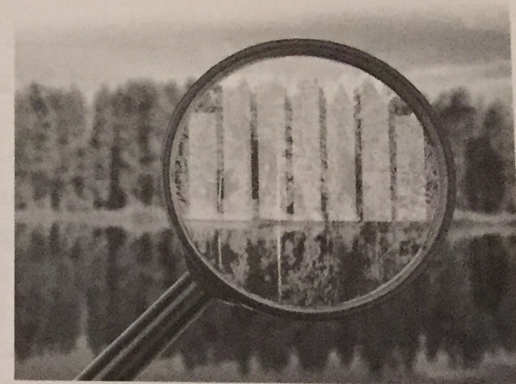
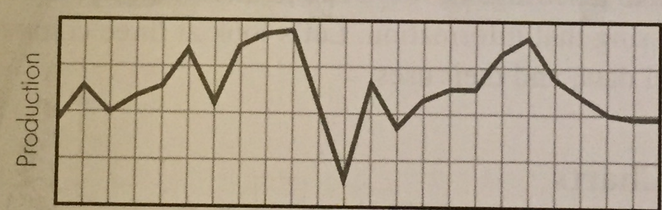


CHAPTER 9

Plots, Graphs, and Pictures

Thought Questions

1. You have seen pie charts and bar graphs and should have some rudimentary idea of how to construct them. Suppose you have been keeping track of your living expenses and find that you spend 50% of your money on rent, 25% on food, and 25% on other expenses. Draw a pie chart and a bar graph to depict this information. Discuss which is more visually appealing and useful.
2. Here is an example of a plot that has some problems. Give two reasons why this is not a good plot.



3. Suppose you had a set of data representing two measurement variables—namely, height and weight—for each of 100 people. How could you put that information into a plot, graph, or picture that illustrated the relationship between the two measurements for each person?
4. Suppose you own a company that produces candy bars, and you want to display two graphs. One graph is for customers and shows the price of a candy bar for each of the past 10 years. The other graph is for stockholders and shows the amount the company was worth for each of the past 10 years. You decide to adjust the dollar amounts in one graph for inflation but to use the actual dollar amounts in the other graph. If you were trying to present the most favorable story in each case, which graph would be adjusted for inflation? Explain.
5. What do you think is meant by the term *time series*?

9.1 Well-Designed Statistical Pictures

There are many ways to present data in pictures. The most common are plots and graphs, but sometimes a unique picture is used to fit a particular situation. The purpose of a plot, graph, or picture of data is to give you a visual summary that is more informative than simply looking at a collection of numbers. Done well, a picture can quickly convey a message that would take you longer to find if you had to study the data on your own. Done poorly, a picture can mislead all but the most observant of readers. Here are some basic characteristics that all plots, graphs, and pictures should exhibit:

1. The data should stand out clearly from the background.
2. There should be clear labeling that indicates
   a. the title or purpose of the picture.
   b. what each of the axes, bars, pie segments, and so on, denotes.
   c. the scale of each axis, including starting points.
3. A source should be given for the data.
4. There should be as little "chart junk"—that is, extraneous material—in the picture as possible.

9.2 Pictures of Categorical Data

Categorical data are easy to represent with pictures. The most frequent use of such data is to determine how the whole divides into categories, and pictures are useful in expressing that information. Let's look at three common types of pictures for categorical data, and their uses.

Pie Charts

Pie charts are useful when only one categorical variable is measured. Pie charts show what percentage of the whole falls into each category. They are simple to understand, and they convey information about the relative size of groups more readily than a table. Side-by-side pie charts can be used to compare groups for a categorical variable, while still illustrating how the pie divides up for each group separately. Figure 9.1 shows pie charts representing the percentage of Caucasian American boys and girls who have various eye colors. Note that for both sexes, roughly one-third have blue eyes and another one-third have brown eyes. You can also see that there are slight differences in the sexes, with more boys having blue eyes and more girls having green eyes.

Bar Graphs

Bar graphs also show percentages or frequencies in various categories, but they can be used to represent two or three categorical variables simultaneously. One categorical

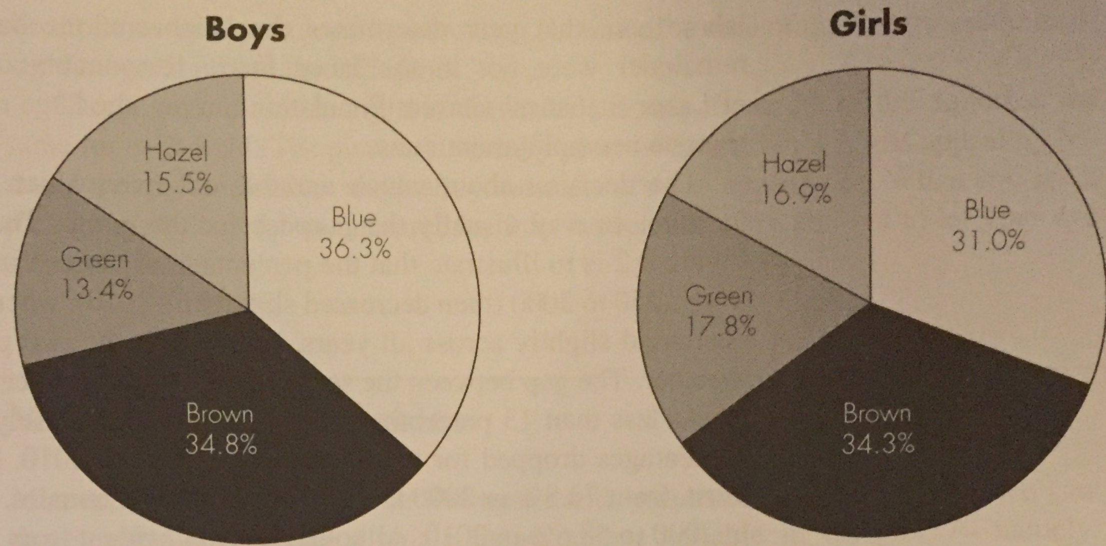
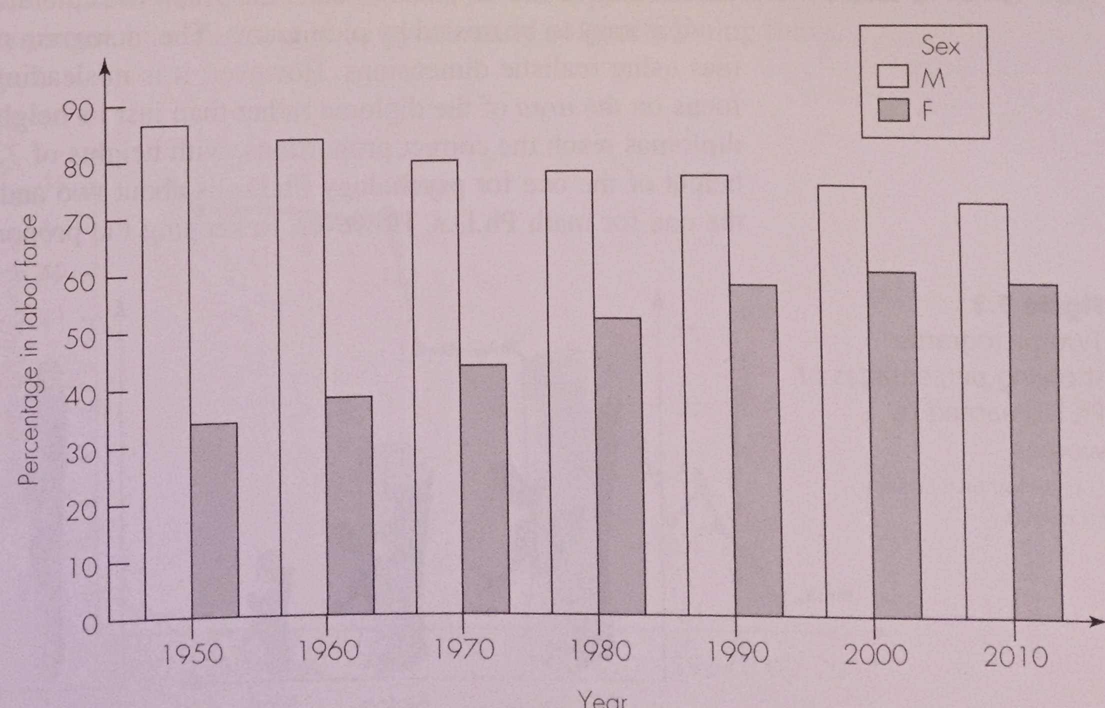


Figure 9.1
Pie charts of eye colors of Caucasian American boys and girls
Source: Malloy, 2008

variable is used to label the horizontal axis. Within each of the categories along that axis, a bar is drawn to represent each category of the second variable. Frequencies or percentages are shown on the vertical axis. A third variable can be included if the graph has only two categories by using percentages on the vertical axis. One category is shown, and the other is implied by the fact that the total must be 100%.

For example, Figure 9.2 illustrates employment trends for men and women across decades. The year in which the information was collected is one categorical variable, represented by the horizontal axis. In each year, people were categorized according to two additional variables: whether they were in the labor force and whether they were male or female. Separate bars are drawn for males and females, and the percentage of all adults of that sex who were in the

Figure 9.2
Percentage of males and females 16 and over in the labor force
Source: Based on data from U.S. Dept. of Labor, Bureau of Labor Statistics, *Current Population Survey.*



labor force that year determines the heights of the bars. It is implicit that the remainder were not in the labor force. Respondents were part of the Bureau of Labor Statistics' Current Population Survey, the large monthly survey used to determine unemployment rates.

The decision about which variable occupies which position should be made to better convey visually the purpose for the graph. The purpose of the graph in Figure 9.2 is to illustrate that the percentage of women in the labor force increased from 1950 to 2000 (then decreased slightly by 2010), whereas the percentage of men decreased slightly across all years, resulting in the two percentages coming closer together. The gap between the sexes in 1950 was 53 percentage points, but by 2010, it was less than 13 percentage points. If you look closely, you will notice that the percentages dropped for both sexes from 2000 to 2010. For males, it dropped the most, from 74.8% in 2000 to 71.2% in 2010. For females, the drop was from 59.9% in 2000 to 58.6% in 2010. Although it is not evident from the graph, which includes only years divisible by 10, the percentage of women in the labor force peaked at 60% in 1999.

Bar graphs are not always as visually appealing as pie charts, but they are much more versatile. They can also be used to represent actual frequencies instead of percentages and to represent proportions that are not required to sum to 100%.

Pictograms

A **pictogram** is like a bar graph except that it uses pictures related to the topic of the graph. Figure 9.3 shows a pictogram illustrating the proportion of Ph.D.s earned by women in the United States in 2011 in three fields—psychology (72%), biological sciences (52%), and mathematics (29%)—as reported by the National Science Foundation. Notice that in place of bars, the graph uses pictures of diplomas.

It is easy to be misled by pictograms. The pictogram on the left shows the diplomas using realistic dimensions. However, it is misleading because the eye tends to focus on the *area* of the diploma rather than just its height. The heights of the three diplomas reach the correct proportions, with heights of 72%, 52%, and 29%, so the height of the one for psychology Ph.D.s is about two and a half times the height of the one for math Ph.D.s. However, in keeping the proportions realistic, the area of

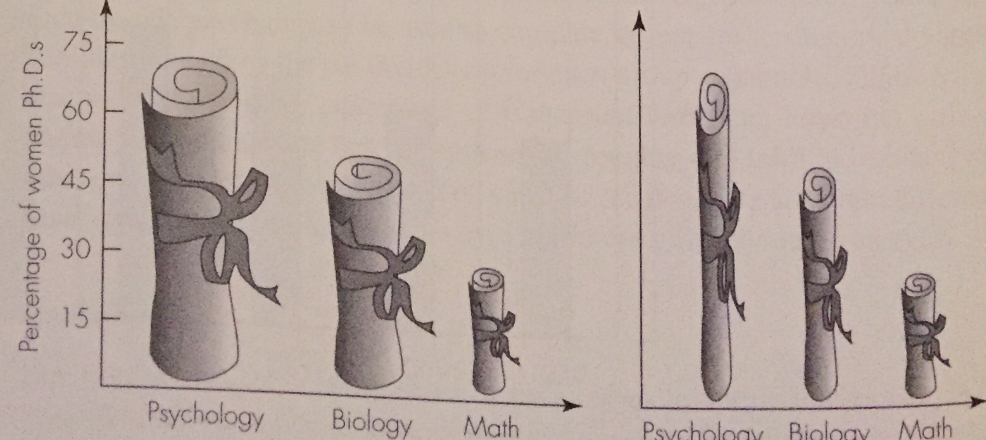


Figure 9.3
Two pictograms showing percentages of Ph.D.s earned by women
Source: National Science Foundation

the diploma for psychology is more than six times the area of the one for math, leading the eye to inflate the difference.

The pictogram on the right is drawn by keeping the width of the diplomas the same for each field. The picture is visually more accurate, but it is less appealing because the diplomas are consequently quite distorted in appearance. When you see a pictogram, be careful to interpret the information correctly and not to let your eye mislead you.

9.3 Pictures of Measurement Variables

Measurement variables can be illustrated with graphs in numerous ways. We saw three ways to illustrate a single measurement variable in Chapter 7—namely, stemplots, histograms, and boxplots. Graphs are most useful for displaying the relationship between two measurement variables or for displaying how a measurement variable changes over time. Two common types of displays for measurement variables are illustrated in Figures 9.4 and 9.5.

Line Graphs

Figure 9.4 is an example of a **line graph** displayed over time. It shows the winning times for the men's 500-meter speed skating event in the Winter Olympics from 1924 to 2010. Notice the distinct downward trend, with only a few upturns over the years (including 2006 and 2010). There was a large drop between 1952 and 1956, followed by a period of relative stability. Also notice that no Olympic games were held in 1944 because of World War II. These patterns are much easier to detect with a picture than they would be by scanning a list of winning times.

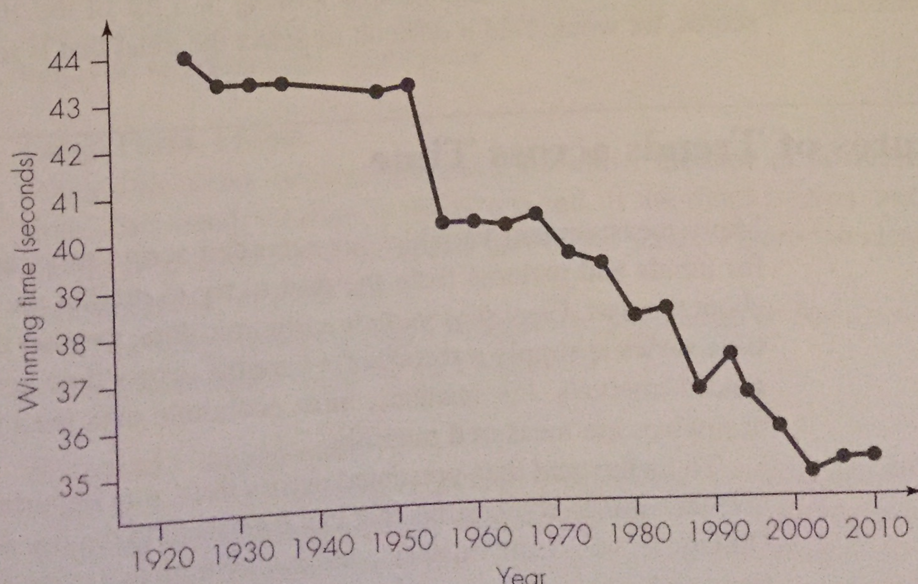


Figure 9.4
Line graph displaying winning time versus year for men's 500-meter Olympic speed skating
Source: http://www.olympic.org/speed-skating-500m-men

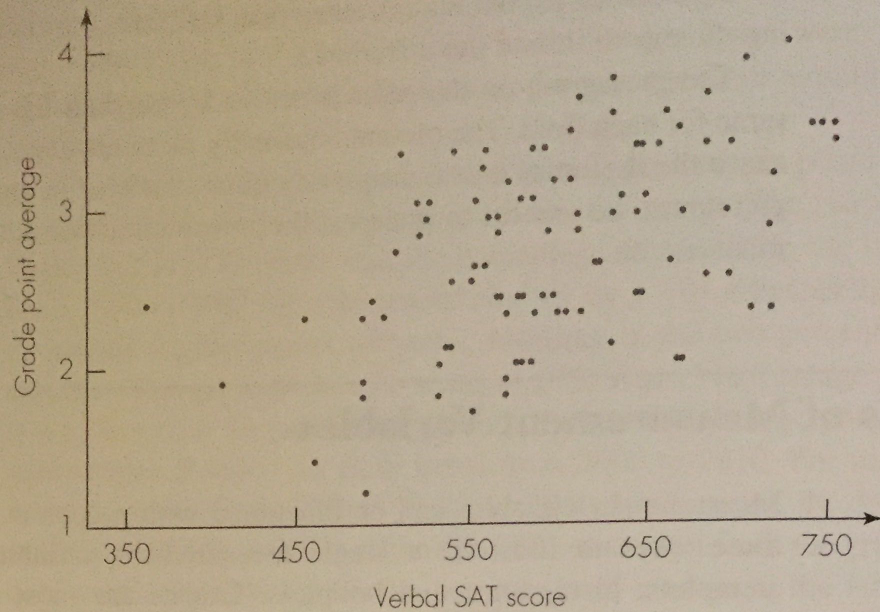


Figure 9.5
Scatterplot of grade point average versus verbal SAT score
Source: Ryan, Joiner, and Ryan, 1985, pp. 309–312.

Scatterplots

Figure 9.5 is an example of a **scatterplot.** Scatterplots are useful for displaying the relationship between two measurement variables. Each dot on the plot represents one individual, unless two or more individuals have the same data, in which case only one point is plotted at that location. (An exception is that some software programs replace the dot with the number of points it represents, if it's more than one.) The plot in Figure 9.5 shows the grade point averages (GPAs) and verbal scholastic achievement test (SAT) scores for a sample of 100 students at a university in the northeastern United States.

Although a scatterplot can be more difficult to read than a line graph, it displays more information. It shows outliers, as well as the degree of variability that exists for one variable at each location of the other variable. In Figure 9.5, we can see an increasing trend toward higher GPAs with higher SAT scores, but we can also still see substantial variability in GPAs at each level of verbal SAT scores. A scatterplot is definitely more useful than the raw data. Simply looking at a list of the 100 pairs of GPAs and SAT scores, we would find it difficult to detect the trend that is so obvious in the scatterplot.

9.4 Pictures of Trends across Time

Many measurement variables are recorded across time, and it is of interest to look for trends and patterns from the past to try to predict the future, or simply to learn about the past. Examples include economic data, weather data, and demographics. A **time series** is simply a record of a variable across time, usually measured at equally spaced intervals. For instance, most economic data used by both governments and businesses are measured monthly.

To understand data presented across time, it is important to know how to recognize the various components that can contribute to the ups and downs in a time series. Otherwise, you could mistake a temporary high in a cycle for a permanent increasing trend and make a very unwise economic decision.

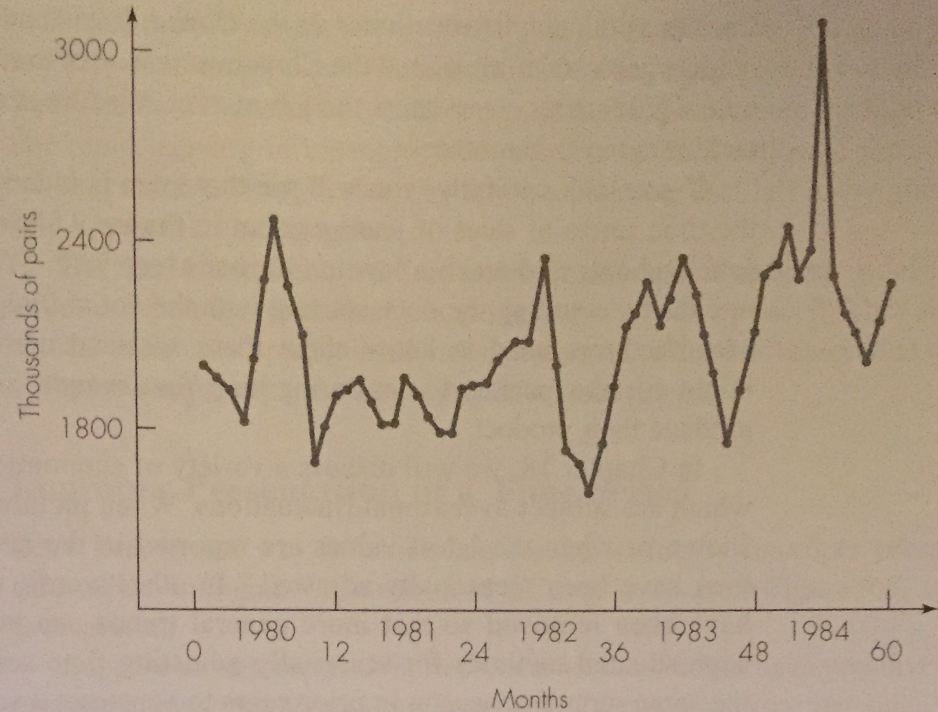


Figure 9.6
An example of a time series plot: Jeans sales in the United Kingdom from 1980 to 1984
Source: Hand et al., 1994, p. 314.

A Time Series Plot

Figure 9.6 illustrates a **time series plot.** The data represent monthly sales of jeans in Britain for the 5-year period from January 1980 to December 1984. Notice that the data points have been connected to make it easier to follow the ups and downs across time. Data are measured in thousands of pairs sold. Month 1 is January 1980, and month 60 is December 1984.

Components of Time Series

Most time series have the same four basic components: *long-term trend, seasonal components, irregular cycles,* and *random fluctuations.* We will examine these in more detail in Chapters 10 and 18, but the basic ideas are discussed here and will help you interpret time series plots.

Long-Term Trend

Many time series measure variables that either increase or decrease steadily across time. This steady change is called a **trend.** If the trend is even moderately large, it should be obvious by looking at a plot of the series. Figure 9.6 clearly shows an increasing trend for jeans sales.

If the long-term trend has a straight line pattern (up or down), we can investigate it using techniques we cover in Chapter 10.

Seasonal Components

Most time series involving economic data or data related to people's behavior have **seasonal components**. In other words, they tend to be high in certain months or seasons and low in others every year. For example, new housing starts are much higher in warmer months. Sales of toys and other standard gifts are much higher just before

Christmas. Unemployment rates in the United States tend to rise in January—when outdoor jobs are minimal and the Christmas season is over—and again in June, when a new graduating class enters the job market. Weddings are more likely to occur in certain months than others.

If you look carefully, you will see that there is indeed a seasonal component to the time series of sales of jeans evident in Figure 9.6. Sales appear to peak during June and July and reach a low in October every year. (You can identify individual months by counting the dots, starting with the dot at 0 representing January, 1980.) Manufacturers need to know about these seasonal components. Otherwise, they might mistake increased sales during June, for example, as a general trend and over-produce their product.

In Chapter 18, we will discuss a variety of economic indicators, almost all of which are subject to seasonal fluctuations. When pictures of these time series are shown or when the latest values are reported in the news, it will be stated that they have been "seasonally adjusted." In other words, the seasonal components have been removed so that more general trends can be seen. Economists have sophisticated methods for seasonally adjusting time series. They use data from the same month or season in prior years to construct a seasonal factor, which is a number either greater than one or less than one by which the current figures are multiplied.

Irregular Cycles and Random Fluctuations

There are two remaining components of time series: the irregular (but smooth) **cycles** that economic systems tend to follow and unexplainable random fluctuations. It is often hard to distinguish between these two components, especially if the cycles are not regular.

Figure 9.7 shows the U.S. unemployment rate, seasonally adjusted, monthly from January, 1990, through May, 2013. Notice the definite irregular cycles during which unemployment rates rise and fall over a number of years. Some of these can

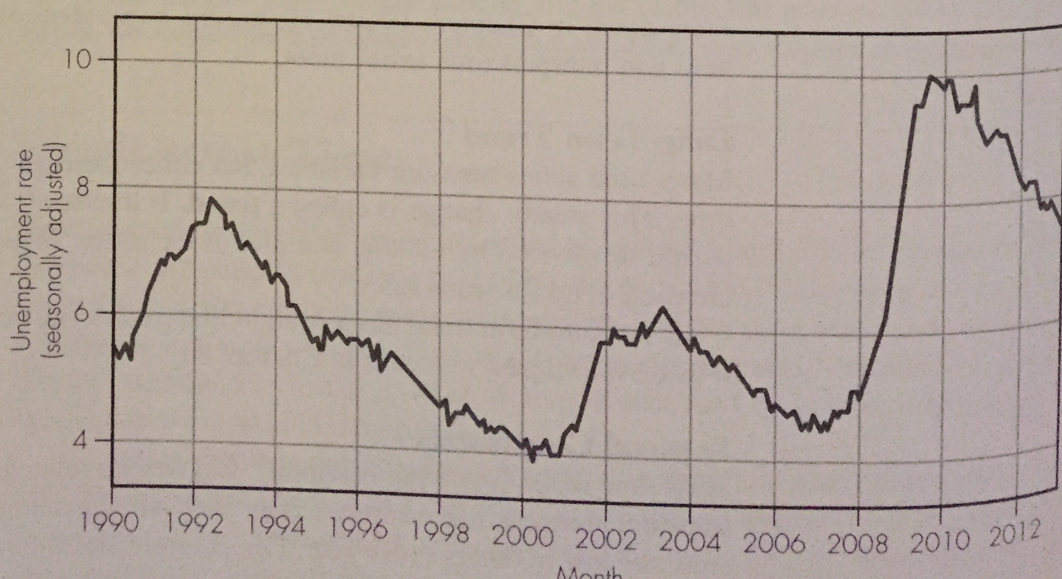


Figure 9.7
Seasonally adjusted monthly unemployment rate in the United States from January 1990 to May 2013
Source: http://data.bls.gov/pdq/SurveyOutputServlet

be at least partially explained by social and political factors. For example, the steady decrease in unemployment during the 1990s has been partially attributed to President Clinton's economic policies and partially to the boom in the high tech industry. The rapid increase in unemployment starting in 2008 followed the collapse of the housing market, brought about by lending practices that left many homeowners unable to pay their mortgages.

The final component in a time series, **random fluctuations**, is defined as what's left over when the other three components have been removed. They are part of the natural variability present in all measurements. It is for this reason that the future can never be completely predicted using data from the past.

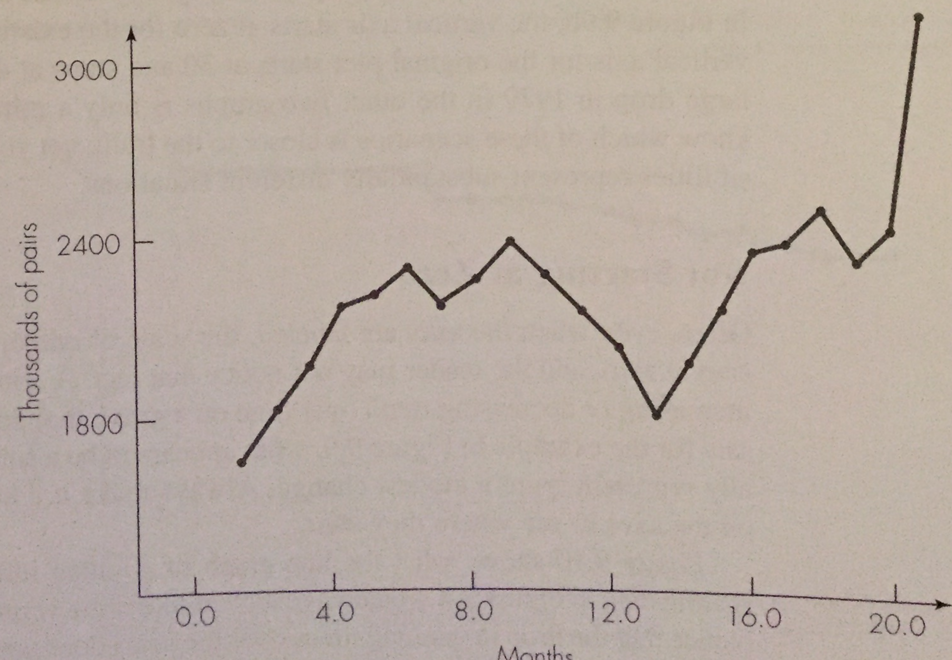
Improper Presentation of a Time Series

Let's look at one way in which you can be fooled by improper presentation of a time series. In Figure 9.8, a subset of the time series from Figure 9.6, sales of jeans, is displayed.

Suppose an unscrupulous entrepreneur was anxious to have you invest your hard-earned savings into his clothing company. To convince you that sales of jeans can only go up, he presents you with a limited set of data—from October 1982 to June 1984. With only those few months shown, it appears that the basic trend is way up!

A less obvious version of this trick is to present data up to the present time but to start the plot of the series at an advantageous point in the past, rather than providing all of the available data. Be suspicious of time series showing returns on investments that look too good to be true. They probably are. Notice when the time series begins and compare that with your knowledge of recent economic cycles.

Figure 9.8
Distortion caused by displaying only part of a time series: Jeans sales for 21 months



9.5 Difficulties and Disasters in Plots, Graphs, and Pictures

A number of common mistakes appear in plots and graphs that may mislead readers. If you are aware of them and watch for them, you will substantially reduce your chances of misreading a statistical picture.

> *The most common problems in plots, graphs, and pictures are*
>
> **1.** No labeling on one or more axes
> **2.** Not starting at zero as a way to exaggerate trends
> **3.** Distorting time series plots
> **4.** Change(s) in labeling on one or more axes
> **5.** Misleading units of measurement
> **6.** Using poor information

No Labeling on One or More Axes

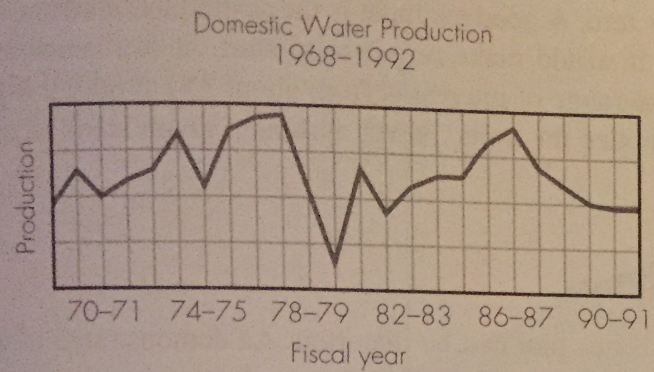
You should always look at the axes in a picture to make sure they are labeled. Figure 9.9a gives an example of a plot for which the units were *not* labeled on the vertical axis. The plot appeared in a newspaper insert titled, "May 1993: Water awareness month." When there is no information about the units used on one of the axes, the plot cannot be interpreted. To see this, consider Figure 9.9b and c, displaying two different scenarios that could have produced the actual graph in Figure 9.9a. In Figure 9.9b, the vertical axis starts at zero for the existing plot. In Figure 9.9c, the vertical axis for the original plot starts at 30 and stops at 40, so what appears to be a large drop in 1979 in the other two graphs is only a minor fluctuation. We do not know which of these scenarios is closer to the truth, yet you can see that the two possibilities represent substantially different situations.

Not Starting at Zero

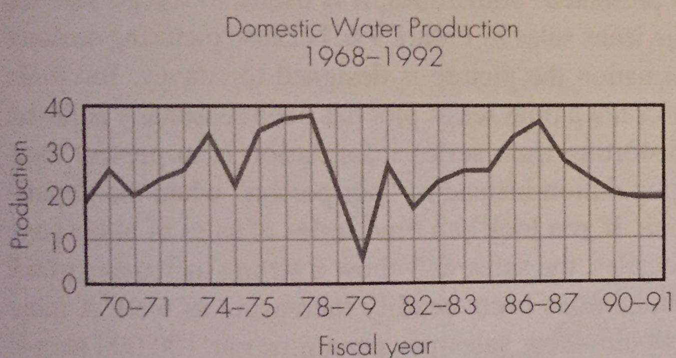
Often, even when the axes are labeled, the scale of one or both of the axes does not start at zero, and the reader may not notice that fact. A common ploy is to present an increasing or decreasing trend over time on a graph that does not start at zero. As we saw for the example in Figure 9.9, what appears to be a substantial change may actually represent quite a modest change. Always make it a habit to check the numbers on the axes to see where they start.

Figure 9.10 shows what the line graph of winning times for the Olympic speed skating data in Figure 9.4 would have looked like if the vertical axis had started at zero. Notice that the drop in winning times over the years does not look nearly as dramatic as it did in Figure 9.4. Be very careful about this form of potential deception if someone is

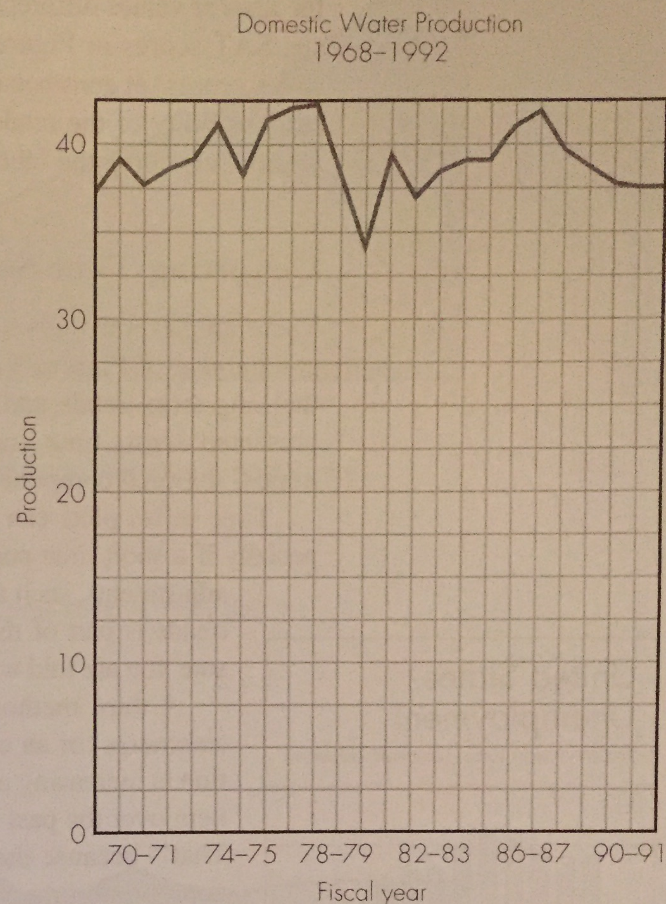
Figure 9.9
Example of a graph with no labeling (a) and possible interpretations (b and c)
Source: Insert in the *California Aggie* (UC Davis), 30 May 1993.



(a) Actual graph

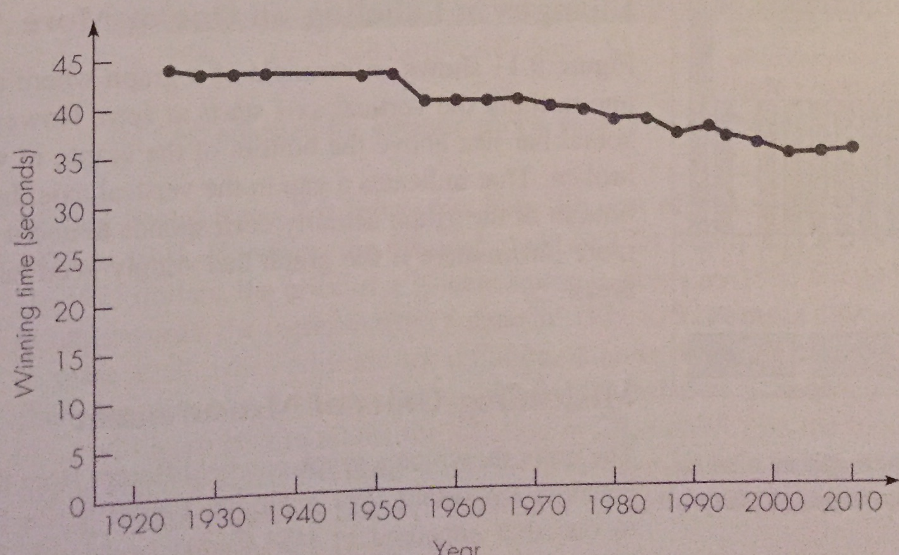


(b) Axis in "actual graph" starts at zero



(c) Axis in "actual graph" does not start at zero as a way to exaggerate trends

Figure 9.10
An example of the change in perception when axes start at zero, showing speed-skating times from Figure 9.4



presenting a graph to display growth in sales of a product, a drop in interest rates, and so on. Be sure to look at the labeling, especially on the vertical axis.

Despite this, be aware that for some graphs it makes sense to start the units on the axes at values different from zero. A good example is the scatterplot of GPAs versus SAT scores in Figure 9.5. It would make no sense to start the horizontal axis (SAT scores) at zero because the range of interest is from about 350 to 800. It is the responsibility of the reader to notice the units. Never assume a graph starts at zero without checking the labeling.

Distorting Time Series Plots

There are several ways a time series plot can be distorted. As demonstrated in the comparison of Figures 9.6 and 9.8, a partial time series can be presented to mask the long-term trends and cycles present in the full series. When you look at data presented across time, make sure the time period is long enough to include likely trends, seasonal components, and irregular cycles.

Time series plots can also be misleading if they are not seasonally adjusted, especially if a short time period is presented. Sometimes, it is useful to forego seasonal adjustments, such as in the jeans sales in Figure 9.6, because picturing seasonal trends is part of the information the picture is designed to convey. But make sure you are told whether or not a time series plot has been seasonally adjusted.

A third method of distorting time series information is to present measurements for an entire population instead of "per person" when the population is increasing over time. If you look at a time series of sales of almost any item over the past few decades, the sales will show a strong increasing trend. That's because the population has been steadily increasing as well. A more appropriate time series would show sales per person or per 100,000 people instead. We will investigate this issue further in Chapter 11.

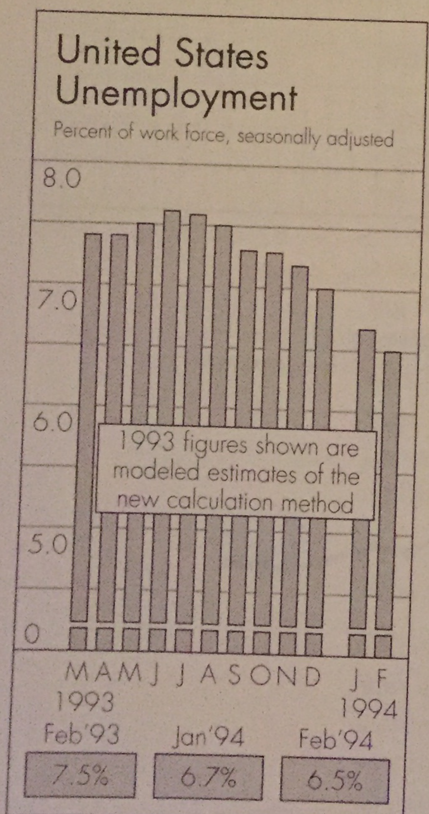


Figure 9.11
A bar graph with gap in labeling
Source: Davis (CA) Enterprise, 4 March 1994, p. A-7.

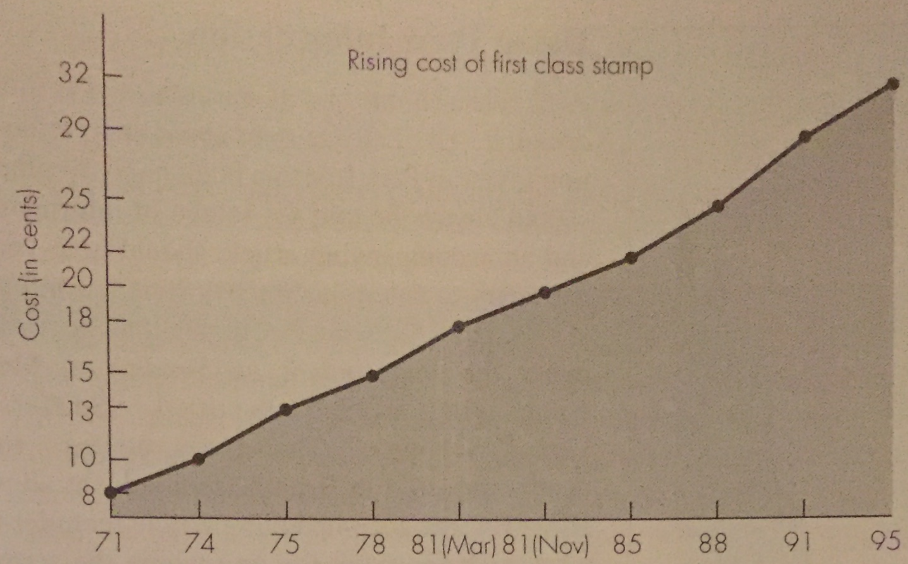
Changes in Labeling on One or More Axes

Figure 9.11 shows an example of a graph where a cursory look would lead one to think the vertical axis starts at zero. However, notice the white horizontal bar just above the bottom of the graph, in which the vertical bars are broken. That indicates a gap in the vertical axis. In fact, you can see that the bottom of the graph actually corresponds to about 4.0%. It would have been more informative if the graph had simply been labeled as such, without the break.

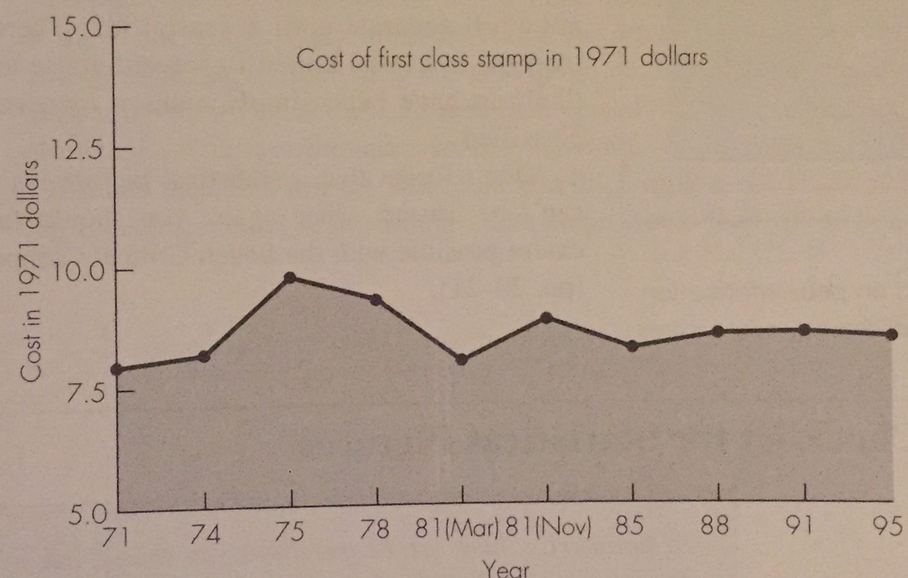
Misleading Units of Measurement

The units shown on a graph can be different from those that the reader would consider important. For example, Figure 9.12a illustrates a graph similar to one that appeared in *USA Today* on March 7, 1994 with the heading:

Figure 9.12
Cost of a first class stamp over time, without and with adjustment for inflation
Source: USA Today, 7 March 1994, p. 13A.



(a) Misleading graph showing steep rise in postal rates



(b) Cost adjusted for inflation shows little or no increase

"Rising Postal Rates." It represents how the cost of a first-class stamp rose from 1971 to 1995. However, the original graph in *USA Today* had a footnote that read "In 1971 dollars, the price of a 32-cent stamp in February 1995 would be 8.4 cents." In other words, the increase from 8 cents in 1971 to 32 cents in 1995 was very little increase at all, when adjusted for inflation. A more truthful picture would show the changing price of a first-class stamp adjusted for inflation, as shown in Figure 9.12b. (You will learn how to adjust for inflation in Chapter 18.) As the footnote implied, such a graph would show little or no rise in postal rates as a function of the worth of a dollar. Both Figures 9.12a and 9.12b have an additional problem, which you will be asked to identify in Exercise 6.

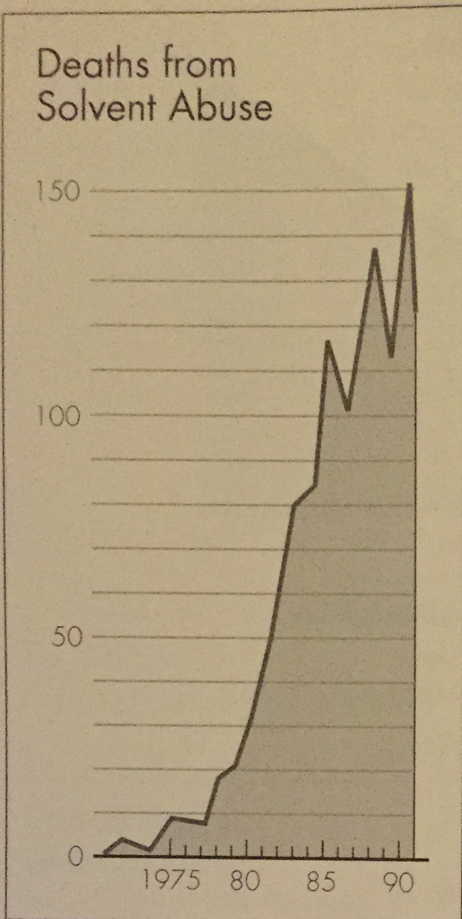


Figure 9.13
A graph based on poor information
Source: The Independent on Sunday (London), 13 March 1994.

Using Poor Information

A picture can only be as accurate as the information that was used to design it. All of the cautions about interpreting the collection of information given in Part 1 of this book apply to graphs and plots as well. You should always be told the source of information presented in a picture, and an accompanying article should give you as much information as necessary to determine the worth of that information.

Figure 9.13 shows a graph that appeared in the London newspaper the *Independent on Sunday* on March 13, 1994. The accompanying article was titled, "Sniffers Quit Glue for More Lethal Solvents." The graph appears to show that very few deaths occurred in Britain from solvent abuse before the late 1970s. However, the accompanying article includes the following quote from a research fellow at the unit where the statistics are kept: "It's only since we have started collecting accurate data since 1982 that we have begun to discover the real scale of the problem" (p. 5). In other words, the article indicates that the information used to create the graph is not at all accurate until at least 1982. Therefore, the apparent sharp increase in deaths linked to solvent abuse around that time period is likely to have been simply a sharp increase in deaths reported and classified.

Don't forget that a statistical picture isn't worth much if the data can't be trusted. Once again, you should familiarize yourself to the extent possible with the Seven Critical Components listed in Chapter 2 (pp. 20–21).

9.6 A Checklist for Statistical Pictures

To summarize, here are 12 questions you should ask when you look at a statistical picture—before you even begin to try to interpret the data displayed.

1. Does the message of interest stand out clearly?
2. Is the purpose or title of the picture evident?
3. Is a source given for the data, either with the picture or in an accompanying article?
4. Did the information in the picture come from a reliable, believable source?
5. Is everything clearly labeled, leaving no ambiguity?
6. Do the axes start at zero or not?
7. For time series data, is a long enough time period shown?
8. Can any observed trends be explained by another variable, such as increasing population?
9. Do the axes maintain a constant scale?

10. Are there any breaks in the numbers on the axes that may be easy to miss?
11. For financial data, have the numbers been adjusted for inflation and/or seasonally adjusted?
12. Is there information cluttering the picture or misleading the eye?

CASE STUDY 9.1

Time to Panic about Illicit Drug Use?

The graph illustrated in Figure 9.14 appeared on the website for the U.S. Department of Justice, Drug Enforcement Agency, in Spring, 1998 (http://www.usdoj.gov/dea/drugdata/cp-316.htm). The headline over the graph reads "Emergency Situation among Our Youth." Look quickly at the graph, and describe what you see. Did it lead you to believe that almost 80% of 8th-graders used illicit drugs in 1996, compared with only about 10% in 1992? The graph is constructed so that you might easily draw that conclusion. Notice that careful reading indicates otherwise, and crucial information is missing. The graph tells us only that in 1996 the rate of use was 80% higher, or 1.8 times what it was in 1991. The actual rate of use is *not* provided at all in the graph. Only after searching the remainder of the website does that emerge. The rate of illicit drug use among 8th-graders in 1991 was about 11%, and thus, in 1996, it was about 1.8 times that, or about 19.8%. Additional information elsewhere on the website indicates that about 8% of 8th-graders used marijuana in 1991, and thus, this was the most common illicit drug used. These are still disturbing statistics, but not as disturbing as the graph would lead you to believe. ■

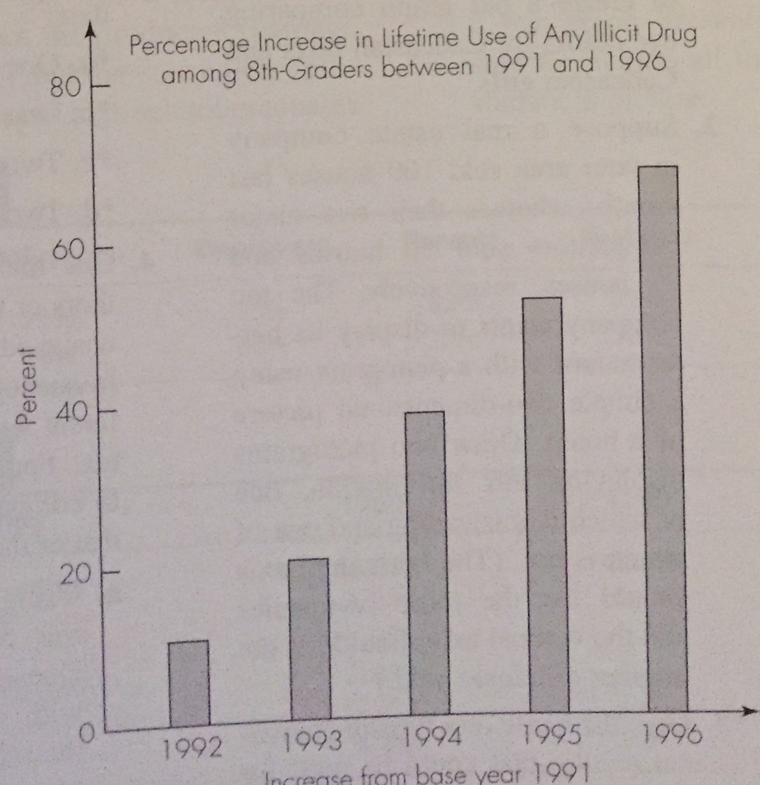


Figure 9.14
Emergency situation among our youth: 8th-grade drug use
Source: U.S. Dept. of Justice.

Thinking About Key Concepts

- *Pie charts* and *bar graphs* are useful for picturing categorical data, with the latter more useful for comparing a categorical variable across groups.
- There are numerous ways that pictures of data can convey misleading messages. Make sure you carefully examine graphs and plots and don't be fooled by distorted visual displays of information.
- *Scatter plots* are useful for displaying the relationship between two measurement variables and can reveal outliers that are not apparent from pictures of each variable on its own.
- A perfectly good graph can convey a misleading message if the information used to create it is not thoughtfully presented. For instance, prices across time should be adjusted for inflation, and measurements for groups of individuals across time should be presented as numbers per capita instead of as totals for the whole group.
- *Time series* plots present measurements across regularly spaced intervals. They can show trends, seasonal patterns, and cycles, but can also be distorted to convey a misleading message—for instance, by showing a short time period.

Exercises

Exercises with numbers divisible by 3 (3, 6, 9, etc.) are included in the Solutions at the back of the book. They are marked with an asterisk (*).

1. Use the pie charts in Figure 9.1 to create a bar graph comparing eye colors for Caucasian boys and Caucasian girls.

2. Suppose a real estate company in your area sold 100 houses last month, whereas their two major competitors sold 50 houses and 25 houses, respectively. The top company wants to display its better record with a pictogram using a simple two-dimensional picture of a house. Draw two pictograms displaying this information, one of which is misleading and one of which is not. (The horizontal axis should list the three companies and the vertical axis should list the number of houses sold.)

***3.** Give the name of a type of statistical picture that could be used for each of the following kinds of data:

***a.** One categorical variable

***b.** One measurement variable

***c.** Two categorical variables

***d.** Two measurement variables

4. One method used to compare authors or to determine authorship on unsigned writing is to look at the frequency with which words of different lengths appear in a piece of text. For this exercise, you are going to compare your own writing with that of the author of this book.

a. Using the first full paragraph of this chapter (not the Thought Questions), create a pie chart with three segments, showing the relative frequency of words of 1 to 3 letters, 4 to 5 letters,