1.4 Calculate the grade. A student whose data do not appear on the spreadsheet scored 83 on Exam1, 82 on Exam2, 77 for Homework, 90 on the Final, and 80 on the Project. Find TotalPoints for this student and give the grade earned.

spreadsheet

The display in Figure 1.2 is from an Excel **spreadsheet.** Spreadsheets are very useful for doing the kind of simple computations that you did in Exercise 1.4. You can type in a formula and have the same computation performed for each row.

Note that the names we have chosen for the variables in our spreadsheet do not have spaces. For example, we could have used the name "Exam 1" for the first-exam score rather than Exam1. In some statistical software packages, however, spaces are not allowed in variable names. *For this reason, when creating spreadsheets for eventual use with statistical software, it is best to avoid spaces in variable names.* Another convention is to use an underscore (_) where you would normally use a space. For our data set, we could use Exam_1, Exam_2, and Final_Exam.

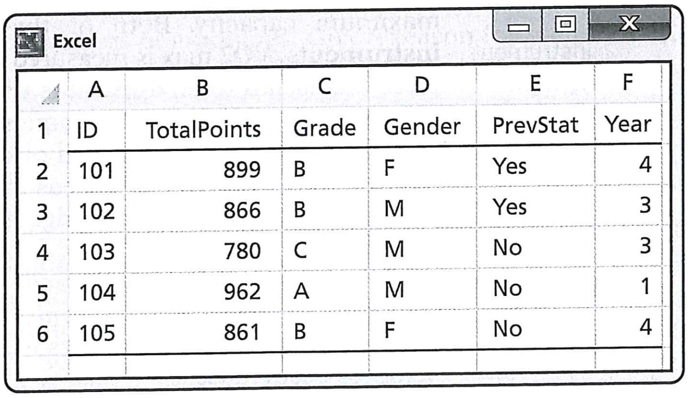
EXAMPLE

1.4 Cases and variables for the statistics class data. The data set in Figure 1.2 was constructed to keep track of the grades for students in an introductory statistics course. The cases are the students in the class. There are eight variables in this data set. These include a label for each student and scores for the various course requirements. There are no units for ID and grade. The other variables all have "points" as the unit.

EXAMPLE

1.5 Statistics class data for a different purpose. Suppose that the data for the students in the introductory statistics class were also to be used to study relationships between student characteristics and success in the course. For this purpose, we might want to use a data set like the spreadsheet in Figure 1.3.

FIGURE 1.3 Spreadsheet for Example 1.5.



	A	B	C	D	E	F
1	ID	TotalPoints	Grade	Gender	PrevStat	Year
2	101	899	B	F	Yes	4
3	102	866	B	M	Yes	3
4	103	780	C	M	No	3
5	104	962	A	M	No	1
6	105	861	B	F	No	4

SECTION 1.2 Summary

Exploratory data analysis uses graphs and numerical summaries to describe the variables in a data set and the relations among them.

The **distribution** of a variable tells us what values it takes and how often it takes these values.

Bar graphs and **pie charts** display the distributions of categorical variables. These graphs use the counts or percents of the categories.

Stemplots and **histograms** display the distributions of quantitative variables. Stemplots separate each observation into a **stem** and a one-digit **leaf.** Histograms plot the **frequencies** (counts) or the percents of equal-width classes of values.

When examining a distribution, look for **shape, center,** and **spread** and for clear **deviations** from the overall shape.

Some distributions have simple shapes, such as **symmetric** or **skewed.** The number of **modes** (major peaks) is another aspect of overall shape. Not all distributions have a simple overall shape, especially when there are few observations.

Outliers are observations that lie outside the overall pattern of a distribution. Always look for outliers and try to explain them.

When observations on a variable are taken over time, make a **time plot** that graphs time horizontally and the values of the variable vertically. A time plot can reveal changes over time.

SECTION 1.2 Exercises

For Exercise 1.16, see page 12; for Exercise 1.17, see page 14; for Exercises 1.18 and 1.19, see page 15; for Exercise 1.20, see page 17; for Exercises 1.21 and 1.22, see page 18; for Exercise 1.23, see page 21; and for Exercise 1.24, see page 22.

1.25 The *Titanic* and class. On April 15, 1912, on her maiden voyage, the *Titanic* collided with an iceberg and sank. The ship was luxurious but did not have enough lifeboats for the 2224 passengers and crew. As a result of the collision, 1502 people died.[9] The ship had three classes of passengers. The level of luxury and the price of the ticket varied with the class, with first class being the most luxurious. There were 323 passengers in first class, 277 in second class, and 709 in third class.[10] TITANIC

(a) Make a bar graph of these data.

(b) Give a short summary of how the number of passengers varied with class.

(c) If you made a bar graph of the percents of passengers in each class, would the general features of the graph differ from the one you made in part (a)? Explain your answer.

1.26 Another look at the *Titanic* and class. Refer to the previous exercise. TITANIC

(a) Make a pie chart to display the data.

(b) Compare the pie chart with the bar graph. Which do you prefer? Give reasons for your answer.

1.27 Who survived? Refer to the two previous exercises. The number of first-class passengers who survived was 200. For second and third class, the numbers were 119 and 181, respectively. Create a graphical summary that shows how the survival of passengers depended on class. TITANIC

1.28 Do you use your Twitter account? Although Twitter has more than 500,000,000 users, only about 170,000,000 are active. A study of Twitter account usage defined an active account as one with at least one message posted within a three-month period. Here are the percents of active accounts for 20 countries:[11] TWITTC

Country	Percent	Country	Percent	Country	Percent
Argentina	25	India	19	South Korea	24
Brazil	25	Indonesia	28	Spain	29
Canada	28	Japan	30	Turkey	25
Chile	24	Mexico	26	United Kingdom	26
Colombia	26	Netherlands	33	United States	28
France	24	Philippines	22	Venezuela	28
Germany	23	Russia	26		

(a) Make a stemplot of the distribution of percents of active accounts.

(b) Describe the overall pattern of the data and any deviations from that pattern.

(c) Identify the shape, center, and spread of the distribution.

(d) Identify and describe any outliers.

1.29 Another look at Twitter account usage. Refer to the previous exercise. TWITTC

(a) Use a histogram to summarize the distribution.

(b) Use this histogram to answer parts (b), (c), and (d) of the previous exercise.

(c) Which graphical display, stemplot or histogram, is more useful for describing this distribution? Give reasons for your answer.

1.30 Energy consumption. The U.S. Energy Information Administration reports data summaries of various energy statistics. Let's look at the total amount of energy consumed, in quadrillions of British thermal units (Btu), for each month in 2011. Here are the data:[12] ENERGY

Month	Energy (quadrillion Btu)	Month	Energy (quadrillion Btu)
January	9.33	July	8.41
February	8.13	August	8.43
March	8.38	September	7.58
April	7.54	October	7.61
May	7.61	November	7.81
June	7.92	December	8.60

(a) Look at the table and describe how the energy consumption varies from month to month.

(b) Make a time plot of the data and describe the patterns.

(c) Suppose you wanted to communicate information about the month-to-month variation in energy consumption. Which would be more effective, the table of the data or the graph? Give reasons for your answer.

1.31 Energy consumption in a different year. Refer to the previous exercise. Here are the data for 2010: ENERGY

Month	Energy (quadrillion Btu)	Month	Energy (quadrillion Btu)
January	9.13	July	8.38
February	8.21	August	8.44
March	8.21	September	7.69
April	7.37	October	7.51
May	7.68	November	7.80
June	8.01	December	9.23

(a) Analyze these data using the questions in the previous exercise as a guide.

(b) Compare the patterns in 2010 with those in 2011. Describe any similarities and differences.

1.32 Favorite colors. What is your favorite color? One survey produced the following summary of responses to that question: blue, 42%; green, 14%; purple, 14%; red, 8%; black, 7%; orange, 5%; yellow, 3%; brown, 3%; gray, 2%; and white, 2%.[13] Make a bar graph of the percents and write a short summary of the major features of your graph. FAVCOL

1.33 Least-favorite colors. Refer to the previous exercise. The same study also asked people about their least-favorite color. Here are the results: orange, 30%; brown, 23%; purple, 13%; yellow, 13%; gray, 12%; green, 4%; white, 4%; red, 1%; black, 0%; and blue, 0%. Make a bar graph of these percents and write a summary of the results. LFAVCOL

1.34 Garbage. The formal name for garbage is "municipal solid waste." Here is a breakdown of the materials that make up American municipal solid waste:[14] GARBAGE

Material	Weight (million tons)	Percent of total (%)
Food scraps	34.8	13.9
Glass	11.5	4.6
Metals	22.4	9.0
Paper, paperboard	71.3	28.5
Plastics	31.0	12.4
Rubber, leather, textiles	20.9	8.4
Wood	15.9	6.4
Yard trimmings	33.4	13.4
Other	8.6	3.2
Total	249.6	100.0

(a) Add the weights and then the percents for the nine types of material given, including "Other." Each entry, including the total, is separately rounded to the nearest tenth. So the sum and the total may slightly because of **roundoff error.**

(b) Make a bar graph of the percents. The graph gives a clearer picture of the main contributors to garbage if you order the bars from tallest to shortest.

(c) Make a pie chart of the percents. Compare the advantages and disadvantages of each graphical summary. Which do you prefer? Give reasons for your answer.

(a) Compute the mean for these data.

(b) Compute the median for these data.

(c) Which measure do you prefer for describing the center of this distribution? Explain your answer. (You may include a graphical summary as part of your explanation.)

1.62 Measures of spread for the double stout data. Refer to the previous exercise. STOUT

(a) Compute the standard deviation for these data.

(b) Compute the quartiles for these data.

(c) Which measure do you prefer for describing the spread of this distribution? Explain your answer. (You may include a graphical summary as part of your explanation.)

1.63 Are there outliers in the double stout data? Refer to Exercise 1.61. STOUT

(a) Find the *IQR* for these data.

(b) Use the 1.5 × *IQR* rule to identify and name any outliers.

(c) Make a boxplot for these data and describe the distribution using only the information in the boxplot.

(d) Make a modified boxplot for these data and describe the distribution using only the information in the boxplot.

(e) Make a stemplot for these data.

(f) Compare the boxplot, the modified boxplot, and the stemplot. Evaluate the advantages and disadvantages of each graphical summary for describing the distribution of the double stout data.

1.64 Smolts. Smolts are young salmon at a stage when their skin becomes covered with silvery scales and they start to migrate from freshwater to the sea. The reflectance of a light shined on a smolt's skin is a measure of the smolt's readiness for the migration. Here are the reflectances, in percents, for a sample of 50 smolts:[25] SMOLTS

57.6	54.8	63.4	57.0	54.7	42.3	63.6	55.5	33.5	63.3
58.3	42.1	56.1	47.8	56.1	55.9	38.8	49.7	42.3	45.6
69.0	50.4	53.0	38.3	60.4	49.3	42.8	44.5	46.4	44.3
58.9	42.1	47.6	47.9	69.2	46.6	68.1	42.8	45.6	47.3
59.6	37.8	53.9	43.2	51.4	64.5	43.8	42.7	50.9	43.8

(a) Find the mean reflectance for these smolts.

(b) Find the median reflectance for these smolts.

(c) Do you prefer the mean or the median as a measure of center for these data? Give reasons for your preference.

1.65 Measures of spread for smolts. Refer to the previous exercise. SMOLTS

(a) Find the standard deviation of the reflectance for these smolts.

(b) Find the quartiles of the reflectance for these smolts.

(c) Do you prefer the standard deviation or the quartiles as a measure of spread for these data? Give reasons for your preference.

1.66 Are there outliers in the smolt data? Refer to Exercise 1.64. SMOLTS

(a) Find the *IQR* for the smolt data.

(b) Use the 1.5 × *IQR* rule to identify any outliers.

(c) Make a boxplot for the smolt data and describe the distribution using only the information in the boxplot.

(d) Make a modified boxplot for these data and describe the distribution using only the information in the boxplot.

(e) Make a stemplot for these data.

(f) Compare the boxplot, the modified boxplot, and the stemplot. Evaluate the advantages and disadvantages of each graphical summary for describing the distribution of the smolt reflectance data.

1.67 The value of brands. A brand is a symbol or images that are associated with a company. An effective brand identifies the company and its products. Using a variety of measures, dollar values for brands can be calculated.[26] The most valuable brand is Apple, with a value of $76.568 million. Apple is followed by Google at $69.726 million, Coca-Cola at $67.839 million, Microsoft at $57.853 million, and IBM at $57.532 million. For this exercise you will use the brand values (in millions of dollars) for the top 100 brands in the data file BRANDS. BRANDS

(a) Graphically display the distribution of the values of these brands.

(b) Use numerical measures to summarize the distribution.

(c) Write a short paragraph discussing the dollar values of the top 100 brands. Include the results of your analysis.

1.68 Alcohol content of beer. Brewing beer involves a variety of steps that can affect the alcohol content. The data file BEER gives the percent alcohol for 153 domestic brands of beer.[27] BEER

(a) Use graphical and numerical summaries of your choice to describe these data. Give reasons for your choices.

(b) Give the alcohol content and the brand of any outliers. Explain how you determined that they were outliers.

1.69 Remove the outliers for alcohol content of beer. Refer to the previous exercise. BEER

(a) Calculate the mean with and without the outliers. Do the same for the median. Explain how these statistics change when the outliers are excluded.

(b) Calculate the standard deviation with and without the outliers. Do the same for the quartiles. Explain how these statistics change when the outliers are excluded.

(c) Write a short paragraph summarizing what you have learned in this exercise.

1.70 Calories in beer. Refer to the previous two exercises. The data file also gives the calories per 12 ounces of beverage. BEER

(a) Analyze the data and summarize the distribution of calories for these 153 brands of beer.

(b) In Exercise 1.68 you identified outliers. To what extent are these brands outliers in the distribution of calories? Explain your answer.

1.71 Potatoes. A quality product is one that is consistent and has very little variability in its characteristics. Controlling variability can be more difficult with agricultural products than with those that are manufactured. The following table gives the weights, in ounces, of the 25 potatoes sold in a 10-pound bag. POTATO

7.6	7.9	8.0	6.9	6.7	7.9	7.9	7.9	7.6	7.8	7.0	4.7	7.6
6.3	4.7	4.7	4.7	6.3	6.0	5.3	4.3	7.9	5.2	6.0	3.7	

(a) Summarize the data graphically and numerically. Give reasons for the methods you chose to use in your summaries.

(b) Do you think that your numerical summaries do an effective job of describing these data? Why or why not?

(c) There appear to be two distinct clusters of weights for these potatoes. Divide the sample into two subsamples based on the clustering. Give the mean and standard deviation for each subsample. Do you think that this way of summarizing these data is better than a numerical summary that uses all the data as a single sample? Give a reason for your answer.

1.72 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 on 584 of these trees.[28] 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: PINES

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

(a) Find the five-number summary for these data.

(b) Make a boxplot.

(c) Make a histogram.

(d) Write a short summary of the major features of this distribution. Do you prefer the boxplot or the histogram for these data?

1.73 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. 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.[29] The units are milligrams per liter (mg/l). Here are the data from a random sample of 40 of these children: CRP

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

(a) Find the five-number summary for these data.

(b) Make a boxplot.

(c) Make a histogram.

(d) Write a short summary of the major features of this distribution. Do you prefer the boxplot or the histogram for these data?

1.74 Does a log transform reduce the skewness? 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. CRP

1.75 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. Here are the data on the same sample of 40 children from this study. The units are micromoles per liter (μmol/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. VITA

1.76 Luck and puzzle solving. Children in a psychology study were asked to solve some puzzles and were then given feedback on their performance. They then were asked to rate how luck played a role in determining their scores.[30] 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. LUCK

1.77 Median versus mean for net worth. A report on the assets of American households says that the median net worth of U.S. families is \$77,300. The mean net worth of these families is \$498,800.[31] What explains the difference between these two measures of center?

1.78 Create a data set. Create a data set with 9 observations for which the median would change by a large amount if the smallest observation were deleted.

1.79 Mean versus median. A small accounting firm pays each of its six clerks \$45,000, two junior accountants \$70,000 each, and the firm's owner \$420,000. What is the mean salary paid at this firm? How many of the employees earn less than the mean? What is the median salary?

1.80 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?

1.81 How does the median change? The firm in Exercise 1.79 gives no raises to the clerks and junior accountants, while the owner's take increases to \$500,000. How does this change affect the mean? How does it affect the median?

1.82 Metabolic rates. Calculate the mean and standard deviation of the metabolic rates in Example 1.33 (page 42), 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. METABOL

1.83 Earthquakes. Each year there are about 900,000 earthquakes of magnitude 2.5 or less that are usually not felt. In contrast, there are about 10 of magnitude 7.0 that cause serious damage.[32] Explain why the average magnitude of earthquakes is not a good measure of their impact.

1.84 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 1.3 (page 29). IQGPA

(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?

1.85 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?

1.86 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)?

1.87 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.

1.88 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:[33]

			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 flowers? HELICON

1.89 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. HELICON

(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?

1.90 Imputation. Various problems with data collection can cause some observations to be missing. Suppose a data set has 20 cases. Here are the values of the variable x for 10 of these cases: IMPUTE

17 6 12 14 20 23 9 12 16 21

The values for the other 10 cases are missing. One way to deal with missing data is called **imputation.** The basic idea is that missing values are replaced, or imputed, with values that are based on an analysis of the data that are not missing. For a data set with a single variable, the usual choice of a value for imputation is the mean of the values that are not missing. The mean for this data set is 15.

(a) Verify that the mean is 15 and find the standard deviation for the 10 cases for which x is not missing.

(b) Create a new data set with 20 cases by setting the values for the 10 missing cases to 15. Compute the mean and standard deviation for this data set.

(c) Summarize what you have learned about the possible effects of this type of imputation on the mean and the standard deviation.

1.91 Create a data set. Give an example of a small set of data for which the mean is smaller than the third quartile.

1.92 Create another data set. Create a set of 5 positive numbers (repeats allowed) that have median 11 and mean 8. What thought process did you use to create your numbers?

1.93 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.

(a) Choose four numbers that have the smallest possible standard deviation.

(b) Choose four numbers that have the largest possible standard deviation.

(c) Is more than one choice possible in either (a) or (b)? Explain.

1.94 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.

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

30,001 30,002 30,003

have mean $\bar{x} = 30{,}002$ and standard deviation $s = 1$. Adding a 0 in the center of each number, the next set becomes

300,001 300,003 300,003

has three peaks, one around \$300, another around \$600, and a third around \$1100. Inspection of the data suggests that these correspond roughly to three different types of seats: lower-level seats, club seats, and special luxury seats.

trimodal distribution

bimodal distribution

Many distributions that we have met have a single peak, or mode. The distribution described in Example 1.48 has three modes and is called a **trimodal distribution.** A distribution that has two modes is called a **bimodal distribution.**

The previous example reminds of a continuing theme for data analysis. We looked at a histogram and a density estimate and saw something interesting. This led us to speculation. Additional data on the type and location of the seats may explain more about the prices than we see in Figure 1.31.

SECTION 1.4 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. All Normal distributions satisfy the **68–95–99.7 rule.**

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.

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 1.4 Exercises

For Exercises 1.101 and 1.102, see page 61; for Exercises 1.103 and 1.104, see page 62; for Exercises 1.105 and 1.106, see page 67; and for Exercises 1.107 and 1.108, see page 68.

1.109 Means and medians.

(a) Sketch a symmetric distribution that is *not* Normal. Mark the location of the mean and the median.

(b) Sketch a distribution that is skewed to the left. Mark the location of the mean and the median.

1.110 The effect of changing the standard deviation.

(a) Sketch a Normal curve that has mean 20 and standard deviation 5.

(b) On the same x axis, sketch a Normal curve that has mean 20 and standard deviation 10.

(c) How does the Normal curve change when the standard deviation is varied but the mean stays the same?

1.111 The effect of changing the mean.

(a) Sketch a Normal curve that has mean 20 and standard deviation 5.

(b) On the same x axis, sketch a Normal curve that has mean 30 and standard deviation 5.

(c) How does the Normal curve change when the mean is varied but the standard deviation stays the same?

1.112 NAEP music scores. In Exercise 1.101 (page 61) we examined the distribution of NAEP scores for the twelfth-grade reading skills assessment. For eighth-grade students the average music score is approximately Normal with mean 150 and standard deviation 35.

(a) Sketch this Normal distribution.

(b) Make a table that includes values of the scores corresponding to plus or minus one, two, and three standard deviations from the mean. Mark these points on your sketch along with the mean.

(c) Apply the 68–95–99.7 rule to this distribution. Give the ranges of reading score values that are within one, two, and three standard deviations of the mean.

1.113 NAEP U.S. history scores. Refer to the previous exercise. The scores for twelfth-grade students on the U.S. history assessment are approximately $N(288, 32)$. Answer the questions in the previous exercise for this assessment.

1.114 Standardize some NAEP music scores. The NAEP music assessment scores for eighth-grade students are approximately $N(150, 35)$. Find z-scores by standardizing the following scores: 150, 140, 100, 180, 230.

1.115 Compute the percentile scores. Refer to the previous exercise. When scores such as the NAEP assessment scores are reported for individual students, the actual values of the scores are not particularly meaningful. Usually, they are transformed into percentile scores. The percentile score is the proportion of students who would score less than or equal to the score for the individual student. Compute the percentile scores for the five scores in the previous exercise. State whether you used software or Table A for these computations.

1.116 Are the NAEP U.S. history scores approximately Normal? In Exercise 1.113, we assumed that the NAEP U.S. history scores for twelfth-grade students are approximately Normal with the reported mean and standard deviation, $N(288, 32)$. Let's check that assumption. In addition to means and standard deviations, you can find selected percentiles for the NAEP assessments (see previous exercise). For the twelfth-grade U.S. history scores, the following percentiles are reported:

Percentile	Score
10%	246
25%	276
50%	290
75%	311
90%	328

Use these percentiles to assess whether or not the NAEP U.S. history scores for twelfth-grade students are approximately Normal. Write a short report describing your methods and conclusions.

1.117 Are the NAEP mathematics scores approximately Normal? Refer to the previous exercise. For the NAEP mathematics scores for twelfth-graders the mean is 153 and the standard deviation is 34. Here are the reported percentiles:

Percentile	Score
10%	110
25%	130
50%	154
75%	177
90%	197

Is the $N(153, 34)$ distribution a good approximation for the NAEP mathematics scores? Write a short report describing your methods and conclusions.

1.118 Do women talk more? Conventional wisdom suggests that women are more talkative than men. One study designed to examine this stereotype collected data on the speech of 42 women and 37 men in the United States.[39] TALK

(a) The mean number of words spoken per day by the women was 14,297 with a standard deviation of 6441. Use the 68–95–99.7 rule to describe this distribution.

(b) Do you think that applying the rule in this situation is reasonable? Explain your answer.

(c) The men averaged 14,060 words per day with a standard deviation of 9056. Answer the questions in parts (a) and (b) for the men.

(d) Do you think that the data support the conventional wisdom? Explain your answer. Note that in Section 7.2 we will learn formal statistical methods to answer this type of question.

1.119 Data from Mexico. Refer to the previous exercise. A similar study in Mexico was conducted with 31 women