Section 1.1 Summary

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

To describe a distribution, begin with a graph. **Bar graphs** and **pie charts** display the distributions of categorical variables. These graphs use the counts or percents of the categories. **Stemplots** and **histograms** display the distributions of quantitative variables. Stemplots separate each observation into a stem and a one-digit leaf. Histograms plot the **frequencies** (counts) or percents of equal-width classes of values.

When examining a distribution, look for **shape, center,** and **spread,** and for clear **deviations** from the overall shape. Some distributions have simple shapes, such as **symmetric** or **skewed.** The number of **modes** (major peaks) is another aspect of overall shape. Not all distributions have a simple overall shape, especially when there are few observations.

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

A **relative cumulative frequency graph** (also called an **ogive**) is a good way to see the relative standing of an observation.

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** or other changes over time.

Section 1.1 Exercises

1.19 Ranking colleges Popular magazines rank colleges and universities on their "academic quality" in serving undergraduate students. Describe five variables that you would like to see measured for each college if you were choosing where to study. Give reasons for each of your choices.

1.20 Shopping spree, III Enter the amount of money spent in a supermarket from Exercise 1.6 (page 48) into your calculator. Then use the information in the Technology Toolbox (page 59) to construct a histogram. Use `ZoomStat/ZoomData` first to see what the calculator chooses for class widths. Then, in the calculator's WINDOW, choose new settings that are more sensible. Compare your histogram with the stemplots you made in Exercise 1.6. List at least one advantage that each plot has that the other plots don't have.

1.21 College costs The Department of Education estimates the "average unmet need" for undergraduate students—the cost of school minus estimated family contributions and financial aid. Here are the averages for full-time students at four types of institution in the 1999–2000 academic year:[16]

Public 2-year	Public 4-year	Private nonprofit 4-year	Private for profit
$4495	$4818	$8257	$8296

Make a bar graph of these data. Write a one-sentence conclusion about the unmet needs of students. Explain clearly why it is incorrect to make a pie chart.

FYI There are four quizzes in the *Platinum Resource Binder* for Section 1.1.

Answers to Exercises 1.19–1.26

1.19 Student answers will vary. Some possibilities: academic reputation, retention rate, graduation rate, class sizes, faculty salaries, etc.

1.20 See the *Teacher's Solutions Manual* for graph.

1.21 See the *Teacher's Solutions Manual* for graph. Unmet need is greater at private institutions than it is at public institutions. A pie chart would be incorrect because these numbers do not represent parts of a single whole.

1.22 See the *Teacher's Solutions Manual* for graph. The time plots show that both manufacturers have generally improved over this period.

1.23 (a) Alaska is 5.7%; Florida is 17.6%.

(b) Roughly symmetric (perhaps slightly skewed to the left) and centered near 13%. Ignoring the outliers, the percentages range from 8.5% to 15.6%.

1.24 See the *Teacher's Solutions Manual* for graph.

1.25 (a) See the *Teacher's Solutions Manual* for graph. The distribution is skewed to the right with a single peak. The center is at 4, with a spread from 1 to 15. There are no gaps or outliers.

(b) There are more 2, 3, and 4 letter words in Shakespeare's plays and more very long words in *Popular Science* articles.

1.22 New-vehicle survey The J. D. Power Initial Quality Study polls more than 50,000 buyers of new motor vehicles 90 days after their purchase. A two-page questionnaire asks about "things gone wrong." Here are data on problems per 100 vehicles for vehicles made by Toyota and by General Motors in recent years. Toyota has been the industry leader in quality. Make two time plots in the same graph to compare Toyota and GM. What are the most important conclusions you can draw from your graph?

	1998	1999	2000	2001	2002	2003	2004
GM	187	179	164	147	130	134	120
Toyota	156	134	116	115	107	115	101

1.23 Senior citizens, I The population of the United States is aging, though less rapidly than in other developed countries. Here is a stemplot of the percents of residents aged 65 and over in the 50 states, according to the 2000 census. The stems are whole percents and the leaves are tenths of a percent.

5	7
6	
7	
8	5
9	6 7 9
10	6
11	0 2 2 3 3 6 7 7
12	0 0 1 1 1 1 3 4 4 5 7 8 9
13	0 0 0 1 2 2 3 3 3 4 5 5 6 8
14	0 3 4 5 7 9
15	3 6
16	
17	6

(a) There are two outliers: Alaska has the lowest percent of older residents, and Florida has the highest. What are the percents for these two states?

(b) Ignoring Alaska and Florida, describe the shape, center, and spread of this distribution.

1.24 Senior citizens, II Make another stemplot of the percent of residents aged 65 and over in the states other than Alaska and Florida by splitting stems in the plot from the previous exercise. Which plot do you prefer? Why?

1.25 The statistics of writing style Numerical data can distinguish different types of writing, and sometimes even individual authors. Here are data on the percent of words of 1 to 15 letters used in articles in *Popular Science* magazine:[17]

Length:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Percent:	3.6	14.8	18.7	16.0	12.5	8.2	8.1	5.9	4.4	3.6	2.1	0.9	0.6	0.4	0.2

(a) Make a histogram of this distribution. Describe its shape, center, and spread.

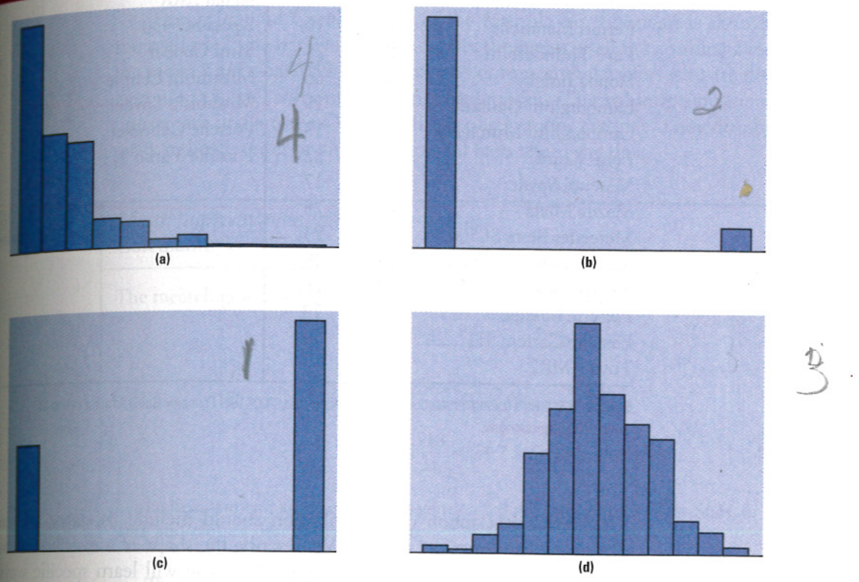
(b) How does the distribution of lengths of words used in *Popular Science* compare with the similar distribution in Figure 1.11 (page 57) for Shakespeare's plays? Look in particular at short words (2, 3, and 4 letters) and very long words (more than 10 letters).

1.26 Student survey A survey of a large high school class asked the following questions:

1. Are you female or male? (In the data, male = 0, female = 1.)
2. Are you right-handed or left-handed? (In the data, right = 0, left = 1.)
3. What is your height in inches?
4. How many minutes do you study on a typical weeknight?

Figure 1.16 shows histograms of the student responses, in scrambled order and without scale markings. Which histogram goes with each variable? Explain your reasoning.

Figure 1.16 *Match each histogram with its variable, for Exercise 1.26.*



1.2 Describing Distributions with Numbers

Interested in a sporty car? Worried that it might use too much gas? The Environmental Protection Agency lists most such vehicles in its "two-seater" or "minicompact" categories. Table 1.6 on the next page gives the city and highway gas mileage for cars in these groups. (The mileages are for the basic engine and transmission combination for each car.) We want to compare two-seaters with minicompacts and city mileage with highway mileage. We can begin with graphs, but numerical summaries make the comparisons more specific.

1.26 From the top left histogram: 4, 2, 1, 3.

Section 1.2 Objectives

- Given a data set, compute the *mean* and *median* as *measures of center.*
- Explain what is meant by a *resistant measure.*
- Identify situations in which the *mean* is the most appropriate measure of center and situations in which the *median* is the most appropriate measure.
- Given a data set, find the *quartiles.*
- Given a data set, find the *five-number summary.*
- Use the *five-number summary* of a data set to construct a *boxplot* for the data.
- Compute the *interquartile range* (*IQR*) of a data set.
- Given a data set, use the *1.5 × IQR rule* to identify outliers.
- Given a data set, compute the *standard deviation* and *variance* as measures of *spread.*
- Give two reasons why we use squared deviations rather just average deviations from the mean.
- Explain what is meant by *degrees of freedom.*
- Identify situations in which the *standard deviation* is the most appropriate measure of spread and situations in which the *interquartile range* is the most appropriate measure.
- Explain the effect of a *linear transformation* of a data set on the *mean, median,* and *standard deviation* of the set.
- Use numerical and graphical techniques to compare two or more data sets.

Getting Started

- Introduce the section by giving the students a data set that contains an outlier (such as the following set of ten exam scores in which one student obviously failed to study enough: 75, 76, 82, 93, 45, 68, 74, 82, 91, 98) and have a discussion about how they might describe the center of the data, how they would describe the spread of the data, and how they might best deal with the outlier.
- Point out that **Section 1.1** dealt with the graphical approach to data analysis, through which we gain information about the *shape* of a distribution. The first step in data analysis is always to *look* at the data. This section deals with the numerical approach to data analysis, through which we gain information about *center* and *spread* of a distribution.