



Using Statistics to Save Lives

Florence Nightingale (1820–1910) won fame as a founder of the nursing profession and as a reformer of health care. As chief nurse for the British army during the Crimean War, from 1854 to 1856, she found that lack of sanitation and disease killed large numbers of soldiers hospitalized by wounds. Her reforms reduced the death rate at her military hospital from 42.7% to 2.2%, and she returned from the war famous. She at once began a fight to reform the entire military health care system, with considerable success.

One of the chief weapons Florence Nightingale used in her efforts was data. She had the facts, because she reformed record keeping as well as medical care. She was a pioneer in using graphs to present data in a vivid form that even generals and members of Parliament could understand. Her inventive graphs are a landmark in the growth of the new science of statistics. She considered statistics essential to understanding any social issue and tried to introduce the study of statistics into higher education.

In beginning our study of statistics, we will follow Florence Nightingale's lead. This chapter and the next will stress the analysis of data as a path to understanding. Like her, we will start with graphs to see what data can teach us. Along with the graphs we will present numerical summaries, just as Florence Nightingale calculated detailed death rates and other summaries. Data for Florence Nightingale were not dry or abstract, because they showed her, and helped her show others, how to save lives. That remains true today.

ACTIVITY 1 How Fast Is Your Heart Beating?

Materials: Clock or watch with second hand

A person's pulse rate provides information about the health of his or her heart. Would you expect to find a difference between male and female pulse rates? In this activity, you and your classmates will collect some data to try to answer this question.

1. To determine your pulse rate, hold the *fingers* of one hand on the artery in your neck or on the inside of the wrist. (The thumb should not be used, because there is a pulse in the thumb.) Count the number of pulse beats in one minute. Do this three times, and *calculate your average individual pulse rate* (add your three pulse rates and divide by 3.) Why is doing this three times better than doing it once?
2. Record the pulse rates for the class in a table, with one column for males and a second column for females. Are there any unusual pulse rates?
3. For now, simply calculate the average pulse rate for the males and the average pulse rate for the females, and compare.

INTRODUCTION

Statistics is the science of data. We begin our study of statistics by mastering the art of examining data. Any set of data contains information about some group of *individuals*. The information is organized in *variables*.

INDIVIDUALS AND VARIABLES

Individuals are the objects described by a set of data. Individuals may be people, but they may also be animals or things.

A **variable** is any characteristic of an individual. A variable can take different values for different individuals.

A college's student data base, for example, includes data about every currently enrolled student. The students are the *individuals* described by the data set. For each individual, the data contain the values of *variables* such as age, gender (female or male), choice of major, and grade point average. In practice, any set of data is accompanied by background information that helps us understand the data.

When you meet a new set of data, ask yourself the following questions:

1. **Who?** What **individuals** do the data describe? **How many** individuals appear in the data?
2. **What?** How many **variables** are there? What are the **exact definitions** of these variables? In what **units** is each variable recorded? Weights, for example, might be recorded in pounds, in thousands of pounds, or in kilograms. Is there any reason to mistrust the values of any variable?
3. **Why?** What is the reason the data were gathered? Do we hope to answer some specific questions? Do we want to draw conclusions about individuals other than the ones we actually have data for?

Some variables, like gender and college major, simply place individuals into categories. Others, like age and grade point average (GPA), take numerical values for which we can do arithmetic. It makes sense to give an average GPA for a college's students, but it does not make sense to give an "average" gender. We can, however, count the numbers of female and male students and do arithmetic with these counts.

CATEGORICAL AND QUANTITATIVE VARIABLES

A **categorical variable** places an individual into one of several groups or categories.

A **quantitative variable** takes numerical values for which arithmetic operations such as adding and averaging make sense.

EXAMPLE 1.1 EDUCATION IN THE UNITED STATES

Here is a small part of a data set that describes public education in the United States:

State	Region	Population (1000)	SAT Verbal	SAT Math	Percent taking	Percent no HS	Teachers' pay (\$1000)
:							
CA	PAC	33,871	497	514	49	23.8	43.7
CO	MTN	4,301	536	540	32	15.6	37.1
CT	NE	3,406	510	509	80	20.8	50.7
:							

case

Let's answer the three "W" questions about these data.

1. Who? The *individuals* described are the states. There are 51 of them, the 50 states and the District of Columbia, but we give data for only 3. Each row in the table describes one individual. You will often see each row of data called a *case*.

2. What? Each column contains the values of one variable for all the individuals. This is the usual arrangement in data tables. Seven variables are recorded for each state. The first column identifies the state by its two-letter post office code. We give data for California, Colorado, and Connecticut. The second column says which region of the country the state is in. The Census Bureau divides the nation into nine regions. These three are Pacific, Mountain, and New England. The third column contains state populations, in thousands of people. Be sure to notice that the *units* are thousands of people. California's 33,871 stands for 33,871,000 people. The population data come from the 2000 census. They are therefore quite accurate as of April 1, 2000, but don't show later changes in population.

The remaining five variables are the average scores of the states' high school seniors on the SAT verbal and mathematics exams, the percent of seniors who take the SAT, the percent of students who did not complete high school, and average teachers' salaries in thousands of dollars. Each of these variables needs more explanation before we can fully understand the data.

3. Why? Some people will use these data to evaluate the quality of individual states' educational programs. Others may compare states on one or more of the variables. Future teachers might want to know how much they can expect to earn.

A variable generally takes values that vary. One variable may take values that are very close together while another variable takes values that are quite spread out. We say that the *pattern of variation* of a variable is its *distribution*.

DISTRIBUTION

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

exploratory data
analysis

Statistical tools and ideas can help you examine data in order to describe their main features. This examination is called *exploratory data analysis*. Like an explorer crossing unknown lands, we first simply describe what we see. Each example we meet will have some background information to help us, but our emphasis is on examining the data. Here are two basic strategies that help us organize our exploration of a set of data:

- Begin by examining each variable by itself. Then move on to study relationships among the variables.
- Begin with a graph or graphs. Then add numerical summaries of specific aspects of the data.

We will organize our learning the same way. Chapters 1 and 2 examine single-variable data, and Chapters 3 and 4 look at relationships among variables. In both settings, we begin with graphs and then move on to numerical summaries.

EXERCISES

1.1 FUEL-EFFICIENT CARS Here is a small part of a data set that describes the fuel economy (in miles per gallon) of 1998 model motor vehicles:

Make and Model	Vehicle type	Transmission type	Number of cylinders	City MPG	Highway MPG
:					
BMW 318I	Subcompact	Automatic	4	22	31
BMW 318I	Subcompact	Manual	4	23	32
Buick Century	Midsize	Automatic	6	20	29
Chevrolet Blazer	Four-wheel drive	Automatic	6	16	20
:					

- What are the individuals in this data set?
- For each individual, what variables are given? Which of these variables are categorical and which are quantitative?

1.2 MEDICAL STUDY VARIABLES Data from a medical study contain values of many variables for each of the people who were the subjects of the study. Which of the following variables are categorical and which are quantitative?

- Gender (female or male)
- Age (years)
- Race (Asian, black, white, or other)
- Smoker (yes or no)
- Systolic blood pressure (millimeters of mercury)
- Level of calcium in the blood (micrograms per milliliter)

1.3 You want to compare the “size” of several statistics textbooks. Describe at least three possible numerical variables that describe the “size” of a book. In what *units* would you measure each variable?

1.4 Popular magazines often rank cities in terms of how desirable it is to live and work in each city. Describe five variables that you would measure for each city if you were designing such a study. Give reasons for each of your choices.

1.1 DISPLAYING DISTRIBUTIONS WITH GRAPHS

Displaying categorical variables: bar graphs and pie charts

The values of a categorical variable are labels for the categories, such as “male” and “female.” The distribution of a categorical variable lists the categories and gives either the **count** or the **percent** of individuals who fall in each category.

EXAMPLE 1.2 THE MOST POPULAR SOFT DRINK

The following table displays the sales figures and market share (percent of total sales) achieved by several major soft drink companies in 1999. That year, a total of 9930 million cases of soft drink were sold.¹

Company	Cases sold (millions)	Market share (percent)
Coca-Cola Co.	4377.5	44.1
Pepsi-Cola Co.	3119.5	31.4
Dr. Pepper/7-Up (Cadbury)	1455.1	14.7
Cott Corp.	310.0	3.1
National Beverage	205.0	2.1
Royal Crown	115.4	1.2
Other	347.5	3.4

How to construct a bar graph:

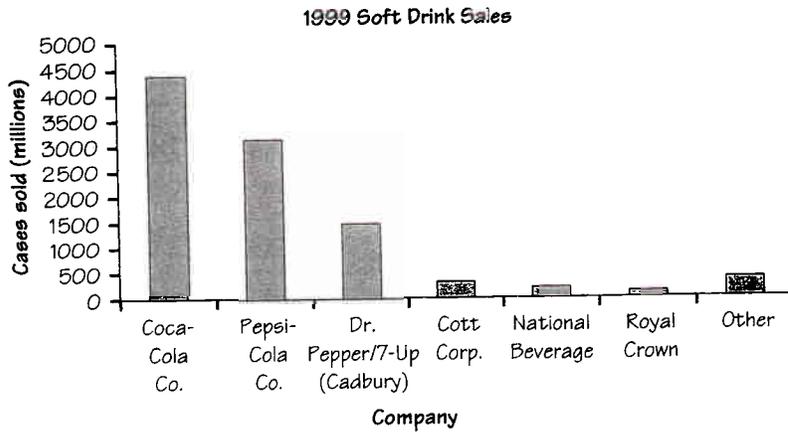
Step 1: Label your axes and title your graph. Draw a set of axes. Label the horizontal axis “Company” and the vertical axis “Cases sold.” Title your graph.

Step 2: Scale your axes. Use the counts in each category to help you scale your vertical axis. Write the category names at equally spaced intervals beneath the horizontal axis.

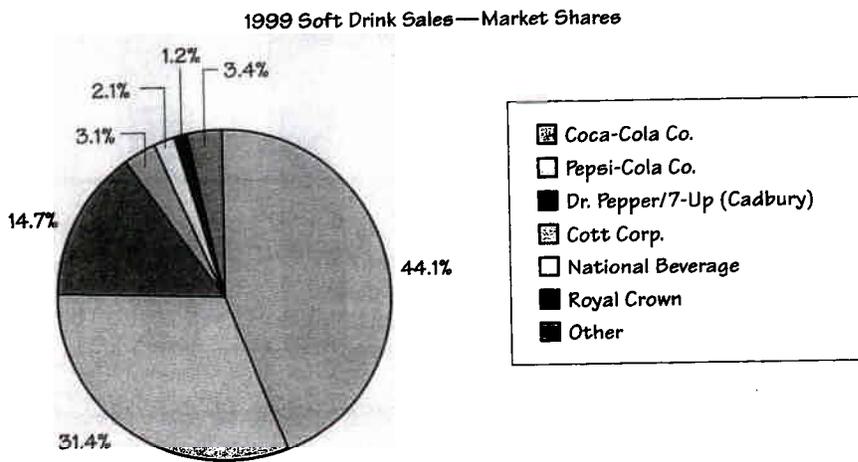
Step 3: Draw a vertical bar above each category name to a height that corresponds to the count in that category. For example, the height of the “Pepsi-Cola Co.” bar should be at 3119.5 on the vertical scale. *Leave a space between the bars in a bar graph.*

Figure 1.1(a) displays the completed bar graph.

How to construct a pie chart: Use a computer! Any statistical software package and many spreadsheet programs will construct these plots for you. Figure 1.1(b) is a pie chart for the soft drink sales data.



(a)



(b)

FIGURE 1.1 A bar graph (a) and a pie chart (b) displaying soft drink sales by companies in 1999.

The **bar graph** in Figure 1.1(a) quickly compares the soft drink sales of the companies. The heights of the bars show the counts in the seven categories. The **pie chart** in Figure 1.1(b) helps us see what part of the whole each group forms. For example, the Coca-Cola “slice” makes up 44.1% of the pie because the Coca-Cola Company sold 44.1% of all soft drinks in 1999.

Bar graphs and pie charts help an audience grasp the distribution quickly. To make a pie chart, you must include all the categories that make up a whole. Bar graphs are more flexible.

EXAMPLE 1.3 DO YOU WEAR YOUR SEAT BELT?

In 1998, the National Highway and Traffic Safety Administration (NHTSA) conducted a study on seat belt use. The table below shows the percentage of automobile drivers who were observed to be wearing their seat belts in each region of the United States.²

Region	Percent wearing seat belts
Northeast	66.4
Midwest	63.6
South	78.9
West	80.8

Figure 1.2 shows a bar graph for these data. Notice that the vertical scale is measured in percents.

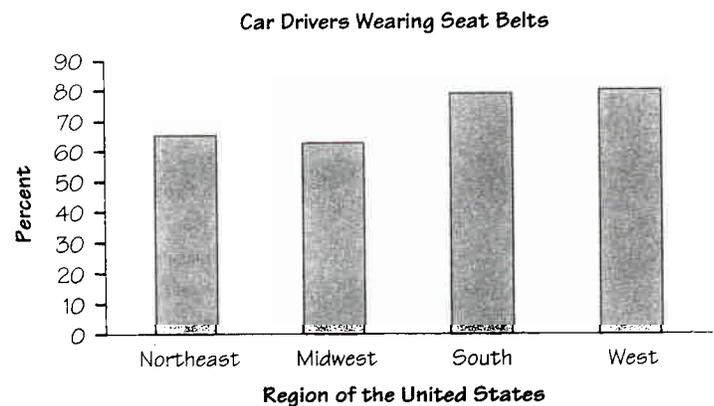


FIGURE 1.2 A bar graph showing the percentage of drivers who wear their seat belts in each of four U.S. regions.

Drivers in the South and West seem to be more concerned about wearing seat belts than those in the Northeast and Midwest. It is not possible to display these data in a single pie chart, because the four percentages cannot be combined to yield a whole (their sum is well over 100%).

EXERCISES

1.5 FEMALE DOCTORATES Here are data on the percent of females among people earning doctorates in 1994 in several fields of study.³

Computer science	15.4%	Life sciences	40.7%
Education	60.8%	Physical sciences	21.7%
Engineering	11.1%	Psychology	62.2%

- (a) Present these data in a well-labeled bar graph.
- (b) Would it also be correct to use a pie chart to display these data? If so, construct the pie chart. If not, explain why not.

1.6 ACCIDENTAL DEATHS In 1997 there were 92,353 deaths from accidents in the United States. Among these were 42,340 deaths from motor vehicle accidents, 11,858 from falls, 10,163 from poisoning, 4051 from drowning, and 3601 from fires.⁴

- (a) Find the percent of accidental deaths from each of these causes, rounded to the nearest percent. What percent of accidental deaths were due to other causes?
- (b) Make a well-labeled bar graph of the distribution of causes of accidental deaths. Be sure to include an "other causes" bar.
- (c) Would it also be correct to use a pie chart to display these data? If so, construct the pie chart. If not, explain why not.

Displaying quantitative variables: dotplots and stemplots

Several types of graphs can be used to display quantitative data. One of the simplest to construct is a **dotplot**.

EXAMPLE 1.4 GOOOOOOOAAAAALLLLLLLLLL!!!

The number of goals scored by each team in the first round of the California Southern Section Division V high school soccer playoffs is shown in the following table.⁵

5	0	1	0	7	2	1	0	4	0	3	0	2	0
3	1	5	0	3	0	1	0	1	0	2	0	3	1

How to construct a dotplot:

Step 1: Label your axis and title your graph. Draw a horizontal line and label it with the variable (in this case, number of goals scored). Title your graph.

Step 2: Scale the axis based on the values of the variable.

Step 3: Mark a dot above the number on the horizontal axis corresponding to each data value. Figure 1.3 displays the completed dotplot.

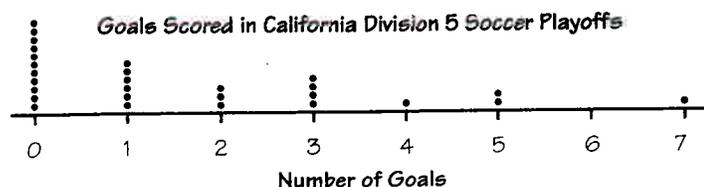


FIGURE 1.3 Goals scored by teams in the California Southern Section Division V high school soccer playoffs.

Making a statistical graph is not an end in itself. After all, a computer or graphing calculator can make graphs faster than we can. The purpose of the graph help us understand the data. After you (or your calculator) make a graph, always ask, “What do I see?” Here is a general tactic for looking at graphs: *Look for an overall pattern and also for striking deviations from that pattern.*

OVERALL PATTERN OF A DISTRIBUTION

To describe the overall pattern of a distribution:

- Give the **center** and the **spread**.
- See if the distribution has a simple **shape** that you can describe in a few words.

Section 1.2 tells in detail how to measure center and spread. For now, describe the *center* by finding a value that divides the observations so that about half take larger values and about half have smaller values. In Figure 1.3, the center is 1. That is, a typical team scored about 1 goal in its playoff soccer game. You can describe the *spread* by giving the smallest and largest values. The spread in Figure 1.3 is from 0 goals to 7 goals scored.

The dotplot in Figure 1.3 shows that in most of the playoff games, Division V soccer teams scored very few goals. There were only four teams that scored 4 or more goals. We can say that the distribution has a “long tail” to the right, or that its *shape* is “skewed right.” You will learn more about describing shape shortly.

Is the one team that scored 7 goals an *outlier*? This value certainly differs from the overall pattern. To some extent, deciding whether an observation is an outlier is a matter of judgment. We will introduce an objective criterion for determining outliers in Section 1.2.

OUTLIERS

An **outlier** in any graph of data is an individual observation that falls outside the overall pattern of the graph.

Once you have spotted outliers, look for an explanation. Many outliers are due to mistakes, such as typing 4.0 as 40. Other outliers point to the special nature of some observations. Explaining outliers usually requires some background information. Perhaps the soccer team that scored seven goals has some very talented offensive players. Or maybe their opponents played poor defense.

Sometimes the values of a variable are too spread out for us to make a reasonable dotplot. In these cases, we can consider another simple graphical display: a **stemplot**.