(c) Now clean up the data so that only true temperature values will be plotted. You can do this by editing either the original text file or the data in your statistical software. Take the time to consider how your specific software handles missing data (if an empty cell won't work, see your software's help function).

(d) Create both a dotplot and a time plot of the cleaned-up annual mean temperatures. Interpret your graphs and conclude in context.

CHAPTER 1 SUMMARY

- A data set contains information on a number of **individuals.** Individuals may be people, animals, or things. For each individual, the data give values for one or more **variables.** A variable describes some characteristic of an individual, such as a person's height, sex, or age.
- Some variables are **categorical** and others are **quantitative.** A categorical variable places each individual into a category, such as male or female. A quantitative variable has numerical values that measure some characteristic of each individual, such as height in centimeters or age in years.
- **Exploratory data analysis** uses graphs and numerical summaries to describe the variables in a data set and the relations among them.
- After you understand the background of your data (individuals, variables, units of measurement), the first thing to do almost always is **plot your data.**
- The **distribution** of a variable describes what values the variable takes and how often it takes these values. **Pie charts** and **bar graphs** display the distribution of a categorical variable. Bar graphs can also compare any set of quantities measured in the same units. **Histograms** and **dotplots** display the distribution of a quantitative variable.
- When examining any graph, look for an **overall pattern** and for notable **deviations** from the pattern.
- **Shape, center, and spread** describe the overall pattern of the distribution of a quantitative variable. Some distributions have simple shapes, such as **symmetric** or **skewed.** Not all distributions have a simple overall shape. Describing the shape of a distribution when there are few observations can be particularly challenging.
- **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 **trends, cycles,** or other changes over time.

THIS CHAPTER IN CONTEXT

Practical statistics is the science of extracting meaning out of validly collected data. We will see in Chapters 6 and 7 that the way we collect data drastically affects the conclusions that we may draw. Some studies examine data from an entire population of interest, as in Example 1.3 on all reported cases of infectious diseases for the state of California in 2014. Other studies select a sample from a given population, as did the study in Example 1.1 about the paw usage preferences of 36 tree shews. The ultimate objective, though, is to draw conclusions about the wider population (for example, all tree shrews), a process called statistical inference that we will describe starting in Chapter 14.

Regardless of the ultimate objective, understanding data starts with exploratory data analysis: the use of graphs and numerical summaries to reveal patterns. This may be done

purely for a descriptive purpose—as highlighted in Part I of this book. Alternatively, it may be done to check whether the data we have are suitable for a specific inference procedure—something we will cover in Parts III and IV. In addition, the discussion box on page 26 illustrates how graphs and numerical summaries are useful tools for finding inconsistencies in the data (typos and other errors) before any real analysis can begin.

In this chapter we showed how categorical data can be summarized graphically using pie charts and bar graphs and how the distribution of a quantitative variable can be displayed with histograms and dotplots—or, for time series, with time plots. These are the simplest and most commonly used types of graphs for inspecting one variable at a time. In Chapters 3 through 5 we will also discuss how graphs can be used to examine the relationship between two quantitative variables or two categorical variables. More complex graphs are sometimes used in the life sciences, such as graphs mapping data to a geographical location to depict patterns of epidemic spread, animal migration, or climate change. Their study is beyond the scope of this introductory textbook.

CHECK YOUR SKILLS

1.15 A study of a very large number of pregnant women in Arkansas reports that the women gained, on average, 14 pounds during their pregnancy and that 18% of the women smoked.[20] Which of the following is *not* a variable in this study?

(a) Pregnancy status

(b) Smoking status

(c) Weight gain

1.16 The two variables in the Arkansas study are

(a) both categorical variables.

(b) both quantitative variables.

(c) one categorical variable and one quantitative variable.

The Statistical Abstract of the United States, *prepared by the Census Bureau, provides the number of single-organ transplants for the year 2010, by organ. The next two exercises are based on the following table:*

Heart	2,333
Lung	1,770
Liver	6,291
Kidney	16,898
Pancreas	350
Intestine	151

1.17 The data on single-organ transplants can be displayed in

(a) a pie chart but not a bar graph.

(b) a bar graph but not a pie chart.

(c) either a pie chart or a bar graph.

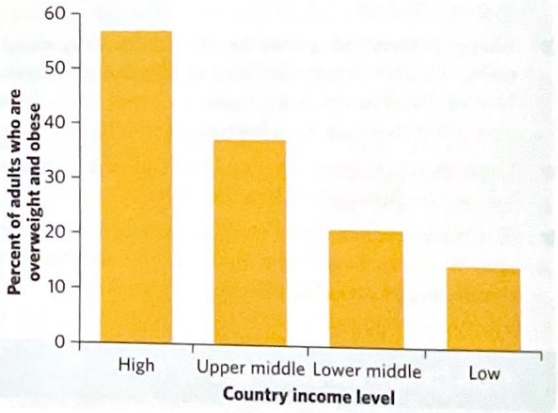
1.18 Kidney transplants represented what percent of single-organ transplants in 2010?

(a) Nearly 61%

(b) One-sixth (nearly 17%)

(c) This percent cannot be calculated from the information provided in the table.

Figure 1.14 shows the percent of adults in the world who are overweight or obese, by type of country of residence based on that country's income level.[21] *The following two exercises are based on this figure.*



▲ **FIGURE 1.14** Percent of adults who are overweight or obese in the world, by country income level.

1.19 The graph in Figure 1.14 is

(a) a bar graph that can be made into one pie chart.

(b) a bar graph that cannot be made into one pie chart.

(c) a histogram with a clear right skew.

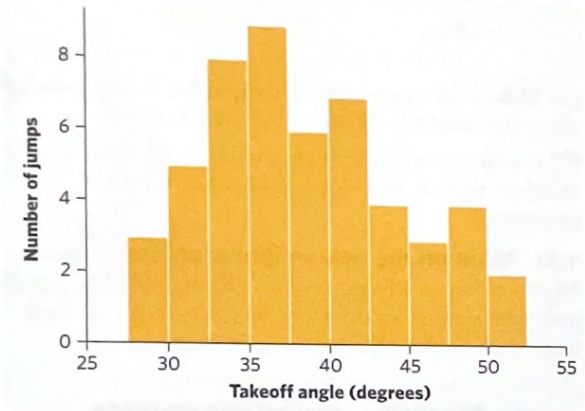
1.20 Which of the following conclusions can be reached from Figure 1.14?

(a) The majority of adults who are overweight and obese live in high-income countries.

(b) The majority of adults who live in high-income countries are overweight and obese.

(c) Both conclusions are correct.

Figure 1.15 is a histogram of the takeoff angles of 51 videotaped jumps of adult hedgehog fleas, Archaeopsyllus erinacei.[22] *The following two exercises are based on this histogram.*



▲ **FIGURE 1.15** Histogram of the takeoff angles (measured in degrees) of 51 flea jumps.

1.21 What percent of jumps have a takeoff angle of 35 degrees or less?

(a) 8% (b) 16% (c) 31%

1.22 The shape of the distribution of takeoff angles in Figure 1.15 is

(a) skewed to the right.

(b) roughly symmetric.

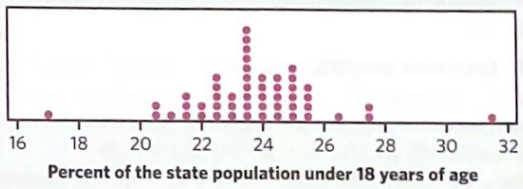
(c) skewed to the left.

1.23 The 2010 U.S. census reported the percent of individuals younger than age 18 in each of the 50 states and the District of Columbia. The data are shown in a doplot in Figure 1.16. This distribution is

(a) single-peaked without outliers.

(b) single-peaked with two outliers.

(c) multiple-peaked with two outliers.



▲ **FIGURE 1.16** Dotplot of the percent of population younger than 18 years of age for the 50 states and D.C.

1.24 The shape of the distribution in the previous exercise is

(a) strongly skewed to the right.

(b) roughly symmetric.

(c) strongly skewed to the left.

CHAPTER 1 EXERCISES

1.25 Endangered species. Bald eagles are an endangered bird species suffering from loss of habitat and pesticide contamination of rivers. A field biologist studying the reproduction of bald eagles records data for the following variables. Which of these variables are categorical, and which are quantitative?

(a) Number of eggs laid

(b) Incubation period (in days)

(c) Parental care (mostly mother, mostly father, both parents)

(d) Nest size (in centimeters)

(e) Presence of pesticides in waters ways (yes/no)

1.26 Eating habits. You are preparing to study the eating habits of elementary schoolchildren. Describe two categorical variables and two quantitative variables that you might record for each child. Give the units of measurement for the quantitative variables.

1.27 Mercury in lakes. Mercury is a metal that is highly toxic to the nervous system. Following is a small part of an EESEE data set ("Mercury in Bass") from a study that assessed the water quality of 53 representative lakes in Florida:[23]

Lake name	pH	Chlorophyll (mg/l)	Avg. mercury in fish (parts per million)	Number of fish sampled	Age of data
Alligator	6.1	0.7	1.23	5	year old
Annie	5.1	3.2	1.33	7	recent
Apopka	9.1	128.3	0.04	6	recent
⋮					

(a) What individuals does this data set describe?

(b) In addition to the lake's name, how many variables does the data set contain? Which of these variables are categorical and which are quantitative?

1.28 Deaths among young people. Here are the number of deaths among persons aged 15 to 24 years in the United States in 2010 due to the leading causes of death for this age group: accidents, 12,015; homicide, 4651; suicide, 4559; cancer, 1594; heart disease, 984; congenital defects, 401.[24]

(a) Make two bar graphs of these data: One with bars ordered alphabetically (by death type) and the other with bars ordered from tallest to shortest. Comparisons are easier if you order the bars by height.

(b) What additional information do you need to make a pie chart?

1.29 Manatee deaths. Manatees are an endangered species of herbivorous, aquatic mammals found primarily in the rivers and estuaries of Florida. As part of its conservation efforts, the Florida Fish and Wildlife Commission records the cause of death for every recovered manatee carcass. Here is a breakdown of the dead manatee counts in Florida for 2012, by cause of death:[25]

Cause of death	Manatees recovered
Watercraft collisions	81
Perinatal	68
Natural	65
Cold stress	28
Flood gate/canal lock	11
Other human	9
Undetermined	
Total	392

(a) Most mortalities recorded as "Undetermined" correspond to manatee carcasses too badly decomposed to make any determination as to the cause of death. How many manatee carcasses had an "undetermined" cause of death?

(b) What is the percent of total manatee deaths caused by collisions with watercraft?

(c) Make a bar graph sorted by manatee count. What does it show?

(d) Could you display these data in a pie chart? Explain why or why not.

1.30 The overweight problem. The 2011 National Health Interview Survey by the National Center for Health Statistics (NCHS) provides weight categorizations for adults 18 years and older based on their body mass index:

Weight category	Percent of adults
Underweight	1.6
Healthy weight	36.3
Overweight	34.2
Obese	27.9

(a) Make a bar graph of these data. What do you conclude about the weight problem in the United States?

(b) Could you make a single pie chart for these data? (Explain why or why not.) If so, what would it emphasize?

1.31 More on the overweight problem. The same NCHS report (see Exercise 1.30) breaks down the sampled individuals by age group. Here are the percents of obese individuals in the 2011 survey for each age group:

Age group	Percent who are obese
18 to 44	26.2
45 to 64	32.2
65 to 74	31.6
75 and over	19.5

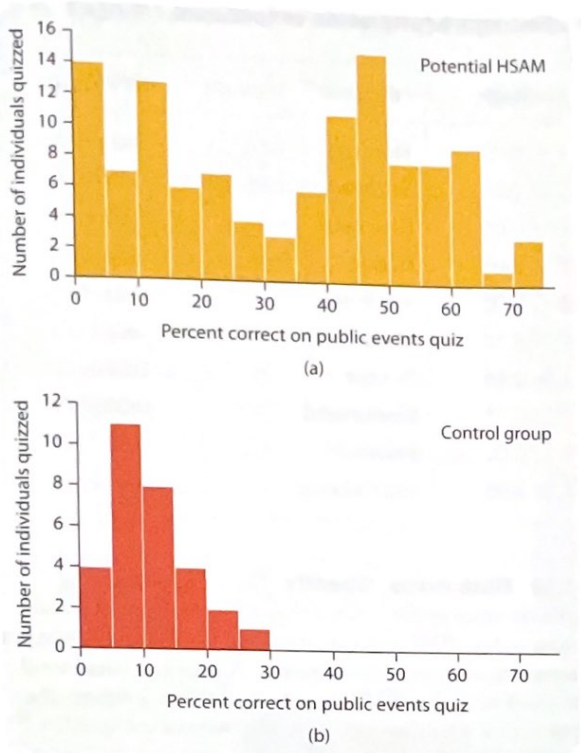
(a) Make a bar graph of these data. What do you conclude about the weight problem in the United States?

(b) Could you make a single pie chart for these data? (Explain why or why not.) If so, what would it emphasize?

1.32 Do you have Highly Superior Autobiographical Memory? Some individuals have the ability to recall accurately vast amounts of autobiographical information without mnemonic tricks or extra practice. This ability is called HSAM, for Highly Superior Autobiographical Memory. A research team administered a quiz on past public events to 115 individuals claiming to have HSAM. The histogram in Figure 1.17(a) shows the distribution of scores on this quiz, expressed as percent of correct answers.[26]

(a) Describe the distribution of quiz scores for individuals claiming to have HSAM.

(b) What does this histogram suggest about the HSAM claims?



▲ **FIGURE 1.17** Histograms of scores on a quiz of past public events (a) for 115 individuals claiming to have Highly Superior Autobiographical Memory (HSAM) and (b) for 30 control individuals.

1.33 Which graph? Of the following data sets, which would you display with a bar graph and which would you display with a histogram? Explain why.

(a) A record of the gender of selected individuals (labeled as male = 0, female = 1)

(b) A record of the age of selected parents (in years) at birth of first child

(c) A record of the height of selected individuals (in inches)

(d) A record of the blood type of selected individuals (A, B, AB, or O)

1.34 Highly Superior Autobiographical Memory, continued. The researchers in Exercise 1.32 also administered the quiz of past public events to 30 individuals who did not claim to have any unusual memory abilities (controls). Their scores are displayed in Figure 1.17(b), using the same horizontal axis scale as that of the histogram of Figure 1.17(a) for easier comparison.

(a) Describe the shape, center, and spread of the distribution of quiz scores for the control individuals.

(b) Compare the histograms in Figure 1.17(a) and (b). How does the distribution of scores among the control individuals support your interpretation of HSAM claims from Exercise 1.32(b)?

1.35 Acid rain. Changing the choice of classes can change the appearance of a histogram. Here is an example in which a small shift in the classes, with no change in the number of classes, has an important effect on the histogram. The data are the acidity levels (measured by pH) in 105 samples of rainwater. Distilled water has pH 7.00. As the water becomes more acid, the pH goes down. The pH of rainwater is important to environmentalists because of the problem of acid rain.[27]

4.33	4.38	4.48	4.48	4.50	4.55	4.59	4.59	4.61	4.61
4.75	4.76	4.78	4.82	4.82	4.83	4.86	4.93	4.94	4.94
4.94	4.96	4.97	5.00	5.01	5.02	5.05	5.06	5.08	5.09
5.10	5.12	5.13	5.15	5.15	5.15	5.16	5.16	5.16	5.18
5.19	5.23	5.24	5.29	5.32	5.33	5.35	5.37	5.37	5.39
5.41	5.43	5.44	5.46	5.46	5.47	5.50	5.51	5.53	5.55
5.55	5.56	5.61	5.62	5.64	5.65	5.65	5.66	5.67	5.67
5.68	5.69	5.70	5.75	5.75	5.75	5.76	5.76	5.79	5.80
5.81	5.81	5.81	5.81	5.85	5.85	5.90	5.90	6.00	6.03
6.03	6.04	6.04	6.05	6.06	6.07	6.09	6.13	6.21	6.34
6.43	6.61	6.62	6.65	6.81					

(a) Make a histogram of pH with 14 classes, using class boundaries 4.2, 4.4, ..., 7.0. Describe this histogram. How many peaks does it show? The presence of more than one peak suggests that the data contain groups that have different distributions.

(b) Make a second histogram, also with 14 classes, using class boundaries 4.14, 4.34, ..., 6.94. The classes are those from (a) moved 0.06 to the left. Describe this new histogram. How many peaks does it show?

1.36 Food oils and health. Fatty acids, despite their unpleasant name, are necessary for human health. Two types of essential fatty acids, called omega-3 and omega-6, are not produced by our bodies and so must be obtained from our food. Food oils, which are widely used in food processing and cooking, are major sources of these compounds. Some evidence suggests that a healthy diet should include more omega-3 than omega-6. Table 1.2 gives the ratio of omega-3 to omega-6 in some common food oils.[28] Values greater than 1 show that an oil has more omega-3 than omega-6.

(a) Make a histogram of these data, using classes bounded by the whole numbers from 0 to 6.

(b) What is the shape of the distribution? How many of the 30 food oils have more omega-3 than omega-6? What does this distribution suggest about the possible health effects of modern food oils?

TABLE 1.2 Omega-3 fatty acids as a fraction of omega-6 fatty acids in food oils

Oil	Ratio	Oil	Ratio	Oil	Ratio
Perilla	5.33	Margarine	0.05	Herring	2.67
Walnut	0.20	Olive	0.08	Soybean, hydrogenated	0.07
Wheat germ	0.13	Shea nut	0.06	Rice bran	0.05
Mustard	0.38	Sunflower (oleic)	0.05	Butter	0.64
Sardine	2.16	Sunflower (linoleic)	0.00	Sunflower	0.03
Salmon	2.50	Flaxseed	3.56	Corn	0.01
Mayonnaise	0.06	Canola	0.46	Sesame	0.01
Cod liver	2.00	Soybean	0.13	Cottonseed	0.00
Shortening (household)	0.11	Grape seed	0.00	Palm	0.02
Shortening (industrial)	0.06	Menhaden	1.96	Cocoa butter	0.04

(c) Table 1.2 contains entries for several fish oils (cod, herring, menhaden, salmon, sardine). How do these values support the idea that eating fish is healthy?

1.37 Stemplots: Healing time. Biologists studying the healing of skin wounds measured the rate at which new cells closed a razor cut made to the skin of an anesthetized newt. Here are the sorted data from 18 newts, measured in micrometers per hour:[29]

11 12 14 18 22 22 23 23 26
27 28 29 30 33 34 35 35 40

(a) Make a dotplot displaying these data and describe the distribution of healing times.

(b) Another type of graphical display used for reasonably small quantitative data sets is the **stemplot**. Stemplots display the information both graphically and numerically by using the numerical values themselves as the basis for the graph. Each data point is split between its final (rightmost) digit and all other digits to its left. These leftmost digits form the *stems* in a vertical column (starting with the smallest at the top) and the last digits are added as *leaves* on a row (left to right, starting with the smallest digit), separated from the stems by a vertical line. Here is how the healing rates appear in a stemplot:

```
1 | 1248
2 | 22336789
3 | 03455
4 | 0
```

Turned 90 degrees counterclockwise, a stemplot looks a lot like a histogram. You can interpret it the same way you would a histogram. Use this stemplot to describe the distribution of healing times.

(c) Compare your dotplot with the stemplot shown here, describing their differences and similarities.

1.38 Data maps: Obesity. The CDC studies the obesity crisis in the United States by obtaining the body mass index (BMI, from self-reported weight and height) of representative sets of individuals. A person is considered obese if his or her BMI exceeds 30. Table 1.3 reports the percent of adults in each state who were obese in 2014.[30]

(a) Define the individuals and the variable shown in Table 1.3. Is the variable quantitative or categorical?

(b) Make a dotplot or a histogram of the obesity data and describe the shape, center, and spread of the distribution. What do the data reveal about obesity in the United States?

(c) Just as data collected over time should be displayed on a time plot to look for evidence of patterns over time, so data collected geographically should be plotted on a map to look for evidence of broad geographical patterns. Figure 1.18 displays the state data on a map of the United States. What geographical pattern can you identify from this map? Would you have been able to identify this parttern from your previous graph or from the data table alone?

1.39 Stemplots: Flower length. Here are the flower lengths (in millimeters) of 16 specimens of the tropical plant *Heliconia bihai*.[31]

46.3 46.4 46.6 46.7 46.8 46.8 46.9 47.1
47.1 47.4 48.1 48.2 48.3 48.4 50.1 50.3

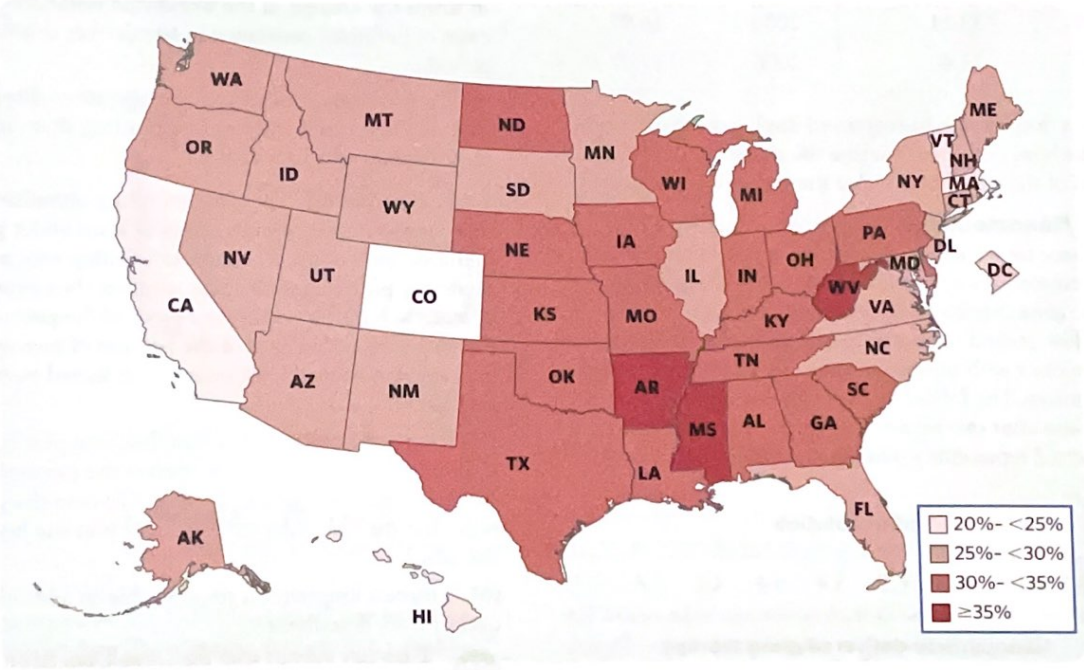
(a) Follow the model in Exercise 1.37 to create a stemplot of these flower lengths.

(b) Use your stemplot to describe the distribution of flower lengths in this tropical plant.

1.40 Fur seals on Saint George Island. Every year hundreds of thousands of northern fur seals return to their

▶ TABLE 1.3 Percent of the population who is obese, for each state and the District of Columbia

State	Percent	State	Percent	State	Percent	State	Percent
Alabama	33.5	Illinois	29.3	Montana	26.4	Rhode Island	27.0
Alaska	29.7	Indiana	32.7	Nebraska	30.2	South Carolina	32.1
Arizona	28.9	Iowa	30.9	Nevada	27.7	South Dakota	29.8
Arkansas	35.9	Kansas	31.3	New Hampshire	27.4	Tennessee	31.2
California	24.7	Kentucky	31.6	New Jersey	26.9	Texas	31.9
Colorado	21.3	Louisiana	34.9	New Mexico	28.4	Utah	25.7
Connecticut	26.3	Maine	28.2	New York	27.0	Vermont	24.8
Delaware	30.7	Maryland	29.6	North Carolina	29.7	Virginia	28.5
D.C.	21.7	Massachusetts	23.3	North Dakota	32.2	Washington	27.3
Florida	26.2	Michigan	30.7	Ohio	32.6	West Virginia	35.7
Georgia	30.5	Minnesota	27.6	Oklahoma	33.0	Wisconsin	31.2
Hawaii	22.1	Mississippi	35.5	Oregon	27.9	Wyoming	29.5
Idaho	28.9	Missouri	30.2	Pennsylvania	30.2		



▲ FIGURE 1.18 Percent of the adult population who is obese, plotted for each state on a map of the United States

haul-out territory in the Pribilof Islands in Alaska to breed, give birth, and teach their pups to swim, hunt, and survive in the Bering Sea. U.S. commercial fur sealing operations ended in 1983, but despite the reduction in harvest, the population of fur seals has continued to decline. Here are data on the number of fur seal pups born on Saint George Island (in thousands) from 1975 to 2006:[32]

Year	Pups born (thousands)	Year	Pups born (thousands)
1975	53.70	1991	24.28
1976	56.16	1992	25.16
1977	43.41	1993	23.70
1978	47.25	1994	22.24
1979	47.47	1995	24.82
1980	39.34	1996	27.39
1981	38.15	1997	24.74
1982	39.29	1998	22.09
1983	31.44	1999	21.13
1984	33.44	2000	20.18
1985	28.87	2001	18.89
1986	32.36	2002	17.59
1987	33.12	2003	17.24
1988	24.82	2004	16.88
1989	33.11	2005	16.97
1990	23.40	2006	17.07

Make a dotplot or a histogram to display the distribution of pups born per year. Describe the shape, center, and spread of the distribution. Are there any outliers?

1.41 Nanomedicine. Researchers examined a new treatment for advanced ovarian cancer in a mouse model. They created a nanoparticle-based delivery system for a suicide gene therapy to be delivered directly to the tumor cells. The grafted tumors were injected either with the new treatment or with only some buffer solution to serve as a comparison. The following data give the fold increase in tumor size after two weeks in 20 mice. A 1 represents no change; a 2 represents a doubling in volume of the tumor.[33]

Buffer solution

9.1 8.1 7.8 7.0 6.8 5.4 5.4 4.1 3.8 3.3

Nanoparticle-delivered gene therapy

4.1 3.5 2.1 2.1 1.8 1.8 1.4 1.2 1.1 1.1

(a) Make two dotplots, one for each group, using the same scale on the horizontal axis for both. Describe the distribution of tumor increase in each treatment group.

(b) Report the approximate midpoints of both groups. What are the most important differences between the two groups? What can you conclude from the study findings?

1.42 Fur seals on Saint George Island, continued. Make a time plot of the number of fur seals born per year from Exercise 1.40. What does the time plot show that your plot in Exercise 1.40 does not show? When you have data collected over time, a time plot is often needed to understand what is happening.

1.43 Herbicide resistance in weeds. Farmers use herbicides to limit the growth of weeds among their crops. Eventually, this technique places evolutionary pressure on weed species, resulting in herbicide resistance. The International Survey of Herbicide Resistant Weeds keeps track of documented cases of herbicide resistance in 80 countries worldwide. Each case is examined to identify the weed species and particular mutation allowing herbicide resistance. Table 1.4 gives the number of unique cases of herbicide resistance (corresponding to a specific weed species and a specific mutation) documented worldwide every year since 1950.[34]

(a) Make a time plot of the data. Does the time plot illustrate only year-to-year variations or are there other patterns apparent? Specifically, is there a trend over any period of years? What about cyclical fluctuation? Explain in words the change in the worldwide number of unique cases of herbicide resistance in weeds over this 63-year period.

(b) Do you think it would be appropriate to display these data in a histogram without first plotting them in a time plot? Explain why or why not.

1.44 Exercising. The Gallup polling organization interviews samples of individuals to learn about public opinions and habits. One question Gallup asks regularly of randomly picked Americans is whether they exercised for at least 30 minutes on three or more of the past seven days. Figure 1.19 is a time plot of the percent of respondents who say that they did, for surveys conducted between 2008 and 2015.[35]

(a) The most notable pattern in this time plot is yearly cycles. In which season of the year is the percent who exercise regularly highest? Lowest? (To read the graph, note that the tick mark for each year is at the beginning of the year.)

(b) Is there a longer-term trend visible in addition to the cycles? If so, describe it.

LARGE DATA SET *Exercises marked with the* Large Data Set *icon guide you through the comprehensive analysis of more complex data sets characterized by a large number of observations, a number of variables, or both. You can find many more such exercises in the book's review chapters (Chapters 8, 16, and 25). Short answers are not available for these exercises to enable their use for class work or assignments.*

LARGE DATA SET **1.45 Everglades.** Everglades National Park is the largest subtropical wilderness in the United States and has been designated a Wetland of International Importance. This important ecosystem is