# Seeing Through Statistics

**JESSICA M. UTTS**

University of California, Davis

*An Alexander Kugushev Book*

**DUXBURY PRESS**
**AN IMPRINT OF WADSWORTH PUBLISHING COMPANY**

I(T)P™ AN INTERNATIONAL THOMSON PUBLISHING COMPANY

Belmont Albany Bonn Boston Cincinnati Detroit London Madrid Melbourne
Mexico City New York Paris San Francisco Singapore Tokyo Toronto Washington

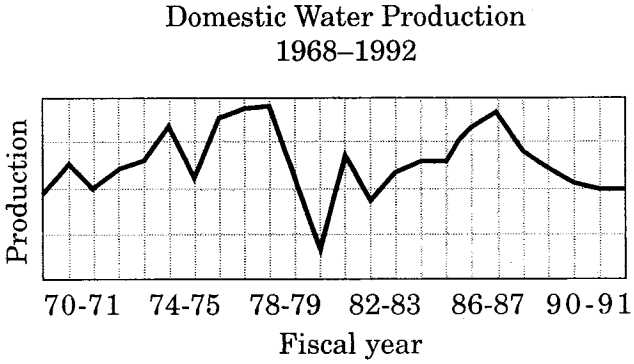
CHAPTER 9

# Plots, Graphs and Pictures

## Thought Questions

**1.** You have certainly seen pie charts and bar graphs before 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 for each of 100 people, namely their height and their weight. How could you put that information into a plot, graph or picture that retained the information for each person?

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 bunch of numbers. Done well, a picture can convey a message that would take awhile for you to construct by studying the data on your own. Done poorly, a picture can mislead all but the most observant of readers.

Here are some basic principles that all plots, graphs and pictures should exhibit:

1. The data should stand out clearly from the background.
2. There should be clear labeling indicating:
   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," i.e. extraneous material, in the picture as possible.

9.2

## Pictures of Categorical Data

Categorical data is 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

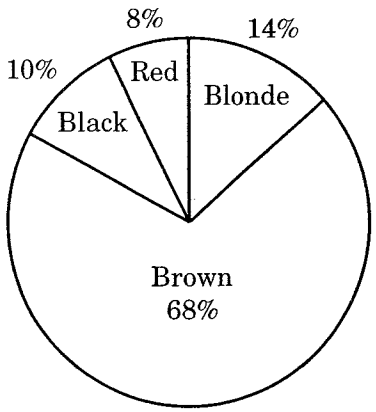
**Pie charts** are useful when only one categorical variable is measured. They show what percentage of the whole falls into each category. They are simple to understand and convey information about relative size of groups more readily than a table. Figure 9.1 shows a pie chart that represents the percentage of Caucasian-American children who have various hair colors.

### Bar Graphs

**Bar graphs** also show percentages or frequencies in various categories, but they can be used to represent two categorical variables simultaneously. One categorical variable can be used to label the horizontal axis and the other to label each of the bars in a group.

**FIGURE 9.1**

Pie Chart of Hair Colors of Caucasian-American Children



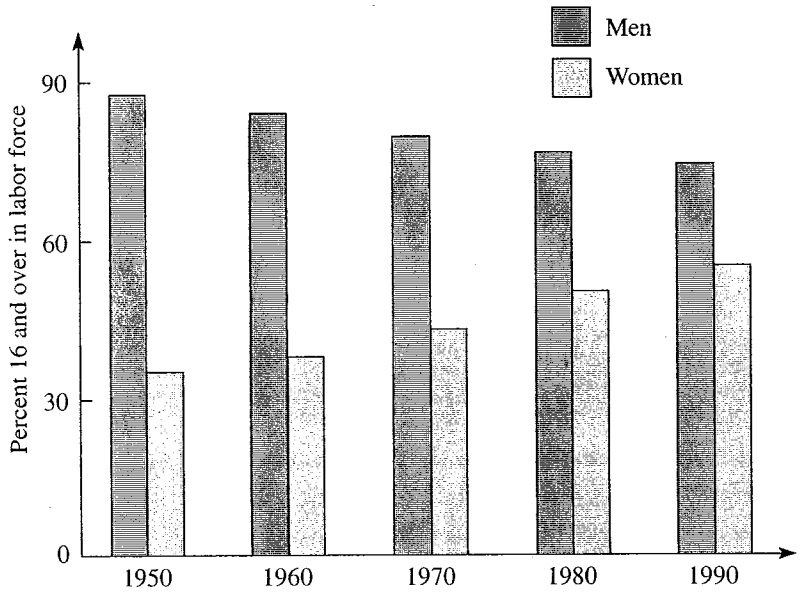

SOURCE: *What the Odds Are*, Les Krantz, 1992, New York: Harper Perennial, p. 118.

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

An example of a bar graph is shown in Figure 9.2. In each year, people were categorized according to two variables—whether or not they were in the labor force and whether they were male or female. Respondents were part of the Bureau of Labor Statistics' Current Population Survey, the large monthly survey used to determine unemployment rates.

**FIGURE 9.2**

Percentage of Men and Women in the Labor Force




SOURCE: Bureau of Labor Statistics.

Notice that the totals are not expected to add to 100%. The purpose of this graph is to illustrate that the percentage of women in the labor force has increased since 1950 while the percentage of men has decreased slightly. The gap in 1950 was 53 percentage points, but by 1990 it was only 19 percentage points.

## Pictograms

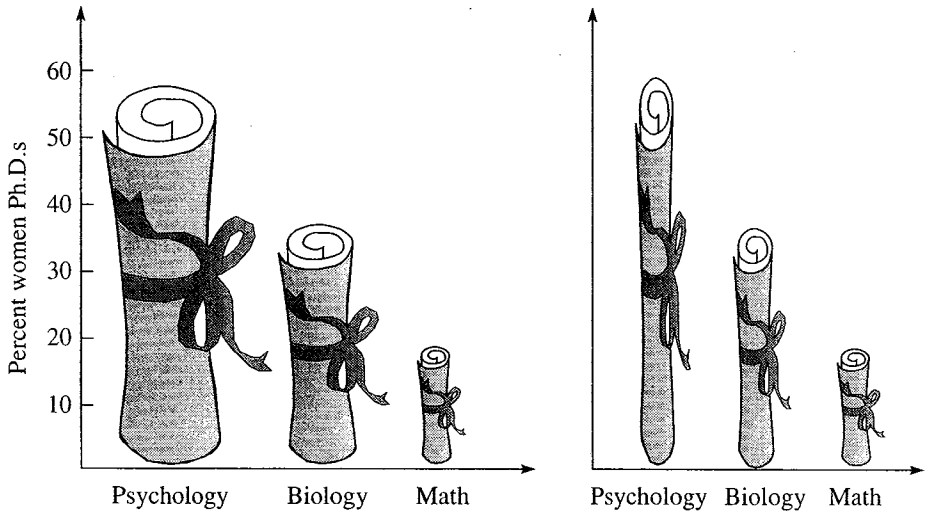
A **pictogram** is like a bar graph except that it uses pictures related to the topic of the graph. Figure 9.3 shows a pictogram illustrating the proportion of Ph.D.s earned by women in three fields, psychology (58%), biology (37%) and mathematics (18%), as reported in *Science* (Vol. 260, April 16, 1993, pg. 409). 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 58%, 37%, and 18%, so for instance the height of the one for psychology Ph.D.s is just over three times the height of the one for math Ph.D.s. But in keeping the proportions realistic, the area of the diploma for psychology is about nine 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 since the diplomas are consequently quite distorted in appearance. When you see

**FIGURE 9.3**

Two Pictograms Showing Percentages of Ph.D.s Earned by Women




SOURCE: *Science*, Vol. 260, April 16, 1993, p. 409.

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 two ways to illustrate a single measurement variable in Chapter 7, namely stemplots and histograms.

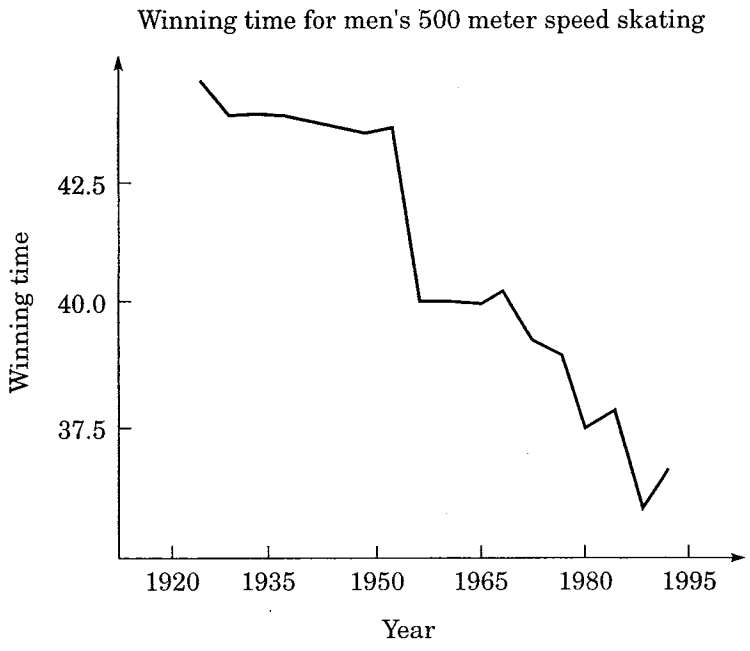
Graphs are most useful for displaying the relationship between two measurement variables or for displaying how a measurement variable changes over time. Two common types of displays for measurement variables are illustrated in Figures 9.4 and 9.5.

### 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 in the Winter Olympics from

**FIGURE 9.4**

Line Graph Displaying Winning Time versus Year for Olympic Speed Skating




SOURCE: *World Almanac and Book of Facts*, 1993, Mark S. Hoffman, ed., New York: Pharos Books, p. 832.

1924 to 1992. Notice the distinct downward trend, with only a few upturns over the years. There was a large drop between 1952 and 1956, followed by a period of relative stability. These patterns are much easier to detect with a picture than they would be by scanning a list of winning times.

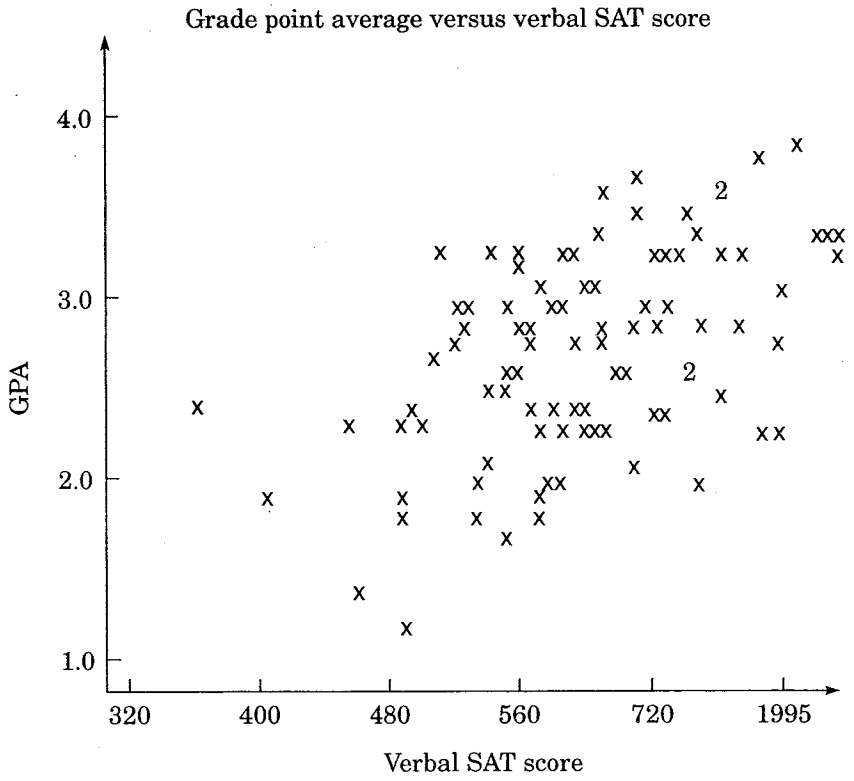
## Scatter Plots

Figure 9.5 is an example of a **scatter plot**. Scatter plots are useful for displaying the relationship between two measurement variables. Each "x" on the plot represents one individual. In the few cases where a "2" appears on the plot, two individuals had the same data. 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 scatter plot can be harder to read than a line graph, it displays more information. It shows outliers and how much variability 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 also that there is still substantial variability in GPAs at each level of verbal SAT scores. A scatter plot is definitely more useful than the raw data. Simply looking at a list of the 100 pairs of GPAs

**FIGURE 9.5**

Scatter Plot of Grade Point Average versus Verbal SAT Score




SOURCE: *Minitab Handbook*, 2nd edition, B. F. Ryan, B. L. Joiner and T. A. Ryan, Jr., 1985, Boston: PWS Kent, pgs. 309–312.

and SAT scores, we would find it difficult to detect the trend that is so obvious in the scatter plot.

9.4

## Difficulties and Disasters in Plots, Graphs and Pictures

A number of common mistakes that appear in plots and graphs allow the reader to be misled. 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
3. Change(s) in labeling on one or more axes
4. Misleading units of measurement
5. Graphs based on 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 9.6a gives an example of a plot for which the units were *not* labeled on the vertical axis. It 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 Figures 9.6b and c, displaying two different scenarios that could have produced the actual graph in Figure 9.6a. In Figure 9.6b, the vertical axis starts at zero for the existing plot. In Figure 9.6c, 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 those 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 9.6, what appears to be a substantial

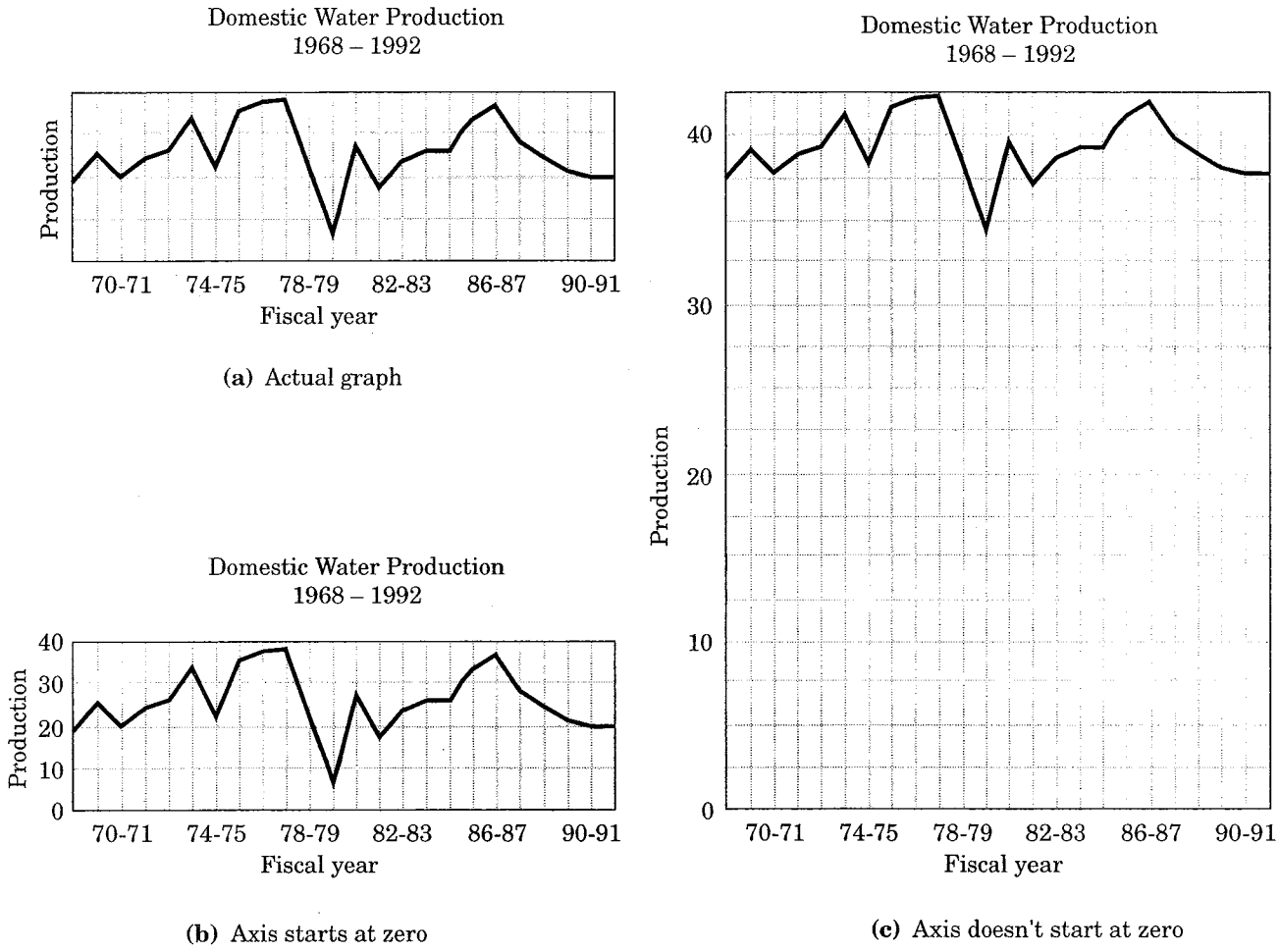



**FIGURE 9.6**

Example of a Graph with no Labeling (a), and Possible Interpretations (b and c)

SOURCE: Insert in the *California Aggie*, May 30, 1993.

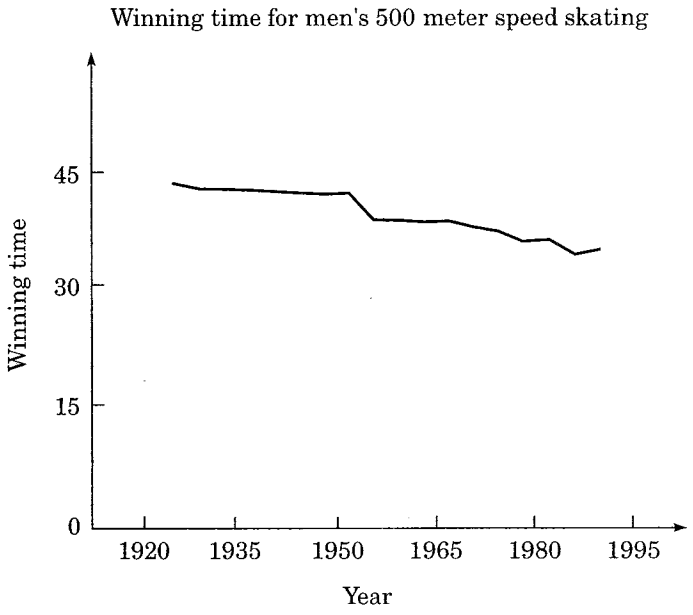
change may actually represent quite a modest change. Always make it a habit to check the numbers on the axes to see where they start.

Figure 9.7 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 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 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.

However, there are graphs for which it makes sense to start the units on the axes at values different from zero. A good example is the scatter plot of GPAs versus SAT scores in Figure 9.5. It would make no sense to start the horizontal axis (SAT scores) at zero, since the range of interest is from about 350 to 800. It is the

**FIGURE 9.7**

An Example of the Change in Perception When Axes Start at Zero




responsibility of the reader to notice the units. Never assume a graph starts at zero without checking the labeling.

## Change(s) in Labeling on One or More Axes

Figure 9.8 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.

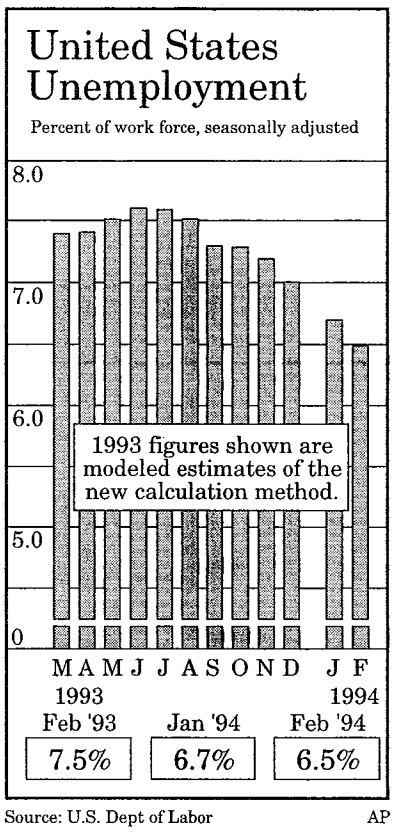
Figure 9.9 shows a much more egregious example of changes in labeling. Notice that the horizontal axis does not maintain consistent distances between years and that varying numbers of years are represented by each of the bars. The distance between the first and second bars on the left is 8 years, while the 5 bars farthest to the right each represent a single year. This is an extremely misleading graph.

## Misleading Units

The units shown on a graph can be different from those that the reader would consider important. For example, Figure 9.10 shows a graph with the heading, "Rising Postal Rates." It accurately represents how the cost of a first-class stamp has risen since 1971. However, notice that the fine print at the bottom reads, "In 1971 dollars, the price of a 32-cent stamp in February 1995 would be 8.4 cents." A more truthful picture would show the changing price of a first-class stamp adjusted for inflation. As the footnote implies, such a graph would show little or no rise in postal rates as a function of the worth of a dollar.

**FIGURE 9.8**

A Bar Graph with Gap in Labeling




SOURCE: *Davis Enterprise*, March 4, 1994, p. A-7.

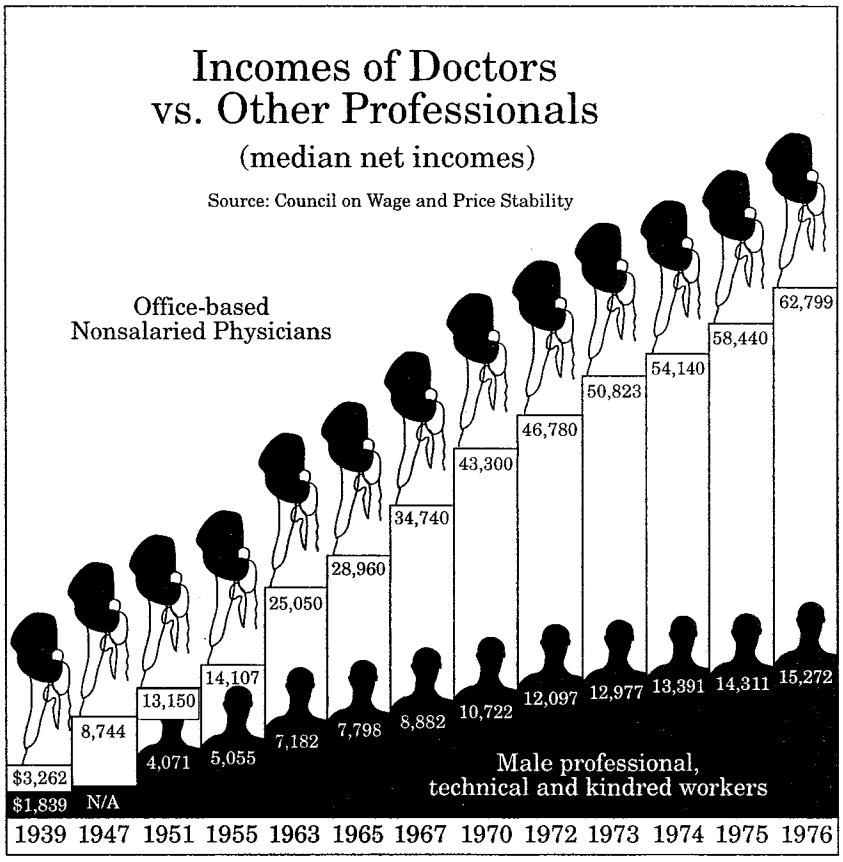
## Graphs Based on Poor Information

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 9.11 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 there were very few deaths in Britain from solvent abuse before the late 1970s. However, the accompanying article includes the following quote, made by 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.

**FIGURE 9.9**

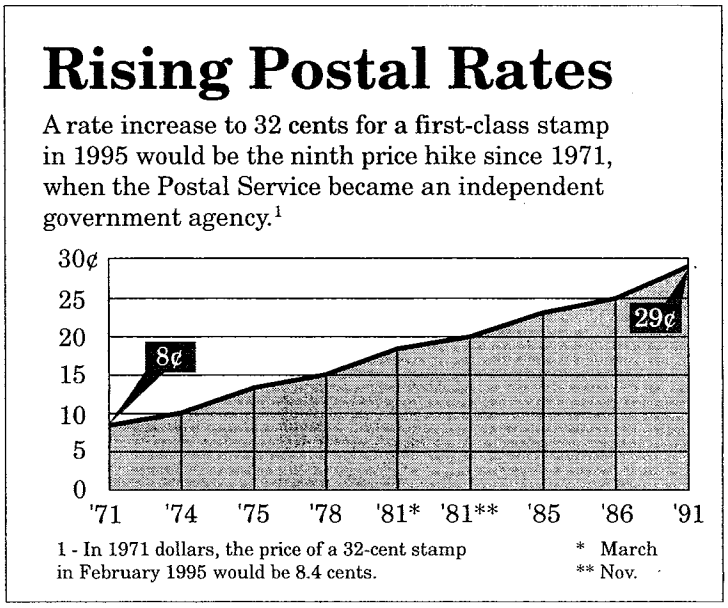
The Distance Between Successive Bars Keeps Changing




SOURCE: *Washington Post* graph reprinted in "How to Display Data Badly," Howard Wainer, 1984, *American Statistician*, Vol. 38.

**FIGURE 9.10**

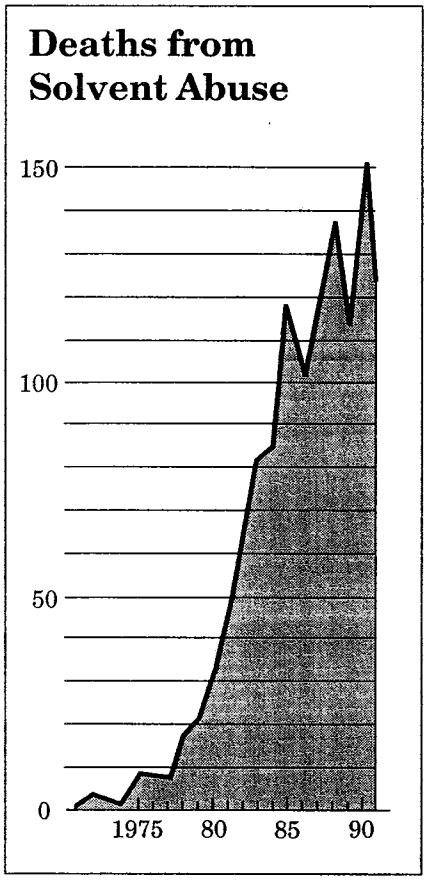
A Graph Using Misleading Units




SOURCE: *USA Today*, March 7, 1994, p. 13A.

**FIGURE 9.11**

A Graph Based on Poor Information




SOURCE: The *Independent on Sunday*, March 13, 1994, London.

9.5

## A Checklist for Statistical Pictures

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

1. Is the message of interest clearly standing out?
2. Is the purpose or title of the picture evident?
3. Is a source quoted 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. Do the axes maintain a constant scale?
8. Are there any breaks in the numbers on the axes that may be easy to miss?
9. For financial data, have the numbers been adjusted for inflation?
10. Is there extraneous information cluttering the picture or misleading the eye?

Don't forget that a statistical picture isn't worth much if the data can't be trusted. As with any statistical information, you should familiarize yourself to the extent possible with the Seven Critical Components listed in Chapter 2 (p. 17).

## *Exercises*

1. 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
2. Suppose a real estate company in your area sold 100 houses last month, while their two major competitors sold 50 houses and 25 houses. The top company wants to display their advantage 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 number of houses sold.)
3. 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 and 6 or more letters in length.
   - **b.** Find a paragraph of your own writing of at least 50 words. Repeat part a of this exercise for your own writing.