearly right-skewed.

(a) Find the mean and median of the amounts recovered. Explain how the relationship between the mean and the median reflects the shape of the distribution.

(b) Give the five-number summary and explain briefly how it reflects the shape of the distribution.

1.67 Mean versus median. A small accounting firm pays each of its five clerks \$35,000, two junior accountants \$80,000 each, and the firm's owner \$320,000. What is the mean salary paid at this firm? How many of the employees earn less than the mean? What is the median salary?

1.68 Be careful about how you treat the zeros. In computing the median income of any group, some federal agencies omit all members of the group who had no income. Give an example to show that the reported median income of a group can go down even though the group becomes economically better off. Is this also true of the mean income?

1.69 How does the median change? The firm in Exercise 1.67 gives no raises to the clerks and junior accountants, while the owner's take increases to \$455,000. How does this change affect the mean? How does it affect the median?

1.70 Metabolic rates. Calculate the mean and standard deviation of the metabolic rates in Example 1.19 (page 41), showing each step in detail. First find the mean \bar{x} by summing the 7 observations and dividing by 7. Then find each of the deviations $x_i - \bar{x}$ and their squares. Check that the deviations have sum 0. Calculate the variance as an average of the squared deviations (remember to divide by $n - 1$). Finally, obtain s as the square root of the variance.

1.71 CHALLENGE Hurricanes and losses. A discussion of extreme weather says: "In most states, hurricanes occur infrequently. Yet, when a hurricane hits, the losses can be catastrophic. Average annual losses are not a meaningful measure of damage from rare but potentially catastrophic events."³¹ Why is this true?

1.72 Distributions for time spent studying. Exercise 1.41 (page 28) presented data on the nightly study time claimed by first-year college men and women. The most common methods for formal comparison of two groups use \bar{x} and s to summarize the data.

We wonder if this is appropriate here. Look at your back-to-back stemplot from Exercise 1.41, or make one now if you have not done so.

(a) What kinds of distributions are best summarized by \bar{x} and s ? It isn't easy to decide whether small data sets with irregular distributions fit the criteria. We will learn a better tool for making this decision in the next section.

(b) Each set of study times appears to contain a high outlier. Are these points flagged as suspicious by the $1.5 \times IQR$ rule? How much does removing the outlier change \bar{x} and s for each group? The presence of outliers makes us reluctant to use the mean and standard deviation for these data unless we remove the outliers on the grounds that these students were exaggerating.

1.73 The density of the earth. Many standard statistical methods that you will study in Part II of this book are intended for use with distributions that are symmetric and have no outliers. These methods start with the mean and standard deviation, \bar{x} and s . Two examples of scientific data for which standard methods should work well are the pH measurements in Exercise 1.36 (page 27) and Cavendish's measurements of the density of the earth in Exercise 1.40 (page 28).

(a) Summarize each of these data sets by giving \bar{x} and s .

(b) Find the median for each data set. Is the median quite close to the mean, as we expect it to be in these examples?


1.74 IQ scores. Many standard statistical methods that you will study in Part II of this book are intended for use with distributions that are symmetric and have no outliers. These methods start with the mean and standard deviation, \bar{x} and s . For example, standard methods would typically be used for the IQ and GPA data in Table 1.9 (page 29).

(a) Find \bar{x} and s for the IQ data. In large populations, IQ scores are standardized to have mean 100 and standard deviation 15. In what way does the distribution of IQ among these students differ from the overall population?

(b) Find the median IQ score. It is, as we expect, close to the mean.


(c) Find the mean and median for the GPA data. The two measures of center differ a bit. What feature of the data (see your stemplot in Exercise 1.43 or make a new stemplot) explains the difference?

1.75



Mean and median for two observations. The *Mean and Median* applet allows you to place observations on a line and see their mean and median visually. Place two observations on the line, by clicking below it. Why does only one arrow appear?


1.76



Mean and median for three observations. In the *Mean and Median* applet, place three observations on the line by clicking below it, two close together near the center of the line and one somewhat to the right of these two.

- (a) Pull the single rightmost observation out to the right. (Place the cursor on the point, hold down a mouse button, and drag the point.) How does the mean behave? How does the median behave? Explain briefly why each measure acts as it does.
- (b) Now drag the rightmost point to the left as far as you can. What happens to the mean? What happens to the median as you drag this point past the other two (watch carefully)?

1.77



Mean and median for five observations. Place five observations on the line in the *Mean and Median* applet by clicking below it.

(a) Add one additional observation *without changing the median*. Where is your new point?

(b) Use the applet to convince yourself that when you add yet another observation (there are now seven in all), the median does not change no matter where you put the seventh point. Explain why this must be true.

1.78

Hummingbirds and flowers. Different varieties of the tropical flower *Heliconia* are fertilized by different species of hummingbirds. Over time, the lengths of the flowers and the form of the hummingbirds' beaks have evolved to match each other. Here are data on the lengths in millimeters of three varieties of these flowers on the island of Dominica:³²

H. bihai								
47.12	46.75	46.81	47.12	46.67	47.43	46.44	46.64	
48.07	48.34	48.15	50.26	50.12	46.34	46.94	48.36	

H. caribaea red								
41.90	42.01	41.93	43.09	41.47	41.69	39.78	40.57	
39.63	42.18	40.66	37.87	39.16	37.40	38.20	38.07	
38.10	37.97	38.79	38.23	38.87	37.78	38.01		


H. caribaea yellow								
36.78	37.02	36.52	36.11	36.03	35.45	38.13	37.1	
35.17	36.82	36.66	35.68	36.03	34.57	34.63		

Make boxplots to compare the three distributions. Report the five-number summaries along with your graph. What are the most important differences among the three varieties of flower?

- 1.79 **Compare the three varieties of flowers.** The biologists who collected the flower length data in the previous exercise compared the three *Heliconia* varieties using statistical methods based on \bar{x} and s .
- (a) Find \bar{x} and s for each variety.

(b) Make a stemplot of each set of flower lengths. Do the distributions appear suitable for use of \bar{x} and s as summaries?

1.80




Effects of logging in Borneo. "Conservationists have despaired over destruction of tropical rainforest by logging, clearing, and burning." These words begin a report on a statistical study of the effects of logging in Borneo. Researchers compared forest plots that had never been logged (Group 1) with similar plots nearby that had been logged 1 year earlier (Group 2) and 8 years earlier (Group 3). All plots were 0.1 hectare in area. Here are the counts of trees for plots in each group:³³

Group 1:	27	22	29	21	19	33	16	20	24	27	28	19
Group 2:	12	12	15	9	20	18	17	14	14	2	17	19
Group 3:	18	4	22	15	18	19	22	12	12			

Give a complete comparison of the three distributions, using both graphs and numerical summaries. To what extent has logging affected the count of trees? The researchers used an analysis based on \bar{x} and s . Explain why this is reasonably well justified.

1.81



Running and heart rate. How does regular running affect heart rate? The RUNNERS data set, described in detail in the Data Appendix, contains heart rates for four groups of people:

Sedentary females

Sedentary males

Female runners (at least 15 miles per week)

Male runners (at least 15 miles per week)

The heart rates were measured after 6 minutes of exercise on a treadmill. There are 200 subjects in

each group. Give a complete comparison of the four distributions, using both graphs and numerical summaries. How would you describe the effect of running on heart rate? Is the effect different for men and women?

The *WORKERS* data set, described in the Data Appendix, contains the sex, level of education, and income of 71,076 people between the ages of 25 and 64 who were employed full-time in 2001.

The boxplots in Figure 1.23 compare the distributions of income for people with five levels of education. This figure is a variation on the boxplot idea: because large data sets often contain very extreme observations, the lines extend from the central box only to the 5th and 95th percentiles. Exercises 1.82 to 1.84 concern these data.

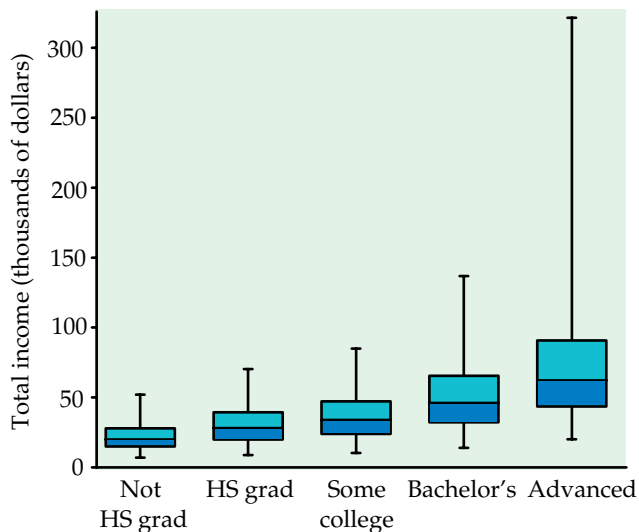


FIGURE 1.23 Boxplots comparing the distributions of income for employed people aged 25 to 64 years with five different levels of education. The lines extend from the quartiles to the 5th and 95th percentiles.

1.82 Income for people with bachelor's degrees. The data include 14,959 people whose highest level of education is a bachelor's degree.

- What is the position of the median in the ordered list of incomes (1 to 14,959)? From the boxplot, about what is the median income of people with a bachelor's degree?
- What is the position of the first and third quartiles in the ordered list of incomes for these people? About what are the numerical values of Q_1 and Q_3 ?
- You answered (a) and (b) from a boxplot that omits the lowest 5% and the highest 5% of incomes.

Explain why leaving out these values has only a very small effect on the median and quartiles.

1.83 Find the 5th and 95th percentiles. About what are the positions of the 5th and 95th percentiles in the ordered list of incomes of the 14,959 people with a bachelor's degree? Incomes outside this range do not appear in the boxplot. About what are the numerical values of the 5th and 95th percentiles of income? (For comparison, the largest income among all 14,959 people was \$481,720. That one person made this much tells us less about the group than does the 95th percentile.)

1.84 How does income change with education? Write a brief description of how the distribution of income changes with the highest level of education reached. Be sure to discuss center, spread, and skewness. Give some specifics read from the graph to back up your statements.

1.85 CHALLENGE Shakespeare's plays. Look at the histogram of lengths of words in Shakespeare's plays, Figure 1.15 (page 25). The heights of the bars tell us what percent of words have each length. What is the median length of words used by Shakespeare? Similarly, what are the quartiles? Give the five-number summary for Shakespeare's word lengths.

1.86 CHALLENGE Create a data set. Create a set of 5 positive numbers (repeats allowed) that have median 10 and mean 7. What thought process did you use to create your numbers?

1.87 Create another data set. Give an example of a small set of data for which the mean is larger than the third quartile.

1.88 CHALLENGE Deviations from the mean sum to zero. Use the definition of the mean \bar{x} to show that the sum of the deviations $x_i - \bar{x}$ of the observations from their mean is always zero. This is one reason why the variance and standard deviation use squared deviations.

1.89 CHALLENGE A standard deviation contest. This is a standard deviation contest. You must choose four numbers from the whole numbers 0 to 20, with repeats allowed.

- Choose four numbers that have the smallest possible standard deviation.
- Choose four numbers that have the largest possible standard deviation.
- Is more than one choice possible in either (a) or (b)? Explain.

- 1.90 Does your software give incorrect answers?** This exercise requires a calculator with a standard deviation button or statistical software on a computer. The observations

20,001 20,002 20,003

have mean $\bar{x} = 20,002$ and standard deviation $s = 1$. Adding a 0 in the center of each number, the next set becomes


200,001 200,002 200,003

The standard deviation remains $s = 1$ as more 0s are added. Use your calculator or computer to calculate the standard deviation of these numbers, adding extra 0s until you get an incorrect answer. How soon did you go wrong? This demonstrates that calculators and computers cannot handle an arbitrary number of digits correctly.

- 1.91 Guinea pigs.** Table 1.8 (page 29) gives the survival times of 72 guinea pigs in a medical study. Survival times—whether of cancer patients after treatment or of car batteries in everyday use—are almost always right-skewed. Make a graph to verify that this is true of these survival times. Then give a numerical summary that is appropriate for such data. Explain in simple language, to someone who knows no statistics, what your summary tells us about the guinea pigs.

- 1.92 Weight gain.** A study of diet and weight gain deliberately overfed 16 volunteers for eight weeks. The mean increase in fat was $\bar{x} = 2.39$ kilograms and the standard deviation was $s = 1.14$ kilograms. What are \bar{x} and s in pounds? (A kilogram is 2.2 pounds.)

- 1.93 Compare three varieties of flowers.** Exercise 1.78 reports data on the lengths in millimeters of flowers of three varieties of *Heliconia*. In Exercise 1.79 you found the mean and standard deviation for each variety. Starting from the \bar{x} - and s -values in millimeters, find the means and standard deviations in inches. (A millimeter is 1/1000 of a meter. A meter is 39.37 inches.)


- 1.94**  **The density of the earth.** Henry Cavendish (see Exercise 1.40, page 28) used \bar{x} to

summarize his 29 measurements of the density of the earth.


(a) Find \bar{x} and s for his data.

(b) Cavendish recorded the density of the earth as a multiple of the density of water. The density of water is almost exactly 1 gram per cubic centimeter, so his measurements have these units. In American units, the density of water is 62.43 pounds per cubic foot. This is the weight of a cube of water measuring 1 foot (that is, 30.48 cm) on each side. Express Cavendish's first result for the earth (5.50 g/cm^3) in pounds per cubic foot. Then find \bar{x} and s in pounds per cubic foot.

- 1.95 Guinea pigs.** Find the **quintiles** (the 20th, 40th, 60th, and 80th percentiles) of the guinea pig survival times in Table 1.8 (page 29). For quite large sets of data, the quintiles or the **deciles** (10th, 20th, 30th, etc. percentiles) give a more detailed summary than the quartiles.

- 1.96**  **Changing units from inches to centimeters.** Changing the unit of length from inches to centimeters multiplies each length by 2.54 because there are 2.54 centimeters in an inch. This change of units multiplies our usual measures of spread by 2.54. This is true of *IQR* and the standard deviation. What happens to the variance when we change units in this way?

- 1.97 A different type of mean.** The **trimmed mean** is a measure of center that is more resistant than the mean but uses more of the available information than the median. To compute the 10% trimmed mean, discard the highest 10% and the lowest 10% of the observations and compute the mean of the remaining 80%. Trimming eliminates the effect of a small number of outliers. Compute the 10% trimmed mean of the guinea pig survival time data in Table 1.8 (page 29). Then compute the 20% trimmed mean. Compare the values of these measures with the median and the ordinary untrimmed mean.

- 1.98**  **Changing units from centimeters to inches.** Refer to Exercise 1.56. Change the measurements from centimeters to inches by multiplying each value by 0.39. Answer the questions from the previous exercise and explain the effect of the transformation on these data.

1.3 Density Curves and Normal Distributions

We now have a kit of graphical and numerical tools for describing distributions. What is more, we have a clear strategy for exploring data on a single quantitative variable:

1. Always plot your data: make a graph, usually a stemplot or a histogram.
2. Look for the overall pattern and for striking deviations such as outliers.
3. Calculate an appropriate numerical summary to briefly describe center and spread.

Technology has expanded the set of graphs that we can choose for Step 1. It is possible, though painful, to make histograms by hand. Using software, clever algorithms can describe a distribution in a way that is not feasible by hand, by fitting a smooth curve to the data in addition to or instead of a histogram. The curves used are called **density curves**. Before we examine density curves in detail, here is an example of what software can do.

EXAMPLE

1.23 Density curves of pH and survival times. Figure 1.24 illustrates the use of density curves along with histograms to describe distributions.³⁴ Figure 1.24(a) shows the distribution of the acidity (pH) of rainwater, from Exercise 1.36 (page 27). That exercise illustrates how the choice of classes can change the shape of a histogram. The density curve and the software's default histogram agree that the distribution has a single peak and is approximately symmetric.

Figure 1.24(b) shows a strongly skewed distribution, the survival times of guinea pigs from Table 1.8 (page 29). The histogram and density curve agree on the overall shape and on the “bumps” in the long right tail. The density curve shows a higher peak near the single mode of the distribution. The histogram divides the observations near the mode into two classes, thus reducing the peak.

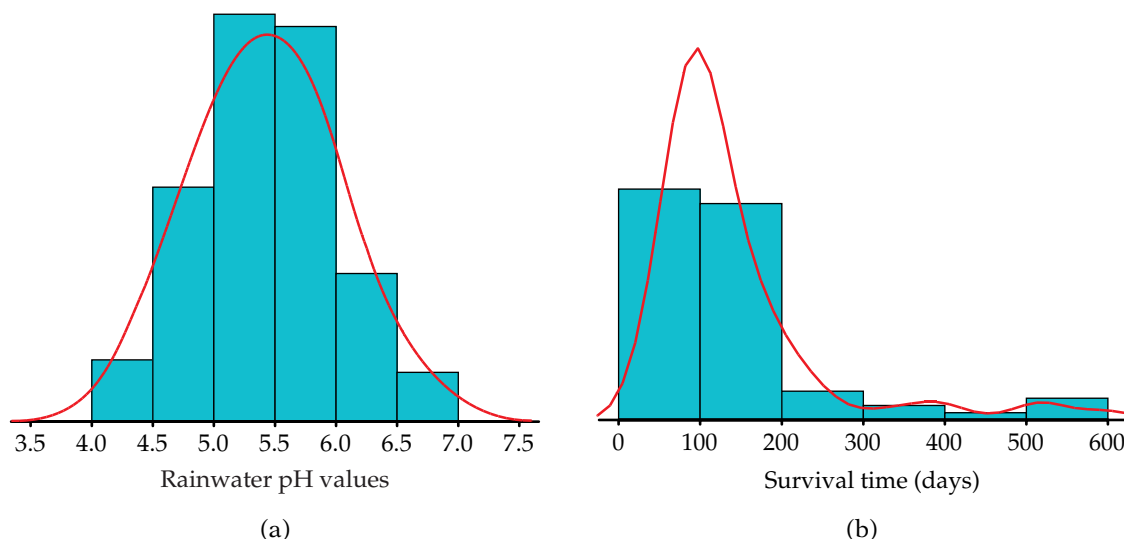


FIGURE 1.24 (a) The distribution of pH values measuring the acidity of 105 samples of rainwater, for Example 1.23. The roughly symmetric distribution is pictured by both a histogram and a density curve. (b) The distribution of the survival times of 72 guinea pigs in a medical experiment, for Example 1.23. The right-skewed distribution is pictured by both a histogram and a density curve.

In general, software that draws density curves describes the data in a way that is less arbitrary than choosing classes for a histogram. A smooth density curve is, however, an idealization that pictures the overall pattern of the data but ignores minor irregularities as well as any outliers. We will concentrate, not on general density curves, but on a special class, the bell-shaped Normal curves.

Density curves

One way to think of a density curve is as a smooth approximation to the irregular bars of a histogram. Figure 1.25 shows a histogram of the scores of all 947 seventh-grade students in Gary, Indiana, on the vocabulary part of the Iowa Test of Basic Skills. Scores of many students on this national test have a very regular distribution. The histogram is symmetric, and both tails fall off quite smoothly from a single center peak. There are no large gaps or obvious outliers. The curve drawn through the tops of the histogram bars in Figure 1.25 is a good description of the overall pattern of the data.

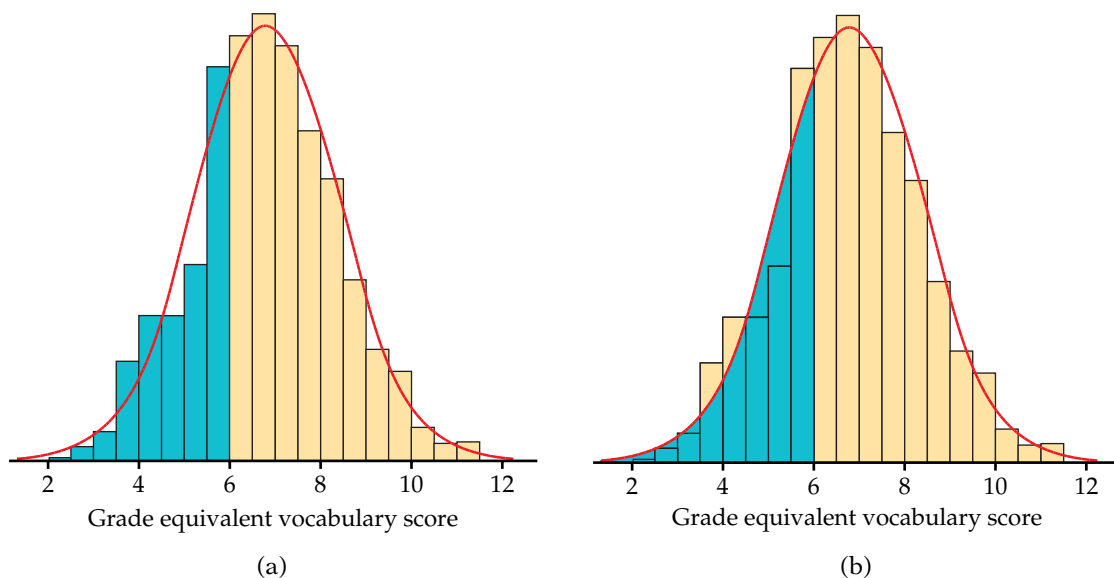


FIGURE 1.25 (a) The distribution of Iowa Test vocabulary scores for Gary, Indiana, seventh-graders. The shaded bars in the histogram represent scores less than or equal to 6.0. The proportion of such scores in the data is 0.303. (b) The shaded area under the Normal density curve also represents scores less than or equal to 6.0. This area is 0.293, close to the true 0.303 for the actual data.

EXAMPLE

1.24 Vocabulary scores. In a histogram, the *areas* of the bars represent either counts or proportions of the observations. In Figure 1.25(a) we have shaded the bars that represent students with vocabulary scores 6.0 or lower. There are 287 such students, who make up the proportion $287/947 = 0.303$ of all Gary seventh-graders. The shaded bars in Figure 1.25(a) make up proportion 0.303 of the total area under all the bars. If we adjust the scale so that the total area of the bars is 1, the area of the shaded bars will be 0.303.

In Figure 1.25(b), we have shaded the *area under the curve* to the left of 6.0. Adjust the scale so that the total area under the curve is exactly 1.

Areas under the curve then represent proportions of the observations. That is, *area = proportion*. The curve is then a density curve. The shaded area under the density curve in Figure 1.25(b) represents the proportion of students with score 6.0 or lower. This area is 0.293, only 0.010 away from the histogram result. You can see that areas under the density curve give quite good approximations of areas given by the histogram.

DENSITY CURVE

A **density curve** is a curve that

- is always on or above the horizontal axis and
- has area exactly 1 underneath it.

A density curve describes the overall pattern of a distribution. The area under the curve and above any range of values is the proportion of all observations that fall in that range.

The density curve in Figure 1.25 is a *Normal curve*. Density curves, like distributions, come in many shapes. Figure 1.26 shows two density curves, a symmetric Normal density curve and a right-skewed curve. A density curve of an appropriate shape is often an adequate description of the overall pattern of a distribution. Outliers, which are deviations from the overall pattern, are not described by the curve.

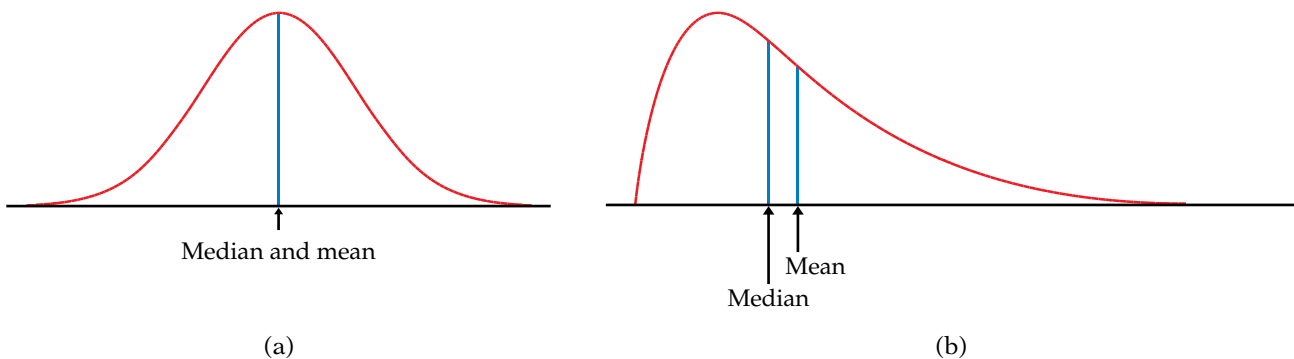


FIGURE 1.26 (a) A symmetric density curve with its mean and median marked. (b) A right-skewed density curve with its mean and median marked.

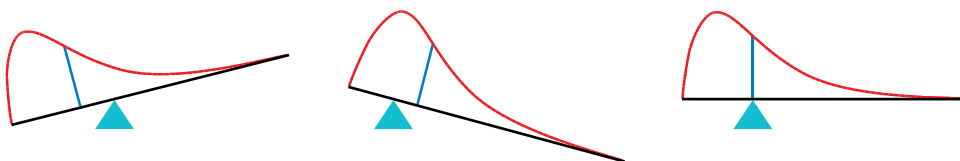
Measuring center and spread for density curves

Our measures of center and spread apply to density curves as well as to actual sets of observations, but only some of these measures are easily seen from the curve. A **mode** of a distribution described by a density curve is a peak point of the curve, the location where the curve is highest. Because areas under a density curve represent proportions of the observations, the **median** is the point with half the total area on each side. You can roughly locate the **quartiles** by

dividing the area under the curve into quarters as accurately as possible by eye. The *IQR* is then the distance between the first and third quartiles. There are mathematical ways of calculating areas under curves. These allow us to locate the median and quartiles exactly on any density curve.

What about the mean and standard deviation? The mean of a set of observations is their arithmetic average. If we think of the observations as weights strung out along a thin rod, the mean is the point at which the rod would balance. This fact is also true of density curves. The mean is the point at which the curve would balance if it were made out of solid material. Figure 1.27 illustrates this interpretation of the mean. We have marked the mean and median on the density curves in Figure 1.26. A symmetric curve, such as the Normal curve in Figure 1.26(a), balances at its center of symmetry. Half the area under a symmetric curve lies on either side of its center, so this is also the median. For a right-skewed curve, such as that shown in Figure 1.26(b), the small area in the long right tail tips the curve more than the same area near the center. The mean (the balance point) therefore lies to the right of the median. It is hard to locate the balance point by eye on a skewed curve. There are mathematical ways of calculating the mean for any density curve, so we are able to mark the mean as well as the median in Figure 1.26(b). The standard deviation can also be calculated mathematically, but it can't be located by eye on most density curves.

FIGURE 1.27 The mean of a density curve is the point at which it would balance.



MEDIAN AND MEAN OF A DENSITY CURVE

The **median** of a density curve is the equal-areas point, the point that divides the area under the curve in half.

The **mean** of a density curve is the balance point, at which the curve would balance if made of solid material.

The median and mean are the same for a symmetric density curve. They both lie at the center of the curve. The mean of a skewed curve is pulled away from the median in the direction of the long tail.

A density curve is an idealized description of a distribution of data. For example, the symmetric density curve in Figure 1.25 is exactly symmetric, but the histogram of vocabulary scores is only approximately symmetric. We therefore need to distinguish between the mean and standard deviation of the density curve and the numbers \bar{x} and s computed from the actual observations. The usual notation for the mean of an idealized distribution is μ (the Greek letter mu). We write the standard deviation of a density curve as σ (the Greek letter sigma).

mean μ
standard deviation σ

Normal distributions

Normal curves

One particularly important class of density curves has already appeared in Figures 1.25 and 1.26(a). These density curves are symmetric, unimodal, and bell-shaped. They are called **Normal curves**, and they describe *Normal distributions*. All Normal distributions have the same overall shape. The exact density curve for a particular Normal distribution is specified by giving its mean μ and its standard deviation σ . The mean is located at the center of the symmetric curve and is the same as the median. Changing μ without changing σ moves the Normal curve along the horizontal axis without changing its spread. The standard deviation σ controls the spread of a Normal curve. Figure 1.28 shows two Normal curves with different values of σ . The curve with the larger standard deviation is more spread out.

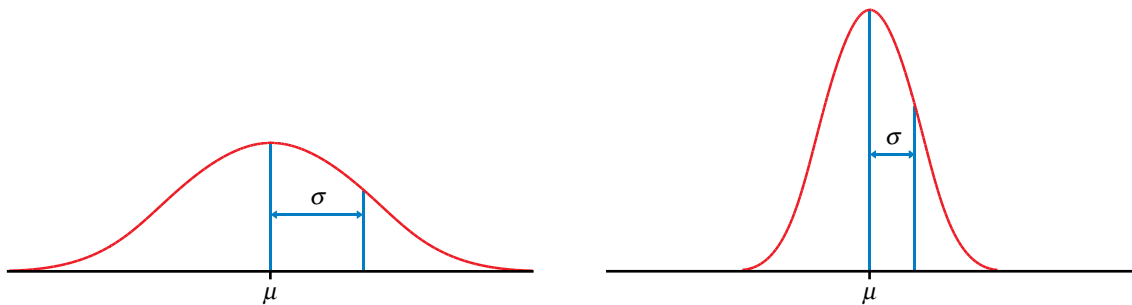
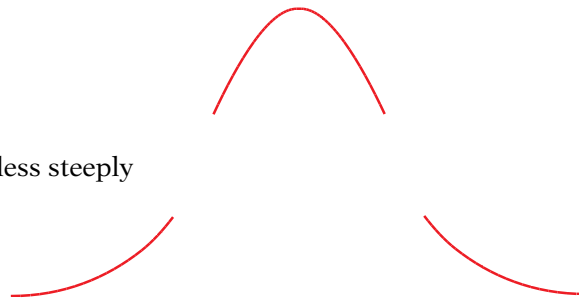


FIGURE 1.28 Two Normal curves, showing the mean μ and standard deviation σ .

The standard deviation σ is the natural measure of spread for Normal distributions. Not only do μ and σ completely determine the shape of a Normal curve, but we can locate σ by eye on the curve. Here's how. As we move out in either direction from the center μ , the curve changes from falling ever more steeply

to falling ever less steeply



The points at which this change of curvature takes place are located at distance σ on either side of the mean μ . You can feel the change as you run your finger along a Normal curve, and so find the standard deviation. Remember that μ and σ alone do not specify the shape of most distributions, and that the shape of density curves in general does not reveal σ . These are special properties of Normal distributions.

There are other symmetric bell-shaped density curves that are not Normal. The Normal density curves are specified by a particular equation. The height

of the density curve at any point x is given by

$$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

We will not make direct use of this fact, although it is the basis of mathematical work with Normal distributions. Notice that the equation of the curve is completely determined by the mean μ and the standard deviation σ .

Why are the Normal distributions important in statistics? Here are three reasons. First, Normal distributions are good descriptions for some distributions of *real data*. Distributions that are often close to Normal include scores on tests taken by many people (such as the Iowa Test of Figure 1.25), repeated careful measurements of the same quantity, and characteristics of biological populations (such as lengths of baby pythons and yields of corn). Second, Normal distributions are good approximations to the results of many kinds of *chance outcomes*, such as tossing a coin many times. Third, and most important, we will see that many *statistical inference* procedures based on Normal distributions work well for other roughly symmetric distributions. **HOWEVER . . . even though many sets of data follow a Normal distribution, many do not.** Most income distributions, for example, are skewed to the right and so are not Normal. Non-Normal data, like non-Normal people, not only are common but are sometimes more interesting than their Normal counterparts.



The 68–95–99.7 rule

Although there are many Normal curves, they all have common properties. Here is one of the most important.

THE 68–95–99.7 RULE

In the Normal distribution with mean μ and standard deviation σ :

- Approximately **68%** of the observations fall within σ of the mean μ .
- Approximately **95%** of the observations fall within 2σ of μ .
- Approximately **99.7%** of the observations fall within 3σ of μ .

Figure 1.29 illustrates the 68–95–99.7 rule. By remembering these three numbers, you can think about Normal distributions without constantly making detailed calculations.

EXAMPLE

1.25 Heights of young women. The distribution of heights of young women aged 18 to 24 is approximately Normal with mean $\mu = 64.5$ inches and standard deviation $\sigma = 2.5$ inches. Figure 1.30 shows what the 68–95–99.7 rule says about this distribution.

Two standard deviations is 5 inches for this distribution. The 95 part of the 68–95–99.7 rule says that the middle 95% of young women are between $64.5 - 5$ and $64.5 + 5$ inches tall, that is, between 59.5 inches and 69.5 inches. This fact is exactly true for an exactly Normal distribution. It is approximately

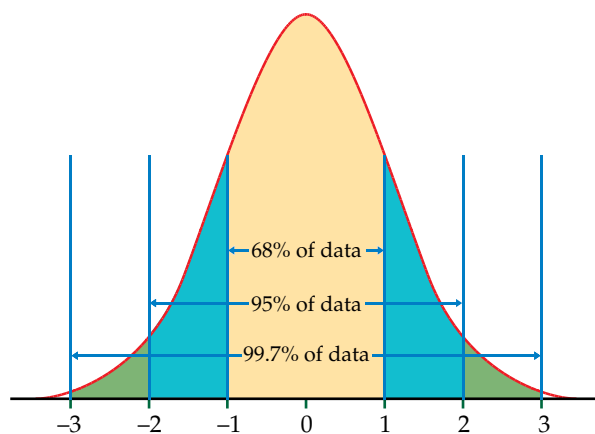


FIGURE 1.29 The 68–95–99.7 rule for Normal distributions.

true for the heights of young women because the distribution of heights is approximately Normal.

The other 5% of young women have heights outside the range from 59.5 to 69.5 inches. Because the Normal distributions are symmetric, half of these women are on the tall side. So the tallest 2.5% of young women are taller than 69.5 inches.

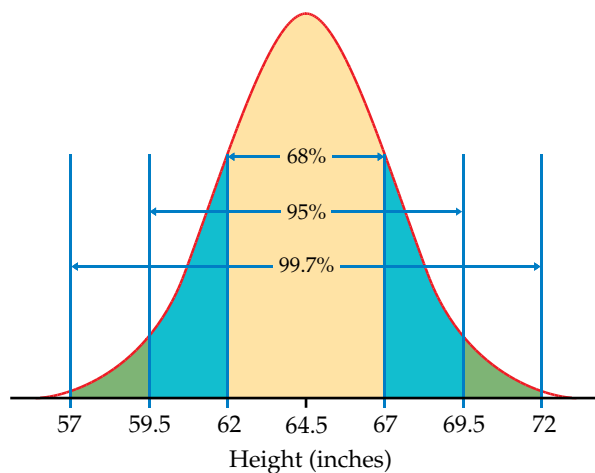


FIGURE 1.30 The 68–95–99.7 rule applied to the heights of young women, for Example 1.25.

Because we will mention Normal distributions often, a short notation is helpful. We abbreviate the Normal distribution with mean μ and standard deviation σ as $N(\mu, \sigma)$. For example, the distribution of young women's heights is $N(64.5, 2.5)$.

USE YOUR KNOWLEDGE

1.99 Test scores. Many states have programs for assessing the skills of students in various grades. The Indiana Statewide Testing for Educational Progress (ISTEP) is one such program.³⁵ In a recent year 76,531 tenth-grade Indiana students took the English/language arts exam. The mean score was 572 and the standard deviation was 51.

Assuming that these scores are approximately Normally distributed, $N(572, 51)$, use the 68–95–99.7 rule to give a range of scores that includes 95% of these students.

1.100 Use the 68–95–99.7 rule. Refer to the previous exercise. Use the 68–95–99.7 rule to give a range of scores that includes 99.7% of these students.

Standardizing observations

As the 68–95–99.7 rule suggests, all Normal distributions share many properties. In fact, all Normal distributions are the same if we measure in units of size σ about the mean μ as center. Changing to these units is called *standardizing*. To standardize a value, subtract the mean of the distribution and then divide by the standard deviation.

STANDARDIZING AND z-SCORES

If x is an observation from a distribution that has mean μ and standard deviation σ , the **standardized value** of x is

$$z = \frac{x - \mu}{\sigma}$$

A standardized value is often called a **z-score**.

A z-score tells us how many standard deviations the original observation falls away from the mean, and in which direction. Observations larger than the mean are positive when standardized, and observations smaller than the mean are negative.

EXAMPLE

1.26 Find some z-scores. The heights of young women are approximately Normal with $\mu = 64.5$ inches and $\sigma = 2.5$ inches. The z-score for height is

$$z = \frac{\text{height} - 64.5}{2.5}$$

A woman's standardized height is the number of standard deviations by which her height differs from the mean height of all young women. A woman 68 inches tall, for example, has z-score

$$z = \frac{68 - 64.5}{2.5} = 1.4$$

or 1.4 standard deviations above the mean. Similarly, a woman 5 feet (60 inches) tall has z-score

$$z = \frac{60 - 64.5}{2.5} = -1.8$$

or 1.8 standard deviations less than the mean height.

USE YOUR KNOWLEDGE

- 1.101 Find the z -score.** Consider the ISTEP scores (see Exercise 1.99), which we can assume are approximately Normal, $N(572, 51)$. Give the z -score for a student who received a score of 600.
- 1.102 Find another z -score.** Consider the ISTEP scores, which we can assume are approximately Normal, $N(572, 51)$. Give the z -score for a student who received a score of 500. Explain why your answer is negative even though all of the test scores are positive.

We need a way to write variables, such as “height” in Example 1.25, that follow a theoretical distribution such as a Normal distribution. We use capital letters near the end of the alphabet for such variables. If X is the height of a young woman, we can then shorten “the height of a young woman is less than 68 inches” to “ $X < 68$.” We will use lowercase x to stand for any specific value of the variable X .

We often standardize observations from symmetric distributions to express them in a common scale. We might, for example, compare the heights of two children of different ages by calculating their z -scores. The standardized heights tell us where each child stands in the distribution for his or her age group.

Standardizing is a linear transformation that transforms the data into the standard scale of z -scores. We know that a linear transformation does not change the shape of a distribution, and that the mean and standard deviation change in a simple manner. In particular, *the standardized values for any distribution always have mean 0 and standard deviation 1.*

If the variable we standardize has a Normal distribution, standardizing does more than give a common scale. It makes all Normal distributions into a single distribution, and this distribution is still Normal. Standardizing a variable that has any Normal distribution produces a new variable that has the *standard Normal distribution*.

THE STANDARD NORMAL DISTRIBUTION

The **standard Normal distribution** is the Normal distribution $N(0, 1)$ with mean 0 and standard deviation 1.

If a variable X has any Normal distribution $N(\mu, \sigma)$ with mean μ and standard deviation σ , then the standardized variable

$$Z = \frac{X - \mu}{\sigma}$$

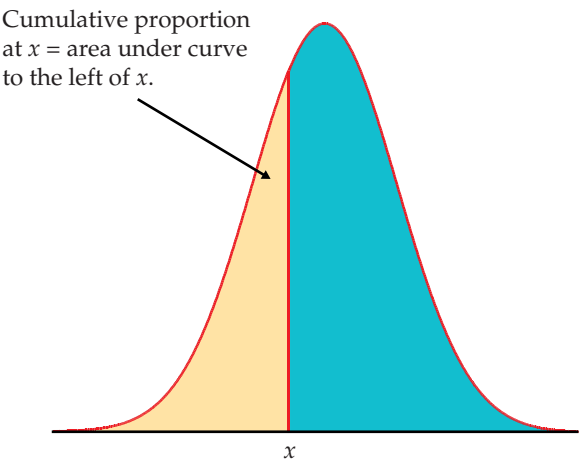
has the standard Normal distribution.

Normal distribution calculations

Areas under a Normal curve represent proportions of observations from that Normal distribution. There is no formula for areas under a Normal curve. Calculations use either software that calculates areas or a table of areas. The table

cumulative proportion and most software calculate one kind of area: **cumulative proportions**. A cumulative proportion is the proportion of observations in a distribution that lie at or below a given value. When the distribution is given by a density curve, the cumulative proportion is the area under the curve to the left of a given value. Figure 1.31 shows the idea more clearly than words do.

FIGURE 1.31 The *cumulative proportion* for a value x is the proportion of all observations from the distribution that are less than or equal to x . This is the area to the left of x under the Normal curve.

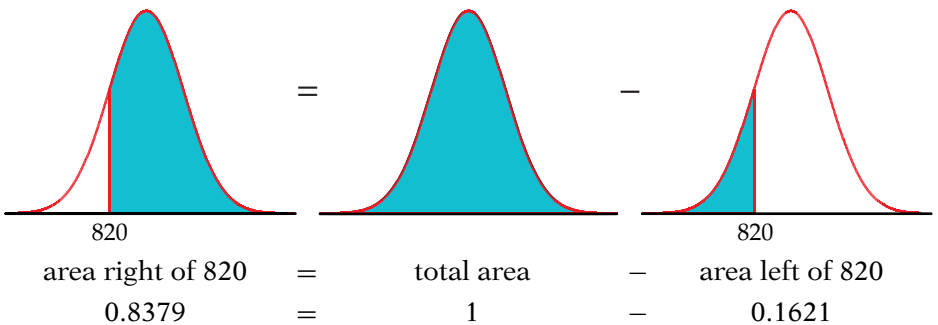


The key to calculating Normal proportions is to match the area you want with areas that represent cumulative proportions. Then get areas for cumulative proportions either from software or (with an extra step) from a table. The following examples show the method in pictures.

EXAMPLE

1.27 The NCAA standard for SAT scores. The National Collegiate Athletic Association (NCAA) requires Division I athletes to get a combined score of at least 820 on the SAT Mathematics and Verbal tests to compete in their first college year. (Higher scores are required for students with poor high school grades.) The scores of the 1.4 million students in the class of 2003 who took the SATs were approximately Normal with mean 1026 and standard deviation 209. What proportion of all students had SAT scores of at least 820?

Here is the calculation in pictures: the proportion of scores above 820 is the area under the curve to the right of 820. That's the total area under the curve (which is always 1) minus the cumulative proportion up to 820.

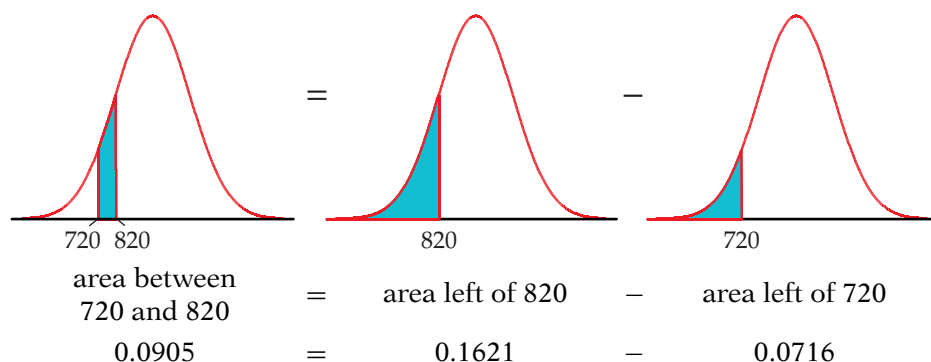


That is, the proportion of all SAT takers who would be NCAA qualifiers is 0.8379, or about 84%.

There is *no* area under a smooth curve and exactly over the point 820. Consequently, the area to the right of 820 (the proportion of scores > 820) is the same as the area at or to the right of this point (the proportion of scores ≥ 820). The actual data may contain a student who scored exactly 820 on the SAT. That the proportion of scores exactly equal to 820 is 0 for a Normal distribution is a consequence of the idealized smoothing of Normal distributions for data.

EXAMPLE

1.28 NCAA partial qualifiers. The NCAA considers a student a “partial qualifier” eligible to practice and receive an athletic scholarship, but not to compete, if the combined SAT score is at least 720. What proportion of all students who take the SAT would be partial qualifiers? That is, what proportion have scores between 720 and 820? Here are the pictures:



About 9% of all students who take the SAT have scores between 720 and 820.

How do we find the numerical values of the areas in Examples 1.27 and 1.28? If you use software, just plug in mean 1026 and standard deviation 209. Then ask for the cumulative proportions for 820 and for 720. (Your software will probably refer to these as “cumulative probabilities.” We will learn in Chapter 4 why the language of probability fits.) If you make a sketch of the area you want, you will never go wrong.



You can use the *Normal Curve* applet on the text CD and Web site to find Normal proportions. The applet is more flexible than most software—it will find any Normal proportion, not just cumulative proportions. The applet is an excellent way to understand Normal curves. But, because of the limitations of Web browsers, the applet is not as accurate as statistical software.

If you are not using software, you can find cumulative proportions for Normal curves from a table. That requires an extra step, as we now explain.

Using the standard Normal table

The extra step in finding cumulative proportions from a table is that we must first standardize to express the problem in the standard scale of z -scores. This allows us to get by with just one table, a table of *standard Normal cumulative*

proportions. Table A in the back of the book gives cumulative proportions for the standard Normal distribution. Table A also appears on the inside front cover. The pictures at the top of the table remind us that the entries are cumulative proportions, areas under the curve to the left of a value z .

EXAMPLE

1.29 Find the proportion from z . What proportion of observations on a standard Normal variable Z take values less than 1.47?

Solution: To find the area to the left of 1.47, locate 1.4 in the left-hand column of Table A, then locate the remaining digit 7 as .07 in the top row. The entry opposite 1.4 and under .07 is 0.9292. This is the cumulative proportion we seek. Figure 1.32 illustrates this area.

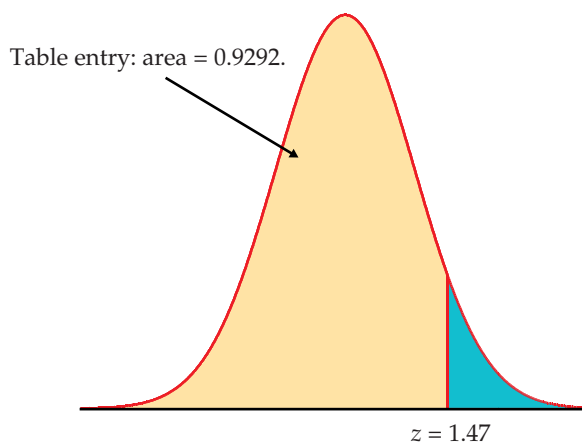


FIGURE 1.32 The area under a standard Normal curve to the left of the point $z = 1.47$ is 0.9292, for Example 1.29. Table A gives areas under the standard Normal curve.

Now that you see how Table A works, let's redo the NCAA Examples 1.27 and 1.28 using the table.

EXAMPLE

1.30 Find the proportion from x . What proportion of all students who take the SAT have scores of at least 820? The picture that leads to the answer is exactly the same as in Example 1.27. The extra step is that we first standardize in order to read cumulative proportions from Table A. If X is SAT score, we want the proportion of students for which $X \geq 820$.

1. *Standardize.* Subtract the mean, then divide by the standard deviation, to transform the problem about X into a problem about a standard Normal Z :

$$\begin{aligned} X &\geq 820 \\ \frac{X - 1026}{209} &\geq \frac{820 - 1026}{209} \\ Z &\geq -0.99 \end{aligned}$$

2. *Use the table.* Look at the pictures in Example 1.27. From Table A, we see that the proportion of observations less than -0.99 is 0.1611. The area to the right of -0.99 is therefore $1 - 0.1611 = 0.8389$. This is about 84%.

The area from the table in Example 1.30 (0.8389) is slightly less accurate than the area from software in Example 1.27 (0.8379) because we must round z to two places when we use Table A. The difference is rarely important in practice.

EXAMPLE

1.31 Proportion of partial qualifiers. What proportion of all students who take the SAT would be partial qualifiers in the eyes of the NCAA? That is, what proportion of students have SAT scores between 720 and 820? First, sketch the areas, exactly as in Example 1.28. We again use X as shorthand for an SAT score.

1. *Standardize.*

$$\begin{aligned} 720 &\leq X < 820 \\ \frac{720 - 1026}{209} &\leq \frac{X - 1026}{209} < \frac{820 - 1026}{209} \\ -1.46 &\leq Z < -0.99 \end{aligned}$$

2. *Use the table.*

$$\begin{aligned} \text{area between } -1.46 \text{ and } -0.99 &= (\text{area left of } -0.99) - (\text{area left of } -1.46) \\ &= 0.1611 - 0.0721 = 0.0890 \end{aligned}$$

As in Example 1.28, about 9% of students would be partial qualifiers.

Sometimes we encounter a value of z more extreme than those appearing in Table A. For example, the area to the left of $z = -4$ is not given directly in the table. The z -values in Table A leave only area 0.0002 in each tail unaccounted for. For practical purposes, we can act as if there is zero area outside the range of Table A.

USE YOUR KNOWLEDGE

1.103 Find the proportion. Consider the ISTEP scores, which are approximately Normal, $N(572, 51)$. Find the proportion of students who have scores less than 600. Find the proportion of students who have scores greater than or equal to 600. Sketch the relationship between these two calculations using pictures of Normal curves similar to the ones given in Example 1.27.

1.104 Find another proportion. Consider the ISTEP scores, which are approximately Normal, $N(572, 51)$. Find the proportion of students who have scores between 600 and 650. Use pictures of Normal curves similar to the ones given in Example 1.28 to illustrate your calculations.

Inverse Normal calculations

Examples 1.25 to 1.29 illustrate the use of Normal distributions to find the proportion of observations in a given event, such as “SAT score between 720 and

820.” We may instead want to find the observed value corresponding to a given proportion.

Statistical software will do this directly. Without software, use Table A backward, finding the desired proportion in the body of the table and then reading the corresponding z from the left column and top row.

EXAMPLE

1.32 How high for the top 10%? Scores on the SAT Verbal test in recent years follow approximately the $N(505, 110)$ distribution. How high must a student score in order to place in the top 10% of all students taking the SAT?

Again, the key to the problem is to draw a picture. Figure 1.33 shows that we want the score x with area above it 0.10. That’s the same as area below x equal to 0.90.

Statistical software has a function that will give you the x for any cumulative proportion you specify. The function often has a name such as “inverse cumulative probability.” Plug in mean 505, standard deviation 110, and cumulative proportion 0.9. The software tells you that $x = 645.97$. We see that a student must score at least 646 to place in the highest 10%.

Without software, first find the standard score z with cumulative proportion 0.9, then “unstandardize” to find x . Here is the two-step process:

1. *Use the table.* Look in the body of Table A for the entry closest to 0.9. It is 0.8997. This is the entry corresponding to $z = 1.28$. So $z = 1.28$ is the standardized value with area 0.9 to its left.
2. *Unstandardize* to transform the solution from z back to the original x scale. We know that the standardized value of the unknown x is $z = 1.28$. So x itself satisfies

$$\frac{x - 505}{110} = 1.28$$

Solving this equation for x gives

$$x = 505 + (1.28)(110) = 645.8$$

This equation should make sense: it finds the x that lies 1.28 standard deviations above the mean on this particular Normal curve. That is the

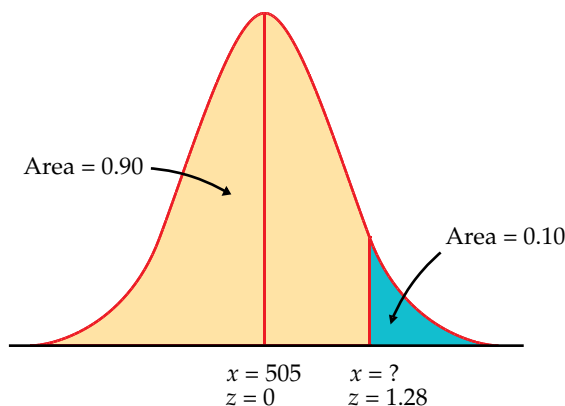


FIGURE 1.33 Locating the point on a Normal curve with area 0.10 to its right, for Example 1.32. The result is $x = 646$, or $z = 1.28$ in the standard scale.

“unstandardized” meaning of $z = 1.28$. The general rule for unstandardizing a z -score is

$$x = \mu + z\sigma$$

USE YOUR KNOWLEDGE

1.105 What score is needed to be in the top 5%? Consider the ISTEP scores, which are approximately Normal, $N(572, 51)$. How high a score is needed to be in the top 5% of students who take this exam?

1.106 Find the score that 60% of students will exceed. Consider the ISTEP scores, which are approximately Normal, $N(572, 51)$. Sixty percent of the students will score above x on this exam. Find x .

Normal quantile plots

The Normal distributions provide good descriptions of some distributions of real data, such as the Gary vocabulary scores. The distributions of some other common variables are usually skewed and therefore distinctly non-Normal. Examples include economic variables such as personal income and gross sales of business firms, the survival times of cancer patients after treatment, and the service lifetime of mechanical or electronic components. While experience can suggest whether or not a Normal distribution is plausible in a particular case, it is risky to assume that a distribution is Normal without actually inspecting the data.

A histogram or stemplot can reveal distinctly non-Normal features of a distribution, such as outliers (the breaking strengths in Figure 1.9, page 17), pronounced skewness (the survival times in Figure 1.24(b), page 54), or gaps and clusters (the Massachusetts college tuitions in Figure 1.16, page 25). If the stemplot or histogram appears roughly symmetric and unimodal, however, we need a more sensitive way to judge the adequacy of a Normal model. The most useful tool for assessing Normality is another graph, the **Normal quantile plot**.

Here is the basic idea of a Normal quantile plot. The graphs produced by software use more sophisticated versions of this idea. It is not practical to make Normal quantile plots by hand.

1. Arrange the observed data values from smallest to largest. Record what percentile of the data each value occupies. For example, the smallest observation in a set of 20 is at the 5% point, the second smallest is at the 10% point, and so on.
2. Do Normal distribution calculations to find the values of z corresponding to these same percentiles. For example, $z = -1.645$ is the 5% point of the standard Normal distribution, and $z = -1.282$ is the 10% point. We call these values of Z **Normal scores**.

3. Plot each data point x against the corresponding Normal score. If the data distribution is close to any Normal distribution, the plotted points will lie close to a straight line.

Normal quantile plot

Normal scores

Any Normal distribution produces a straight line on the plot because standardizing turns any Normal distribution into a standard Normal distribution. Standardizing is a linear transformation that can change the slope and intercept of the line in our plot but cannot turn a line into a curved pattern.

USE OF NORMAL QUANTILE PLOTS

If the points on a Normal quantile plot lie close to a straight line, the plot indicates that the data are Normal. Systematic deviations from a straight line indicate a non-Normal distribution. Outliers appear as points that are far away from the overall pattern of the plot.

Figures 1.34 to 1.36 are Normal quantile plots for data we have met earlier. The data x are plotted vertically against the corresponding standard Normal z -score plotted horizontally. The z -score scale extends from -3 to 3 because almost all of a standard Normal curve lies between these values. These figures show how Normal quantile plots behave.

EXAMPLE

granularity

1.33 Breaking strengths are Normal. Figure 1.34 is a Normal quantile plot of the breaking strengths in Example 1.11 (page 17). Lay a transparent straightedge over the center of the plot to see that most of the points lie close to a straight line. A Normal distribution describes these points quite well. The only substantial deviations are short horizontal runs of points. Each run represents repeated observations having the same value—there are five measurements at 1150, for example. This phenomenon is called **granularity**. It is caused by the limited precision of the measurements and does not represent an important deviation from Normality.

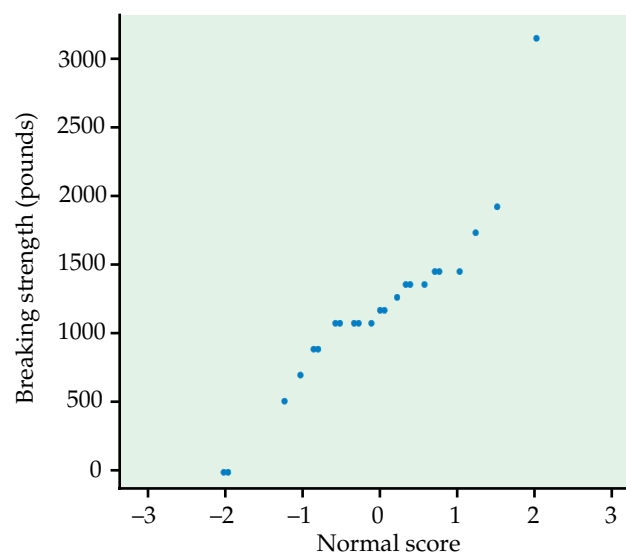


FIGURE 1.34 Normal quantile plot of the breaking strengths of wires bonded to a semiconductor wafer, for Example 1.33. This distribution has a Normal shape except for outliers in both tails.

The high outlier at 3150 pounds lies above the line formed by the center of the data—it is farther out in the high direction than we expect Normal data to be.

The two low outliers at 0 lie below the line—they are suspiciously far out in the low direction. Compare Figure 1.34 with the histogram of these data in Figure 1.9 (page 17).

EXAMPLE

1.34 Survival times are not Normal. Figure 1.35 is a Normal quantile plot of the guinea pig survival times from Table 1.8 (page 29). Figure 1.24(b) (page 54) shows that this distribution is strongly skewed to the right.

To see the right-skewness in the Normal quantile plot, draw a line through the leftmost points, which correspond to the smaller observations. The larger observations fall systematically above this line. That is, the right-of-center observations have larger values than in a Normal distribution. *In a right-skewed distribution, the largest observations fall distinctly above a line drawn through the main body of points.* Similarly, left-skewness is evident when the smallest observations fall below the line. Unlike Figure 1.34, there are no individual outliers.

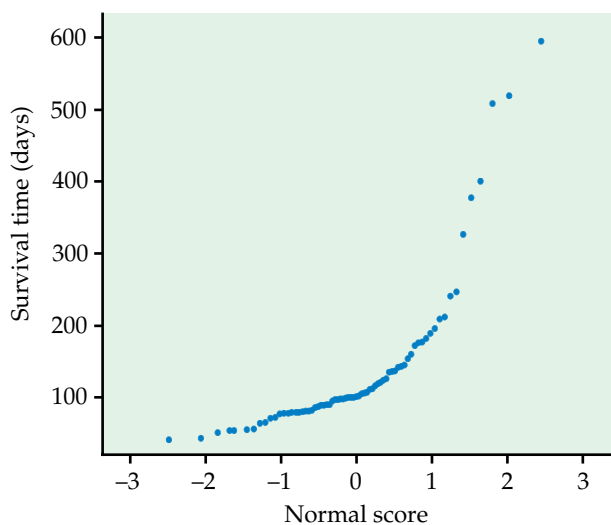


FIGURE 1.35 Normal quantile plot of the survival times of guinea pigs in a medical experiment, for Example 1.34. This distribution is skewed to the right.

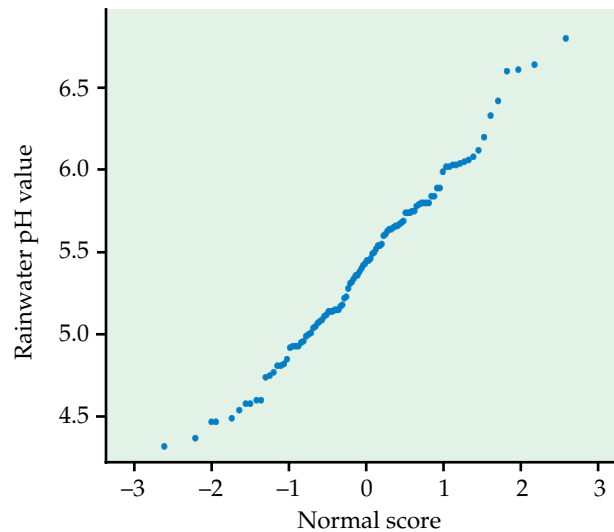
EXAMPLE

1.35 Acidity of rainwater is approximately Normal. Figure 1.36 is a Normal quantile plot of the 105 acidity (pH) measurements of rainwater from Exercise 1.36 (page 27). Histograms don't settle the question of approximate Normality of these data, because their shape depends on the choice of classes. The Normal quantile plot makes it clear that a Normal distribution is a good description—there are only minor wiggles in a generally straight-line pattern.



As Figure 1.36 illustrates, real data almost always show some departure from the theoretical Normal model. *When you examine a Normal quantile plot, look for shapes that show clear departures from Normality. Don't overreact to*

FIGURE 1.36 Normal quantile plot of the acidity (pH) values of 105 samples of rainwater, for Example 1.35. This distribution is approximately Normal.



minor wiggles in the plot. When we discuss statistical methods that are based on the Normal model, we will pay attention to the sensitivity of each method to departures from Normality. Many common methods work well as long as the data are approximately Normal and outliers are not present.

BEYOND THE BASICS

Density Estimation

density estimator

A density curve gives a compact summary of the overall shape of a distribution. Many distributions do not have the Normal shape. There are other families of density curves that are used as mathematical models for various distribution shapes. Modern software offers a more flexible option, illustrated by the two graphs in Figure 1.24 (page 54). A **density estimator** does not start with any specific shape, such as the Normal shape. It looks at the data and draws a density curve that describes the overall shape of the data. Density estimators join stemplots and histograms as useful graphical tools for exploratory data analysis.

SECTION 1.3 Summary

The overall pattern of a distribution can often be described compactly by a **density curve**. A density curve has total area 1 underneath it. Areas under a density curve give proportions of observations for the distribution.

The **mean** μ (balance point), the **median** (equal-areas point), and the **quartiles** can be approximately located by eye on a density curve. The **standard deviation** σ cannot be located by eye on most density curves. The mean and median are equal for symmetric density curves, but the mean of a skewed curve is located farther toward the long tail than is the median.

The **Normal distributions** are described by bell-shaped, symmetric, unimodal density curves. The mean μ and standard deviation σ completely specify the Normal distribution $N(\mu, \sigma)$. The mean is the center of symmetry, and σ is the distance from μ to the change-of-curvature points on either side.

To **standardize** any observation x , subtract the mean of the distribution and then divide by the standard deviation. The resulting **z-score** $z = (x - \mu)/\sigma$ says how many standard deviations x lies from the distribution mean. All Normal distributions are the same when measurements are transformed to the standardized scale. In particular, all Normal distributions satisfy the **68–95–99.7 rule**.

If X has the $N(\mu, \sigma)$ distribution, then the standardized variable $Z = (X - \mu)/\sigma$ has the **standard Normal distribution** $N(0, 1)$. Proportions for any Normal distribution can be calculated by software or from the **standard Normal table** (Table A), which gives the **cumulative proportions** of $Z < z$ for many values of z .

The adequacy of a Normal model for describing a distribution of data is best assessed by a **Normal quantile plot**, which is available in most statistical software packages. A pattern on such a plot that deviates substantially from a straight line indicates that the data are not Normal.

SECTION 1.3 Exercises

For Exercises 1.99 and 1.100, see pages 60 and 61; for Exercises 1.101 and 1.102, see page 62; for Exercises 1.103 and 1.104, see page 66; and for Exercises 1.105 and 1.106, see page 68.

1.107 Sketch some density curves. Sketch density curves that might describe distributions with the following shapes:

- (a) Symmetric, but with two peaks (that is, two strong clusters of observations).
- (b) Single peak and skewed to the right.

1.108 A uniform distribution. If you ask a computer to generate “random numbers” between 0 and 1, you will get observations from a **uniform distribution**. Figure 1.37 graphs the density curve for a uniform distribution. Use areas under this density curve to answer the following questions.

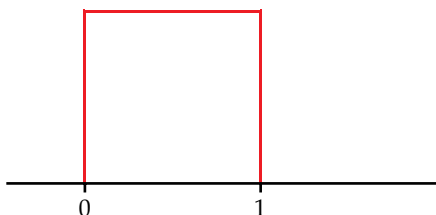


FIGURE 1.37 The density curve of a uniform distribution, for Exercise 1.108.

- (a) Why is the total area under this curve equal to 1?
- (b) What proportion of the observations lie below 0.35?
- (c) What proportion of the observations lie between 0.35 and 0.65?

1.109 Use a different range for the uniform distribution. Many random number generators allow users to specify the range of the random numbers to be produced. Suppose that you specify that the outcomes are to be distributed uniformly between 0 and 4. Then the density curve of the outcomes has constant height between 0 and 4, and height 0 elsewhere.

- (a) What is the height of the density curve between 0 and 4? Draw a graph of the density curve.
- (b) Use your graph from (a) and the fact that areas under the curve are proportions of outcomes to find the proportion of outcomes that are less than 1.
- (c) Find the proportion of outcomes that lie between 0.5 and 2.5.

1.110 Find the mean, the median, and the quartiles. What are the mean and the median of the uniform distribution in Figure 1.37? What are the quartiles?

1.111 Three density curves. Figure 1.38 displays three density curves, each with three points marked on

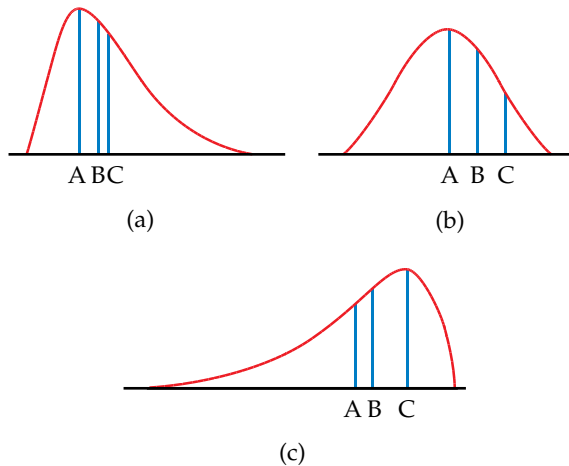



FIGURE 1.38 Three density curves, for Exercise 1.111.

it. At which of these points on each curve do the mean and the median fall?

- 1.112 Length of pregnancies.** The length of human pregnancies from conception to birth varies according to a distribution that is approximately Normal with mean 266 days and standard deviation 16 days. Draw a density curve for this distribution on which the mean and standard deviation are correctly located.

- 1.113**  **Use the Normal Curve applet.** The 68–95–99.7 rule for Normal distributions is a useful approximation. You can use the *Normal Curve* applet on the text CD and Web site to see how accurate the rule is. Drag one flag across the other so that the applet shows the area under the curve between the two flags.

(a) Place the flags one standard deviation on either side of the mean. What is the area between these two values? What does the 68–95–99.7 rule say this area is?

(b) Repeat for locations two and three standard deviations on either side of the mean. Again compare the 68–95–99.7 rule with the area given by the applet.

- 1.114 Pregnancies and the 68–95–99.7 rule.** The length of human pregnancies from conception to birth varies according to a distribution that is approximately Normal with mean 266 days and standard deviation 16 days. Use the 68–95–99.7 rule to answer the following questions.

(a) Between what values do the lengths of the middle 95% of all pregnancies fall?

(b) How short are the shortest 2.5% of all pregnancies? How long do the longest 2.5% last?

- 1.115 Horse pregnancies are longer.** Bigger animals tend to carry their young longer before birth. The length of horse pregnancies from conception to birth varies according to a roughly Normal distribution with mean 336 days and standard deviation 3 days. Use the 68–95–99.7 rule to answer the following questions.

(a) Almost all (99.7%) horse pregnancies fall in what range of lengths?


(b) What percent of horse pregnancies are longer than 339 days?

- 1.116 Binge drinking survey.** One reason that Normal distributions are important is that they describe how the results of an opinion poll would vary if the poll were repeated many times. About 20% of college students say they are frequent binge drinkers. Think about taking many randomly chosen samples of 1600 students. The proportions of college students in these samples who say they are frequent binge drinkers will follow the Normal distribution with mean 0.20 and standard deviation 0.01. Use this fact and the 68–95–99.7 rule to answer these questions.

(a) In many samples, what percent of samples give results above 0.2? Above 0.22?

(b) In a large number of samples, what range contains the central 95% of proportions of students who say they are frequent binge drinkers?

- 1.117 Heights of women.** The heights of women aged 20 to 29 are approximately Normal with mean 64 inches and standard deviation 2.7 inches. Men the same age have mean height 69.3 inches with standard deviation 2.8 inches. What are the z -scores for a woman 6 feet tall and a man 6 feet tall? What information do the z -scores give that the actual heights do not?

- 1.118**  **Use the Normal Curve applet.** Use the *Normal Curve* applet for the standard Normal distribution to say how many standard deviations above and below the mean the quartiles of any Normal distribution lie.

- 1.119 Acidity of rainwater.** The Normal quantile plot in Figure 1.36 (page 71) shows that the acidity (pH) measurements for rainwater samples in Exercise 1.36 are approximately Normal. How well do these scores satisfy the 68–95–99.7 rule?

To find out, calculate the mean \bar{x} and standard deviation s of the observations. Then calculate the percent of the 105 measurements that fall between $\bar{x} - s$ and $\bar{x} + s$ and compare your result with 68%. Do the same for the intervals covering two and three standard deviations on either side of the mean. (The 68–95–99.7 rule is exact for any theoretical Normal distribution. It will hold only approximately for actual data.)

1.120 Find some proportions. Using either Table A or your calculator or software, find the proportion of observations from a standard Normal distribution that satisfies each of the following statements. In each case, sketch a standard Normal curve and shade the area under the curve that is the answer to the question.

- (a) $Z < 1.65$
- (b) $Z > 1.65$
- (c) $Z > -0.76$
- (d) $-0.76 < Z < 1.65$

1.121 Find more proportions. Using either Table A or your calculator or software, find the proportion of observations from a standard Normal distribution for each of the following events. In each case, sketch a standard Normal curve and shade the area representing the proportion.

- (a) $Z \leq -1.9$
- (b) $Z \geq -1.9$
- (c) $Z > 1.55$
- (d) $-1.9 < Z < 1.55$

1.122 Find some values of z . Find the value z of a standard Normal variable Z that satisfies each of the following conditions. (If you use Table A, report the value of z that comes closest to satisfying the condition.) In each case, sketch a standard Normal curve with your value of z marked on the axis.

- (a) 25% of the observations fall below z .
- (b) 35% of the observations fall above z .

1.123 Find more values of z . The variable Z has a standard Normal distribution.

- (a) Find the number z that has cumulative proportion 0.85.
- (b) Find the number z such that the event $Z > z$ has proportion 0.40.

1.124 Find some values of z . The Wechsler Adult Intelligence Scale (WAIS) is the most common “IQ test.” The scale of scores is set separately for each age group and is approximately Normal with mean 100 and standard deviation 15. People with WAIS scores below 70 are considered mentally retarded when, for example, applying for Social Security disability benefits. What percent of adults are retarded by this criterion?

1.125 High IQ scores. The Wechsler Adult Intelligence Scale (WAIS) is the most common “IQ test.” The scale of scores is set separately for each age group and is approximately Normal with mean 100 and standard deviation 15. The organization MENSA, which calls itself “the high IQ society,” requires a WAIS score of 130 or higher for membership. What percent of adults would qualify for membership?

There are two major tests of readiness for college, the ACT and the SAT. ACT scores are reported on a scale from 1 to 3. The distribution of ACT scores for more than 1 million students in a recent high school graduating class was roughly Normal with mean $\mu = 20.8$ and standard deviation $\sigma = 4.8$. SAT scores are reported on a scale from 400 to 1600. The SAT scores for 1.4 million students in the same graduating class were roughly Normal with mean $\mu = 1026$ and standard deviation $\sigma = 209$. Exercises 1.126 to 1.135 are based on this information.

1.126 Compare an SAT score with an ACT score.

Tonya scores 1320 on the SAT. Jermaine scores 28 on the ACT. Assuming that both tests measure the same thing, who has the higher score? Report the z -scores for both students.

1.127 Make another comparison. Jacob scores 17 on the ACT. Emily scores 680 on the SAT. Assuming that both tests measure the same thing, who has the higher score? Report the z -scores for both students.

1.128 Find the ACT equivalent. Jose scores 1380 on the SAT. Assuming that both tests measure the same thing, what score on the ACT is equivalent to Jose’s SAT score?

1.129 Find the SAT equivalent. Maria scores 29 on the ACT. Assuming that both tests measure the same thing, what score on the SAT is equivalent to Maria’s ACT score?

1.130 Find the SAT percentile. Reports on a student’s ACT or SAT usually give the percentile as well as the actual score. The percentile is just the cumulative proportion stated as a percent: the percent of all

scores that were lower than this one. Tonya scores 1320 on the SAT. What is her percentile?

- 1.131 Find the ACT percentile.** Reports on a student's ACT or SAT usually give the percentile as well as the actual score. The percentile is just the cumulative proportion stated as a percent: the percent of all scores that were lower than this one. Jacob scores 17 on the ACT. What is his percentile?
- 1.132 How high is the top 10%?** What SAT scores make up the top 10% of all scores?
- 1.133 How low is the bottom 20%?** What SAT scores make up the bottom 20% of all scores?
- 1.134 Find the ACT quartiles.** The quartiles of any distribution are the values with cumulative proportions 0.25 and 0.75. What are the quartiles of the distribution of ACT scores?
- 1.135 Find the SAT quintiles.** The quintiles of any distribution are the values with cumulative proportions 0.20, 0.40, 0.60, and 0.80. What are the quintiles of the distribution of SAT scores?
- 1.136 Proportion of women with high cholesterol.** Too much cholesterol in the blood increases the risk of heart disease. Young women are generally less afflicted with high cholesterol than other groups. The cholesterol levels for women aged 20 to 34 follow an approximately Normal distribution with mean 185 milligrams per deciliter (mg/dl) and standard deviation 39 mg/dl.³⁶
- (a) Cholesterol levels above 240 mg/dl demand medical attention. What percent of young women have levels above 240 mg/dl?
- (b) Levels above 200 mg/dl are considered borderline high. What percent of young women have blood cholesterol between 200 and 240 mg/dl?
- 1.137 Proportion of men with high cholesterol.** Middle-aged men are more susceptible to high cholesterol than the young women of the previous exercise. The blood cholesterol levels of men aged 55 to 64 are approximately Normal with mean 222 mg/dl and standard deviation 37 mg/dl. What percent of these men have high cholesterol (levels above 240 mg/dl)? What percent have borderline high cholesterol (between 200 and 240 mg/dl)?
- 1.138 Diagnosing osteoporosis.** Osteoporosis is a condition in which the bones become brittle due to loss of minerals. To diagnose osteoporosis, an elaborate apparatus measures bone mineral density

(BMD). BMD is usually reported in standardized form. The standardization is based on a population of healthy young adults. The World Health Organization (WHO) criterion for osteoporosis is a BMD 2.5 standard deviations below the mean for young adults. BMD measurements in a population of people similar in age and sex roughly follow a Normal distribution.

- (a) What percent of healthy young adults have osteoporosis by the WHO criterion?
- (b) Women aged 70 to 79 are of course not young adults. The mean BMD in this age is about -2 on the standard scale for young adults. Suppose that the standard deviation is the same as for young adults. What percent of this older population has osteoporosis?

- 1.139 Length of pregnancies.** The length of human pregnancies from conception to birth varies according to a distribution that is approximately Normal with mean 266 days and standard deviation 16 days.

- (a) What percent of pregnancies last less than 240 days (that's about 8 months)?
- (b) What percent of pregnancies last between 240 and 270 days (roughly between 8 months and 9 months)?
- (c) How long do the longest 20% of pregnancies last?

- 1.140 CHALLENGE Quartiles for Normal distributions.** The quartiles of any distribution are the values with cumulative proportions 0.25 and 0.75.

- (a) What are the quartiles of the standard Normal distribution?
- (b) Using your numerical values from (a), write an equation that gives the quartiles of the $N(\mu, \sigma)$ distribution in terms of μ and σ .

- (c) The length of human pregnancies from conception to birth varies according to a distribution that is approximately Normal with mean 266 days and standard deviation 16 days. Apply your result from (b): what are the quartiles of the distribution of lengths of human pregnancies?

- 1.141 CHALLENGE IQR for Normal distributions.** Continue your work from the previous exercise. The interquartile range *IQR* is the distance between the first and third quartiles of a distribution.

- (a) What is the value of the *IQR* for the standard Normal distribution?

(b) There is a constant c such that $IQR = c\sigma$ for any Normal distribution $N(\mu, \sigma)$. What is the value of c ?

1.142 CHALLENGE Outliers for Normal distributions.

Continue your work from the previous two exercises. The percent of the observations that are suspected outliers according to the $1.5 \times IQR$ rule is the same for any Normal distribution. What is this percent?

1.143 Heart rates of runners. Figure 1.39 is a Normal quantile plot of the heart rates of the 200 male runners in the study described in Exercise 1.81 (page 51). The distribution is close to Normal. How can you see this? Describe the nature of the small deviations from Normality that are visible in the plot.

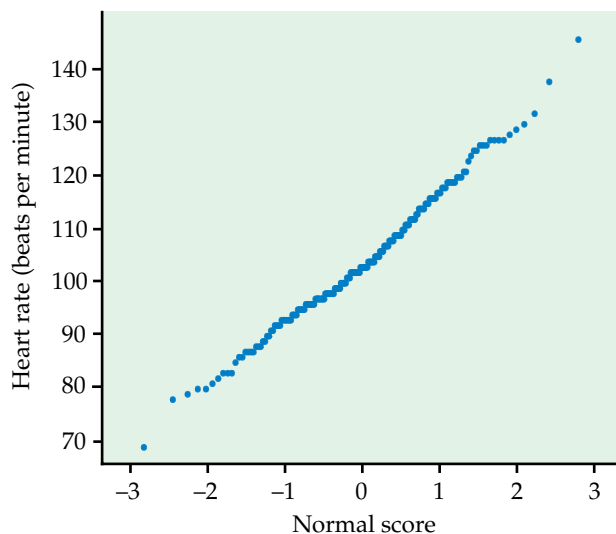


FIGURE 1.39 Normal quantile plot of the heart rates of 200 male runners, for Exercise 1.143.

1.144 Carbon dioxide emissions. Figure 1.40 is a Normal quantile plot of the emissions of carbon dioxide (CO_2) per person in 48 countries, from Table 1.6 (page 26). In what way is this distribution non-Normal? Comparing the plot with Table 1.6, which countries would you call outliers?

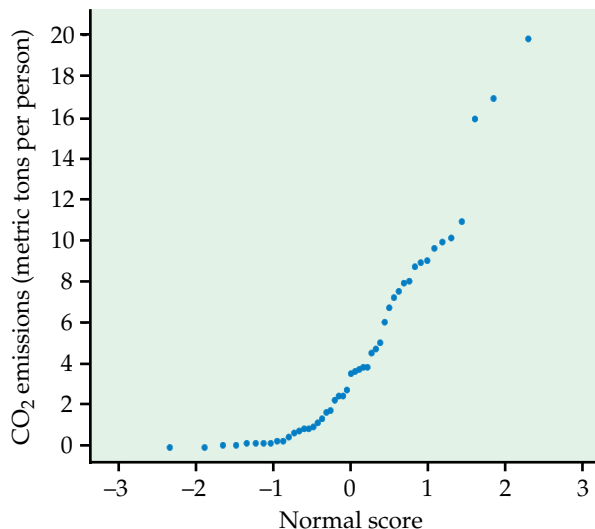


FIGURE 1.40 Normal quantile plot of CO_2 emissions in 48 countries, for Exercise 1.144.

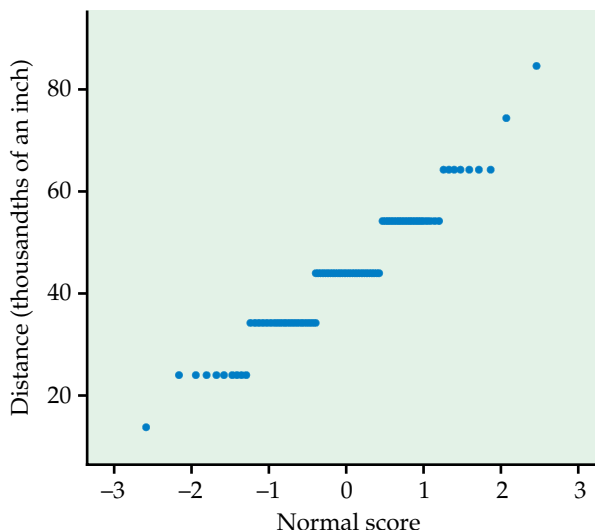


FIGURE 1.41 Normal quantile plot of distance between mounting holes, for Exercise 1.145.

1.145 Electrical meters. The distance between two mounting holes is important to the performance of an electrical meter. The manufacturer measures this distance regularly for quality control purposes, recording the data as thousandths of an inch more than 0.600 inches. For example, 0.644 is recorded as 44. Figure 1.41 is a Normal quantile plot of

the distances for the last 90 electrical meters measured.³⁷ Is the overall shape of the distribution approximately Normal? Why does the plot have a “stair-step” appearance?

1.146 CHALLENGE Four Normal quantile plots. Figure 1.42 shows four Normal quantile plots for data that you have seen before, without scales for the variables plotted. In scrambled order, they are:

1. The IQ scores in the histogram of Figure 1.7 (page 14).

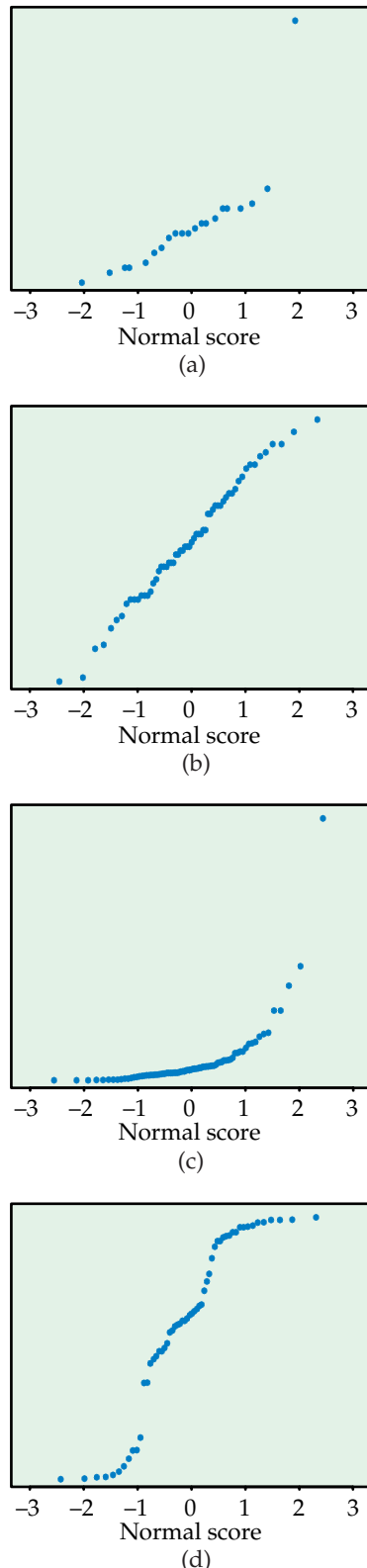


FIGURE 1.42 Four Normal quantile plots, for Exercise 1.146.

2. The tuition and fee charges of Massachusetts colleges in the histogram of Figure 1.16 (page 25).
3. The highway gas mileages of two-seater cars, including the Honda Insight, from Table 1.10 (page 31).
4. The 80 customer service call lengths from Table 1.1, displayed in the stemplot of Figure 1.6 (page 12).

Which Normal quantile plot goes with each data set? Explain the reasons for your choices.

The remaining exercises for this section require the use of software that will make Normal quantile plots.

1.147 Density of the earth. We expect repeated careful measurements of the same quantity to be approximately Normal. Make a Normal quantile plot for Cavendish's measurements in Exercise 1.40 (page 28). Are the data approximately Normal? If not, describe any clear deviations from Normality.

1.148 Three varieties of flowers. The study of tropical flowers and their hummingbird pollinators (Exercise 1.78, page 51) measured lengths for three varieties of *Heliconia* flowers. We expect that such biological measurements will have roughly Normal distributions.

(a) Make Normal quantile plots for each of the three flower varieties. Which distribution is closest to Normal?

(b) The other two distributions show the same kind of mild deviation from Normality. In what way are these distributions non-Normal?

1.149 Logging in Borneo. The study of the effects of logging on tree counts in the Borneo rain forest (Exercise 1.80, page 51) used statistical methods that are based on Normal distributions. Make Normal quantile plots for each of the three groups of forest plots. Are the three distributions roughly Normal? What are the most prominent deviations from Normality that you see?

1.150 Use software to generate some data. Use software to generate 100 observations from the standard Normal distribution. Make a histogram of these observations. How does the shape of the histogram compare with a Normal density curve? Make a Normal quantile plot of the data. Does the plot suggest any important deviations from Normality? (Repeating this exercise several times is a good way to become familiar with how

histograms and Normal quantile plots look when data actually are close to Normal.)

- 1.151 Use software to generate more data.** Use software to generate 100 observations from the uniform distribution described in Exercise 1.108.

Make a histogram of these observations. How does the histogram compare with the density curve in Figure 1.37? Make a Normal quantile plot of your data. According to this plot, how does the uniform distribution deviate from Normality?

CHAPTER 1 Exercises

- 1.152 Park space and population.** Below are data on park and open space in several U.S. cities with high population density.³⁸ In this table, population is reported in thousands of people, and park and open space is called open space, with units of acres.

City	Population	Open space
Baltimore	651	5,091
Boston	589	4,865
Chicago	2,896	11,645
Long Beach	462	2,887
Los Angeles	3,695	29,801
Miami	362	1,329
Minneapolis	383	5,694
New York	8,008	49,854
Oakland	399	3,712
Philadelphia	1,518	10,685
San Francisco	777	5,916
Washington, D.C.	572	7,504

- Make a bar graph for population. Describe what you see in the graph.
- Do the same for open space.
- For each city, divide the open space by population. This gives rates: acres of open space per thousand residents.
- Make a bar graph of the rates.
- Redo the bar graph that you made in part (d) by ordering the cities by their open space to population rate.
- Which of the two bar graphs in (d) and (e) do you prefer? Give reasons for your answer.

- 1.153 Compare two Normal curves.** In Exercise 1.99, we worked with the distribution of ISTEP scores on the English/language arts portion of the exam for tenth-graders. We used the fact that the distribution of scores for the 76,531 students who took the exam was approximately $N(572, 51)$. These students were classified in a variety of ways, and summary statistics were reported for these different subgroups. When classified by gender,

the scores for the women are approximately $N(579, 49)$, and the scores for the men are approximately $N(565, 55)$. Figure 1.43 gives the Normal density curves for these two distributions. Here is a possible description of these data: women score about 14 points higher than men on the ISTEP English/language arts exam. Critically evaluate this statement and then write your own summary based on the distributions displayed in Figure 1.43.

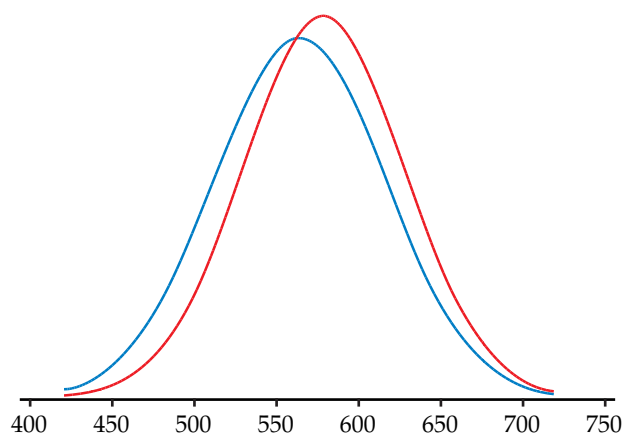


FIGURE 1.43 Normal density curves for ISTEP scores of women and men, for Example 1.53.

- 1.154 Leisure time for college students.** You want to measure the amount of “leisure time” that college students enjoy. Write a brief discussion of two issues:
- How will you define “leisure time”?
 - How will you measure leisure time?

- 1.155 Biological clocks.** Many plants and animals have “biological clocks” that coordinate activities with the time of day. When researchers looked at the length of the biological cycle in the plant *Arabidopsis* by measuring leaf movements, they found that the length of the cycle is not always 24 hours. Further study discovered that cycle length changes systematically with north-south location.

TABLE 1.11

Biological clock cycle lengths for a plant species in different locations

23.89	23.72	23.74	24.35	25.05	24.56	23.69	22.33	23.79	22.12
25.39	23.08	25.64	23.98	25.84	25.46	24.37	24.13	24.40	24.74
24.44	24.82	23.56	24.96	24.21	23.85	24.57	23.44	23.64	24.23
24.01	24.58	25.57	23.73	24.11	23.21	25.08	24.03	24.62	23.51
23.21	23.41	23.69	22.97	24.65	24.65	24.29	23.89	25.08	23.89
24.95	23.09	23.21	24.66	23.88	25.33	24.38	24.68	25.34	25.22
23.45	23.39	25.43	23.16	23.95	23.25	24.72	24.89	24.88	24.71
23.58	25.98	24.28	24.25	23.16	24.19	27.22	23.77	26.21	24.33
24.34	24.89	24.32	24.14	24.00	23.48	25.81	24.99	24.18	22.73
24.18	23.95	24.48	23.89	24.24	24.96	24.58	24.29	24.31	23.64
23.87	23.68	24.87	23.00	23.48	24.26	23.34	25.11	24.69	24.97
24.64	24.49	23.61	24.07	26.60	24.91	24.76	25.09	26.56	25.13
24.81	25.63	25.63	24.69	24.41	23.79	22.88	22.00	23.33	25.12
24.00	24.31	23.03	24.51	28.55	22.96	23.61	24.72	24.04	25.18
24.30	24.22	24.39	24.73	24.68	24.14	24.57	24.42	25.62	

Table 1.11 contains cycle lengths for 149 locations around the world.³⁹ Describe the distribution of cycle lengths with a histogram and numerical summaries. In particular, how much variation is there among locations?

- 1.156 Product preference.** Product preference depends in part on the age, income, and gender of the consumer. A market researcher selects a large sample of potential car buyers. For each consumer, she records gender, age, household income, and automobile preference. Which of these variables are categorical and which are quantitative?
- 1.157 Distance-learning courses.** The 222 students enrolled in distance-learning courses offered by a college ranged from 18 to 64 years of age. The mode of their ages was 19. The median age was 31.⁴⁰ Explain how this can happen.
- 1.158 Internet service.** Late in 2003, there were 77.4 million residential subscribers to Internet service in the United States. The numbers of subscribers claimed by the top 10 providers of service were as follows.⁴¹ (There is some doubt about the accuracy of these claims.)

Service provider	Subscribers (millions)	Service provider	Subscribers (millions)
America Online	24.7	SBC	3.1
MSN	8.7	Verizon	2.1
United Online	5.2	Cox	1.8
EarthLink	5.0	Charter	1.5
Comcast	4.9	BellSouth	1.3

Display these data in a graph. How many subscribers do the many smaller providers have? Add an “Other” entry in your graph. Business

people looking at this graph see an industry that offers opportunities for larger companies to take over.

- 1.159 Weights are not Normal.** The heights of people of the same sex and similar ages follow Normal distributions reasonably closely. Weights, on the other hand, are not Normally distributed. The weights of women aged 20 to 29 have mean 141.7 pounds and median 133.2 pounds. The first and third quartiles are 118.3 pounds and 157.3 pounds. What can you say about the shape of the weight distribution? Explain your reasoning.
- 1.160 What graph would you use?** What type of graph or graphs would you plan to make in a study of each of the following issues?
- (a) What makes of cars do students drive? How old are their cars?
 - (b) How many hours per week do students study? How does the number of study hours change during a semester?
 - (c) Which radio stations are most popular with students?
 - (d) When many students measure the concentration of the same solution for a chemistry course laboratory assignment, do their measurements follow a Normal distribution?
- 1.161 Household size and household income.** Rich and poor households differ in ways that go beyond income. Figure 1.44 displays histograms that compare the distributions of household size (number of people) for low-income and high-income households in 2002.⁴² Low-income households had incomes less than \$15,000, and

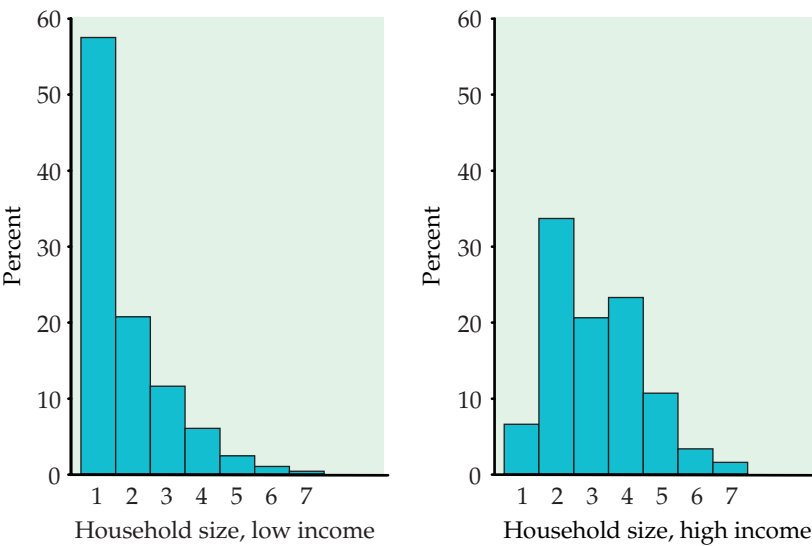


FIGURE 1.44 The distributions of household size for households with incomes less than \$15,000 (left) and households with incomes of at least \$100,000 (right), for Exercise 1.161.

high-income households had incomes of at least \$100,000.

- (a) About what percent of each group of households consisted of two people?
- (b) What are the important differences between these two distributions? What do you think explains these differences?

1.162 Spam filters. A university department installed a spam filter on its computer system. During a 21-day period, 6693 messages were tagged as spam. How much spam you get depends on what your online habits are. Here are the counts for some students and faculty in this department (with log-in IDs changed, of course):

ID	Count	ID	Count	ID	Count	ID	Count
AA	1818	BB	1358	CC	442	DD	416
EE	399	FF	389	GG	304	HH	251
II	251	JJ	178	KK	158	LL	103

All other department members received fewer than 100 spam messages. How many did the others receive in total? Make a graph and comment on what you learn from these data.

1.163 Two distributions. If two distributions have exactly the same mean and standard deviation, must their histograms have the same shape? If they have the same five-number summary, must their histograms have the same shape? Explain.

1.164 By-products from DDT. By-products from the pesticide DDT were major threats to the survival of birds of prey until use of DDT was banned at the end of 1972. Can time plots show the effect of the ban? Here are two sets of data for bald eagles nesting in the forests of northwestern Ontario.⁴³ The data set below gives the mean number of young per breeding area.

Year	1966	1967	1968	1969	1970	1971	1972	1973
Young	1.26	0.73	0.89	0.84	0.54	0.60	0.54	0.78
Year	1974	1975	1976	1977	1978	1979	1980	1981
Young	0.46	0.77	0.86	0.96	0.82	0.98	0.93	1.12

The following data are measurements of the chemical DDE (the by-product of DDT that most threatens birds of prey) from bald eagle eggs in the same area of Canada. These are in parts per million (ppm). There are often several measurements per year.

Year	1967	1967	1968	1971	1971	1972	1976
DDE	44	95	121	125	95	87	13.3
Year	1976	1976	1976	1976	1977	1977	1980
DDE	16.4	50.4	59.8	56.4	0.6	23.8	16.6
Year	1980	1980	1981	1981	1981		
DDE	14.5	24.0	7.8	48.2	53.4		

Make time plots of eagle young and of mean DDE concentration in eggs. How does the effect of banning DDT appear in your plots?

- 1.165 Babe Ruth and Mark McGwire.** Babe Ruth hit 60 home runs in 1927, a record that stood until Mark McGwire hit 70 in 1998. A proper comparison of Ruth and McGwire should include their historical context. Here are the number of home runs by the major league leader for each year in baseball history, 1876 to 2003, in order from left to right. Make a time plot. (Be sure to add the scale of years.)

5	3	4	9	6	7	7	10	27	11	11	17	14	20	14	16	13
19	18	17	13	11	15	25	12	16	16	13	10	9	12	10	12	9
10	21	14	19	19	24	12	12	11	29	54	59	42	41	46	39	47
60	54	46	56	46	58	48	49	36	49	46	58	35	43	37	36	34
33	28	44	51	40	54	47	42	37	47	49	51	52	44	47	46	41
61	49	45	49	52	49	44	44	49	45	48	40	44	36	38	38	52
46	48	48	31	39	40	43	40	40	49	42	47	51	44	43	46	43
50	52	58	70	65	50	73	57	47								

- (a) Describe the effect of World War II (1942 to 1945 seasons).
- (b) Ruth led in the 11 years in boldface between 1918 and 1931. McGwire led in the 5 boldface years between 1987 and 1999. Briefly compare the achievements of Ruth and McGwire in the context of their times.
- 1.166 Barry Bonds.** The single-season home run record was broken by Barry Bonds of the San Francisco Giants in 2001, when he hit 73 home runs. Here are Bonds's home run totals from 1986 (his first year) to 2003:

16	25	24	19	33	25	34	46	37
33	42	40	37	34	49	73	46	45

Make a stemplot of these data. Bonds's record year is a high outlier. How do his career mean and median number of home runs change when we drop the record 73? What general fact about the mean and median does your result illustrate?

- 1.167 CHALLENGE Norms for reading scores.** Raw scores on behavioral tests are often transformed for easier comparison. A test of reading ability has mean 75 and standard deviation 10 when given to third-graders. Sixth-graders have mean score 82 and standard deviation 11 on the same test. To provide separate "norms" for each grade, we want scores in each grade to have mean 100 and standard deviation 20.

(a) What linear transformation will change third-grade scores x into new scores $x_{\text{new}} = a + bx$ that have the desired mean and standard deviation? (Use $b > 0$ to preserve the order of the scores.)

(b) Do the same for the sixth-grade scores.

(c) David is a third-grade student who scores 78 on the test. Find David's transformed score. Nancy is a sixth-grade student who scores 78. What is her transformed score? Who scores higher within his or her grade?

(d) Suppose that the distribution of scores in each grade is Normal. Then both sets of transformed scores have the $N(100, 20)$ distribution. What percent of third-graders have scores less than 78? What percent of sixth-graders have scores less than 78?

- 1.168 Damage caused by tornados.** The average damage caused by tornados in the states (Table 1.5, page 25) and the estimated amount of oil recovered from different oil wells (Exercise 1.39, page 28) both have right-skewed distributions. Choose one of these data sets. Make a Normal quantile plot. How is the skewness of the distribution visible in the plot? Based on the plot, which observations (if any) would you call outliers?
- 1.169 Proportions older than 65.** We know that the distribution of the percents of state residents over 65 years of age has a low outlier (Alaska) and a high outlier (Florida). The stemplot in Exercise 1.21 (page 24) looks unimodal and roughly symmetric.
- (a) Sketch what a Normal quantile plot would look like for a distribution that is Normal except for two outliers, one in each direction.
- (b) If your software includes Normal quantile plots, make a plot of the percent-over-65 data and discuss what you see.
- 1.170 Returns on stocks.** Returns on common stocks are "heavy tailed." That is, they have more values far from the center in both the low and the high tails than a Normal distribution would have. However, average returns for many individual stocks over longer periods of time become more Normal.
- (a) Sketch the appearance of a Normal quantile plot for a distribution having roughly Normal center and heavy tails. Explain the reasoning behind your sketch.
- (b) The data include the annual returns for the years 1950 to 2003, pictured in the stemplot in Figure 1.22(a). If your software allows, make a Normal quantile plot of these returns. Is the distribution clearly heavy tailed? Are there other clear deviations from Normality?

1.171 Use software to generate some data. Most statistical software packages have routines for generating values of variables having specified distributions. Use your statistical software to generate 25 observations from the $N(20, 5)$ distribution. Compute the mean and standard deviation \bar{x} and s of the 25 values you obtain. How close are \bar{x} and s to the μ and σ of the distribution from which the observations were drawn? Repeat 19 more times the process of generating 25 observations from the $N(20, 5)$ distribution and recording \bar{x} and s . Make a stemplot of the 20 values of \bar{x} and another stemplot of the 20 values of s . Make Normal quantile plots of both sets of data. Briefly describe each of these distributions. Are they symmetric or skewed? Are they roughly Normal? Where are their centers? (The distributions of measures like \bar{x} and s when repeated sets of observations are made from the same theoretical distribution will be very important in later chapters.)

1.172 Distribution of income. Each March, the Bureau of Labor Statistics collects detailed information about more than 50,000 randomly selected households. The WORKERS data set contains

data on 71,076 people from the March 2002 survey. All of these people were between 25 and 64 years of age and worked throughout the year. The Data Appendix describes this data set in detail. Describe the distribution of incomes for these people. Use graphs and numbers, and briefly state your main findings. Because this is a very large randomly selected sample, your results give a good description of incomes for all working Americans aged 25 to 64.

1.173 SAT mathematics scores and grade point averages. The CSDATA data set described in the Data Appendix contains information on 234 computer science students. We are interested in comparing the SAT Mathematics scores and grade point averages of female students with those of male students. Make two sets of side-by-side boxplots to carry out these comparisons. Write a brief discussion of the male-female comparisons. Then make Normal quantile plots of grade point averages and SAT Math scores separately for men and women. Which students are clear outliers? Which of the four distributions are approximately Normal if we ignore outliers?