EXPLORING DATA

Introduction to Data

- Statistics is the science (and art) of collecting and analyzing observations (called data) and communicating your discoveries to others.

- Often, we are interested in learning about a population on the basis of a sample taken from that population.

- Here are some questions to consider when first examining a data set:
  - Who, or what, was observed?
  - What variables were measured?
  - How were they measured?
  - What are the units of measurement?
  - Who collected the data?
  - How did they collect the data?
  - Where were the data collected?
  - Why were the data collected?
  - When were the data collected?

- With categorical variables, we are often concerned with comparing rates or frequencies between groups. A two-way table is sometimes a useful summary. Always be sure that you are making valid comparisons by comparing proportions or percentages of groups, or that you are comparing the appropriate rates.

- Many studies are focused on questions of causality: If we make a change to one variable, will we see a change in the other? Anecdotes are not useful for answering such questions. Observational studies can be used to determine whether associations exist between treatment and outcome variables, but because of the possibility of confounding variables, observational studies cannot support conclusions about causality. Controlled experiments, if they are well designed, do allow us to draw conclusions about causality.

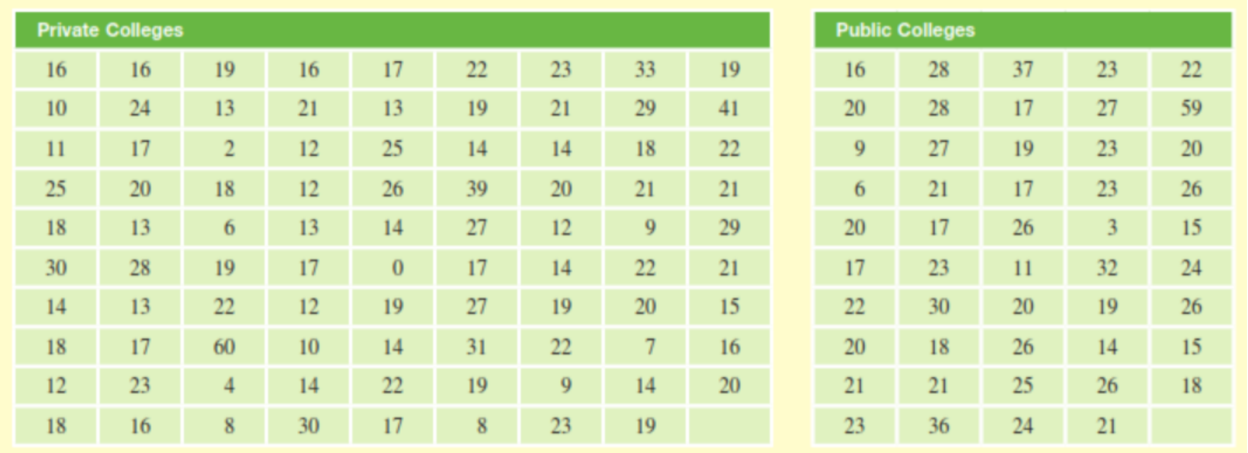
- A well-designed controlled experiment should have the following attributes:
  - A large sample size
  - Random assignment of subjects to a treatment group and to a control group
  - A double-blind format
  - A placebo

Picturing Variation with Graphs

- Any collection of data exhibits variation.
- The most important tool for organizing this variation is called the distribution of the sample, and visualizing this distribution is the first step in every statistical investigation.
- We can learn much about a numerical variable by focusing on three components of the distribution:
  - the shape,
  - the center, and
  - the variability, or horizontal spread.
- Examining a graph of a distribution can lead us to deeper understanding of the situation that produced the data.

CASE STUDY : Student-to-Teacher Ratio at Colleges

- Are private four-year colleges better than public four-year colleges? That depends on what you mean by "better."
- One measure of quality that many people find useful (and there are many other ways to measure quality) is the student-to-teacher ratio: the number of students enrolled divided by the number of teachers.
- For schools with small student-to-teacher ratios, we expect class sizes to be small; students can get extra attention in a small class.
- The data in Table 2.1 on the next page were collected from some schools that award four-year degrees. The data are for the 2010–2011 academic year; 89 private colleges and 49 state-supported (public) colleges were sampled. Each ratio was rounded to the nearest whole number for simplicity.



Private Colleges								
16	16	19	16	17	22	23	33	19
10	24	13	21	13	19	21	29	41
11	17	2	12	25	14	14	18	22
25	20	18	12	26	39	20	21	21
18	13	6	13	14	27	12	9	29
30	28	19	17	0	17	14	22	21
14	13	22	12	19	27	19	20	15
18	17	60	10	14	31	22	7	16
12	23	4	14	22	19	9	14	20
18	16	8	30	17	8	23	19	

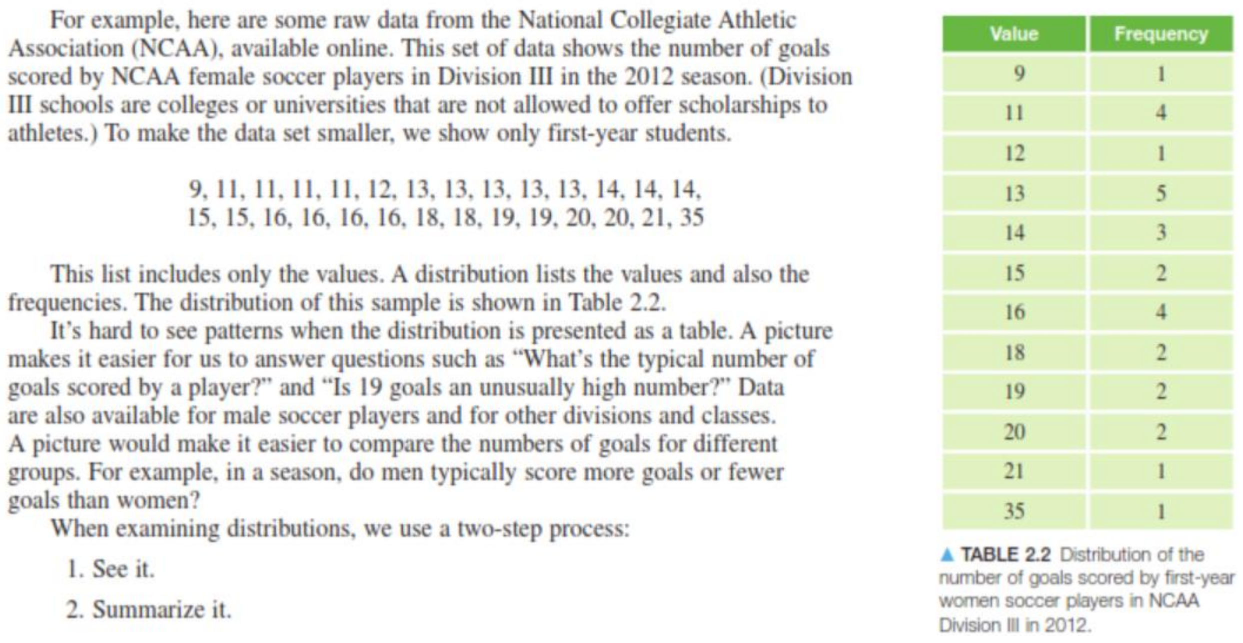
Public Colleges				
16	28	37	23	22
20	28	17	27	59
9	27	19	23	20
6	21	17	23	26
20	17	26	3	15
17	23	11	32	24
22	30	20	19	26
20	18	26	14	15
21	21	25	26	18
23	36	24	21	

▲ **TABLE 2.1** Ratio of students to teachers at private and public colleges.
(Source: http://nces.ed.gov/ipeds/)

- For example, the first private college listed has a student-to-teacher ratio of 16, which means that there are about 16 students for every teacher.
  - What differences do you expect between the two groups?
  - What similarities do you anticipate?
- It is nearly impossible to compare the two groups without imposing some kind of organization on the data.

2.1 Visualizing Variation in Numerical Data

- The distribution of a sample is one of the central organizational concepts of data analysis.
- The distribution organizes data by recording all of the values observed in a sample, as well as how many times each value was observed.



For example, here are some raw data from the National Collegiate Athletic Association (NCAA), available online. This set of data shows the number of goals scored by NCAA female soccer players in Division III in the 2012 season. (Division III schools are colleges or universities that are not allowed to offer scholarships to athletes.) To make the data set smaller, we show only first-year students.

9, 11, 11, 11, 11, 12, 13, 13, 13, 13, 13, 14, 14, 14,
15, 15, 16, 16, 16, 16, 18, 18, 19, 19, 20, 20, 21, 35

This list includes only the values. A distribution lists the values and also the frequencies. The distribution of this sample is shown in Table 2.2.

It's hard to see patterns when the distribution is presented as a table. A picture makes it easier for us to answer questions such as "What's the typical number of goals scored by a player?" and "Is 19 goals an unusually high number?" Data are also available for male soccer players and for other divisions and classes. A picture would make it easier to compare the numbers of goals for different groups. For example, in a season, do men typically score more goals or fewer goals than women?

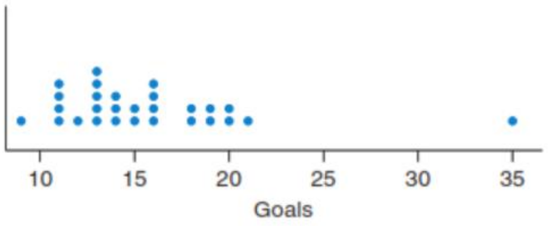
When examining distributions, we use a two-step process:

1. See it.
2. Summarize it.

Value	Frequency
9	1
11	4
12	1
13	5
14	3
15	2
16	4
18	2
19	2
20	2
21	1
35	1

▲ **TABLE 2.2** Distribution of the number of goals scored by first-year women soccer players in NCAA Division III in 2012.

- **Dotplots**



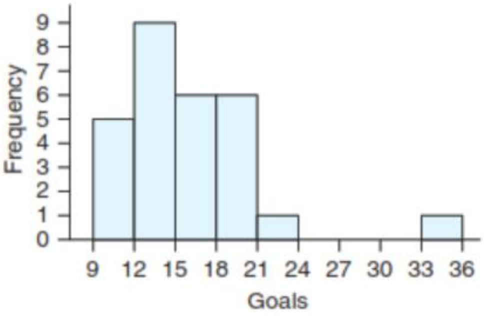
Tech

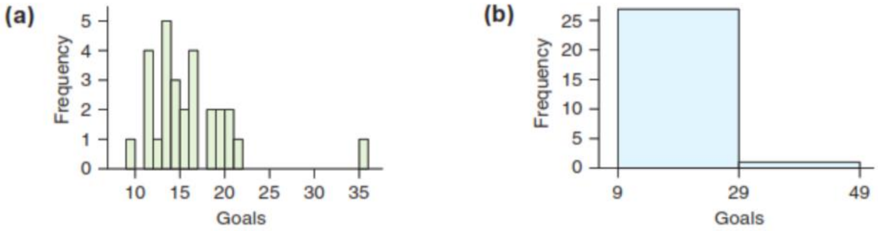
◀ **FIGURE 2.2** Dotplot of the number of goals scored by first-year women soccer players in NCAA Division III, 2012. Each dot represents a soccer player. Note that the horizontal axis begins at 9.

- **Histograms**

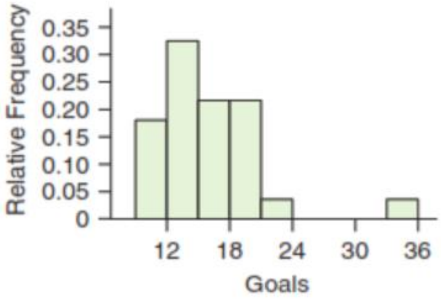
▶ **FIGURE 2.3** Histogram of goals for female first-year soccer players in NCAA Division III, 2012. The first bar, for example, tells us that five players scored between 9 and 12 goals during the season.

Tech





◀ **FIGURE 2.4** Two more histograms of goals scored in one season the same data as in Figure 2.3. **(a)** This histogram has narrow bins and is spiky. **(b)** This histogram has wide bins and offers less detail.



◀ **FIGURE 2.5** Relative frequency histogram of goals scored by first-year women soccer players in NCAA Division III, 2012.

WHAT IS IT? ▶ A graphical summary for numerical data.

WHAT DOES IT DO? ▶ Shows a picture of the distribution of a numerical variable.

HOW DOES IT DO IT? ▶ Observations are grouped into bins, and bars are drawn to show how many observations (or what proportion of observations) lie in each bin.

HOW IS IT USED? ▶ By smoothing over details, histograms help our eyes pick up more important, large-scale patterns. Be aware that making the bins wider hides detail, and making the bins smaller can show too much detail. The vertical axis can display frequency, relative frequency, or percents.

- **Stemplots**

Stemplots, which are also called stem-and-leaf plots, are useful for visualizing numerical variables when you don't have access to technology and the data set is not large. Stemplots are also useful if you want to be able to easily see the actual values of the data.

To make a **stemplot**, divide each observation into a "stem" and a "leaf." The **leaf** is the last digit in the observation. The **stem** contains all the digits that precede the leaf. For the number 60, the *6* is the stem and the *0* is the leaf. For the number 632, the *63* is the stem and the *2* is the leaf. For the number 65.4, the *65* is the stem and the *4* is the leaf.

A stem-and-leaf plot can help us understand data such as drinking behaviors. Alcohol is a big problem at many colleges and universities. For this reason, a collection of college students who said that they drink alcohol were asked how many alcoholic drinks they had consumed in the last seven days. Their answers were

1, 1, 1, 1, 1, 2, 2, 2, 3, 3, 3, 3, 3, 4, 5, 5, 5, 6, 6, 6, 8, 10, 10, 15, 17, 20, 25, 30, 30, 40

For one-digit numbers, imagine a 0 at the front. The observation of 1 drink becomes 01, the observation of 2 drinks becomes 02, and so on. Then each observation is just two digits; the first digit is the stem, and the last digit is the leaf. Figure 2.7 shows a stemplot of these data.

Stem	Leaves
0	111112223333345556668
1	0057
2	05
3	00
4	0

▲ **FIGURE 2.7** A stemplot for alcoholic drinks consumed by college students. Each digit on the right (the leaves) represents a student. Together, the stem and the leaf indicate the number of drinks for an individual student.

WHAT IS IT? ► A graphical summary for numerical data.

WHAT DOES IT DO? ► Shows a picture of the distribution of a numerical variable.

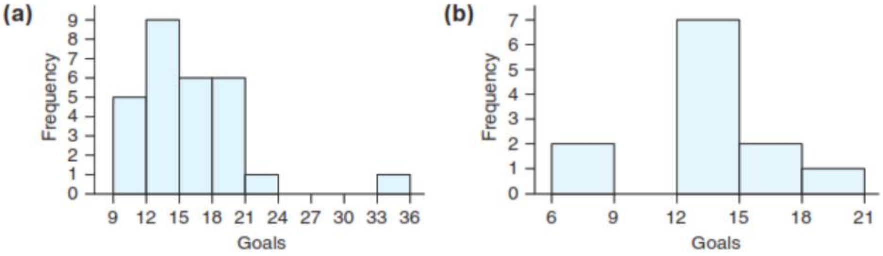
HOW DOES IT DO IT? ► Numbers are divided into leaves (the last digit) and stems (the preceding digits). Stems are written in a vertical column, and associated leaves are "attached."

HOW IS IT USED? ► In very much the same way as a histogram. It is often useful when technology is not available and the data set is not large.

2.2 Summarizing Important Features of a Numerical Distribution

- When examining distributions of numerical data, pay attention to the shape, center, and horizontal spread.

Figure 2.9 compares distributions for two groups. You've already seen histogram (a)—it's the histogram for the goals scored in 2012 by first-year women soccer players in Division III. Histogram (b) shows goals scored for first-year male soccer players in Division III in the same year. How do these two distributions compare?



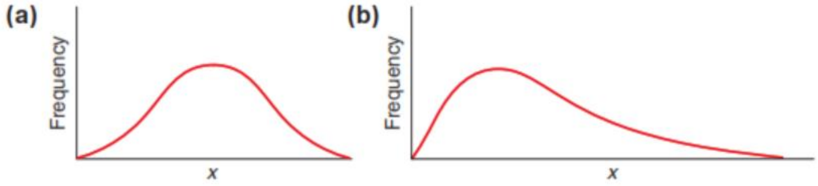
► **FIGURE 2.9** Distributions of the goals scored for **(a)** first-year women and **(b)** first-year men in Division III soccer in 2012.

1. *Shape*. Are there any interesting or unusual features about the distributions? Are the shapes very different? (If so, this might be evidence that men play the game differently than women.)
2. *Center*. What is the typical value of each distribution? Is the typical number of goals scored per game different for men than for women?
3. *Spread*. The horizontal spread presents the variation in goals per game for each group. How do the amounts of variation compare? If one group has low variation, it suggests that the soccer skills of the members are pretty much the same. Lots of variability might mean that there is a wider variety of skill levels.

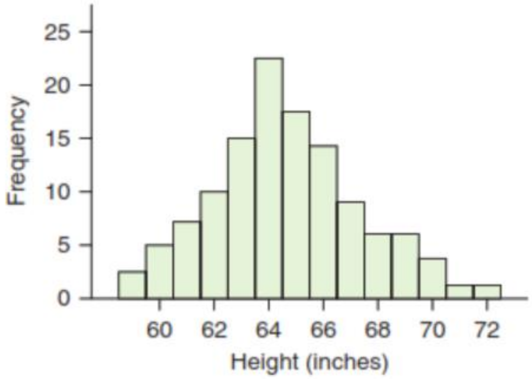
- **Shape**

You should look for three basic characteristics of a distribution's shape:

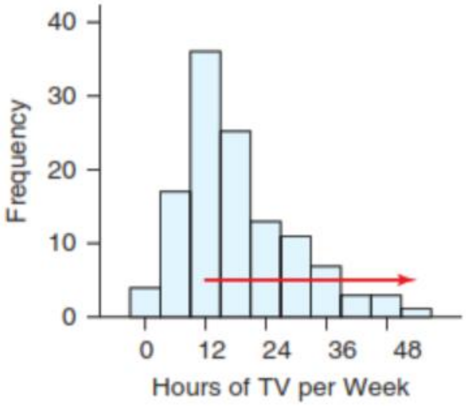
1. Is the distribution symmetric or skewed?
2. How many mounds appear? One? Two? None? Many?
3. Are unusually large or small values present?



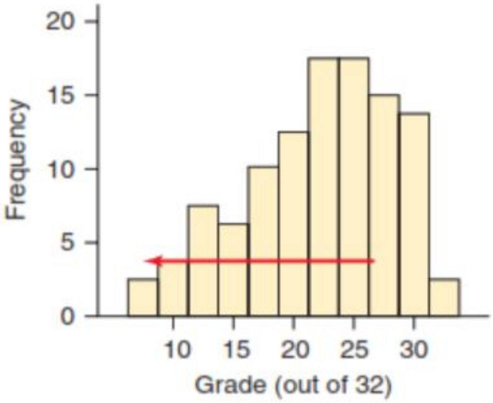
▶ **FIGURE 2.10** Sketches of **(a)** a symmetric distribution and **(b)** a right-skewed distribution.



◀ **FIGURE 2.11** Histogram of heights of women. (Source: Brian Joiner in Tufte 1983)



▲ **FIGURE 2.12** This data set on TV hours viewed per week is skewed to the right. (Source: Minitab Program)

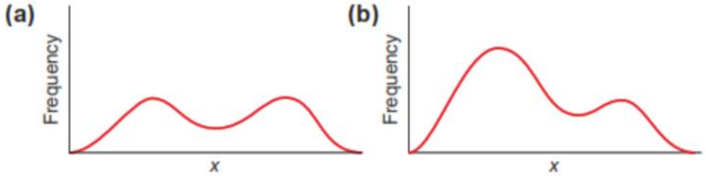


▲ **FIGURE 2.13** This data set on test scores is skewed to the left.

- **Mounds (gundukan)**

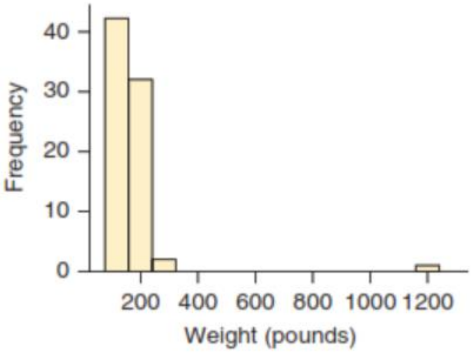
The statistical term for a one-mound distribution is **unimodal distribution**, and a two-mound distribution is called a **bimodal distribution**. Figure 2.14a shows a bimodal distribution. A **multimodal distribution** has more than two modes. The modes do not have to be the same height (in fact, they rarely are). Figure 2.14b is perhaps the more typical bimodal distribution.

▶ **FIGURE 2.14** Idealized bimodal distributions. **(a)** Modes of roughly equal height. **(b)** Modes that differ in height.



- Outliers are values so large or small that they do not fit into the pattern of the distribution. There is no precise definition of the term outlier. Outliers can be caused by mistakes in data entry, but genuine outliers are sometimes unusually interesting observations.

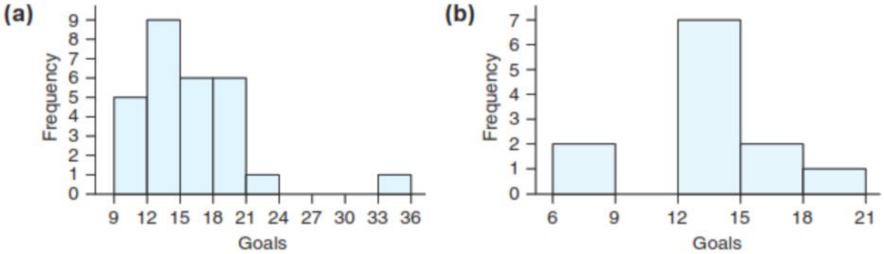
▶ **FIGURE 2.17** Histogram of weights with an extreme value.



- **Center**
  - An important question to ask about any data set is "What is the typical value?" The typical value is the one in the center, but we use the word center here in a deliberately vague way. We might not all agree on precisely where the center of a graph is. The idea here is to get a rough impression so that we can make comparisons later. For example, judging on the basis of the histogram shown in Figure 2.9, the center for thewomen soccer players is about 16 goals. Thus we could say that the typical first-year woman soccer player scored ab ut 16 goals in 2012. In contrast, the center of the distribution for the male soccer players is about 13 goals.

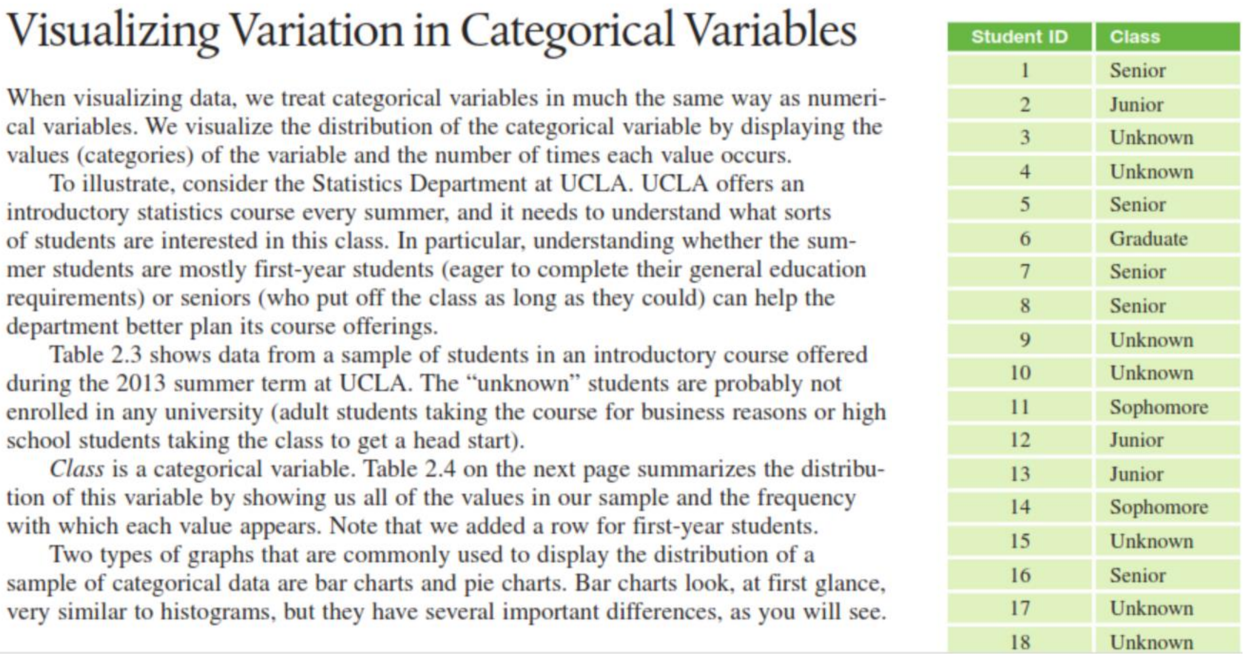
Figure 2.9 compares distributions for two groups. You've already seen histogram (a)—it's the histogram for the goals scored in 2012 by first-year women soccer players in Division III. Histogram (b) shows goals scored for first-year male soccer players in Division III in the same year. How do these two distributions compare?

▶ **FIGURE 2.9** Distributions of the goals scored for **(a)** first-year women and **(b)** first-year men in Division III soccer in 2012.



- It would seem that the typical male soccer player scores fewer goals in a season than the typical female player, perhaps indicating that men’s soccer is stronger on defense than women’s soccer, at least among Division III first-year players.

2.3 Visualizing Variation in Categorical Variables



Visualizing Variation in Categorical Variables

When visualizing data, we treat categorical variables in much the same way as numerical variables. We visualize the distribution of the categorical variable by displaying the values (categories) of the variable and the number of times each value occurs.

To illustrate, consider the Statistics Department at UCLA. UCLA offers an introductory statistics course every summer, and it needs to understand what sorts of students are interested in this class. In particular, understanding whether the summer students are mostly first-year students (eager to complete their general education requirements) or seniors (who put off the class as long as they could) can help the department better plan its course offerings.

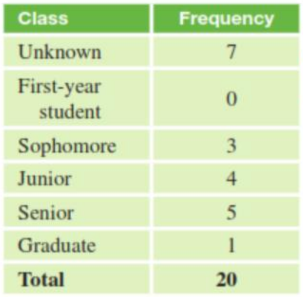
Table 2.3 shows data from a sample of students in an introductory course offered during the 2013 summer term at UCLA. The "unknown" students are probably not enrolled in any university (adult students taking the course for business reasons or high school students taking the class to get a head start).

Class is a categorical variable. Table 2.4 on the next page summarizes the distribution of this variable by showing us all of the values in our sample and the frequency with which each value appears. Note that we added a row for first-year students.

Two types of graphs that are commonly used to display the distribution of a sample of categorical data are bar charts and pie charts. Bar charts look, at first glance, very similar to histograms, but they have several important differences, as you will see.

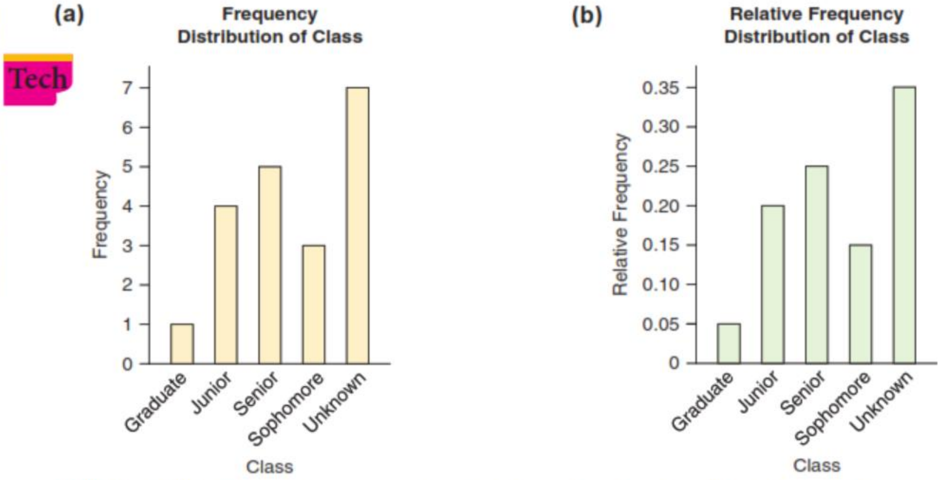
Student ID	Class
1	Senior
2	Junior
3	Unknown
4	Unknown
5	Senior
6	Graduate
7	Senior
8	Senior
9	Unknown
10	Unknown
11	Sophomore
12	Junior
13	Junior
14	Sophomore
15	Unknown
16	Senior
17	Unknown
18	Unknown

- **Bar Charts**



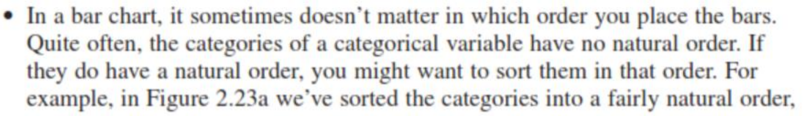
Class	Frequency
Unknown	7
First-year student	0
Sophomore	3
Junior	4
Senior	5
Graduate	1
Total	**20**

▲ **TABLE 2.4** Summary of classes for students in statistics.

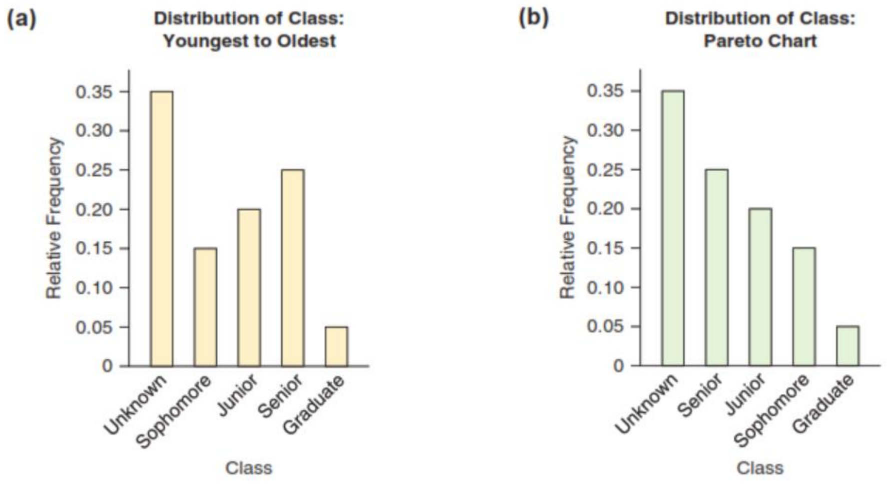


▲ **FIGURE 2.22** **(a)** Bar chart showing numbers of students in each class enrolled in an introductory statistics section. The largest "class" is the group made up of seven unknowns. First-year students are not shown because there were none in the data set. **(b)** The same information as shown in part (a), but now with relative frequencies. The unknowns are about 0.35 (35%) of the sample.

- **Bar Charts vs. Histograms**



- In a bar chart, it sometimes doesn't matter in which order you place the bars. Quite often, the categories of a categorical variable have no natural order. If they do have a natural order, you might want to sort them in that order. For example, in Figure 2.23a we've sorted the categories into a fairly natural order,



▶ **FIGURE 2.23** **(a)** Bar chart of classes using natural order. **(b)** Pareto chart of the same data. Categories are ordered with the largest frequency on the left and arranged so the frequencies decrease to the right.

- Another difference between histograms and bar charts is that in a bar chart, it doesn't matter how wide or narrow the bars are. The widths of the bars have no meaning.
- A final important difference is that a bar chart has gaps between the bars. This indicates that it is impossible to have observations between the categories. In a histogram, a gap indicates that no values were observed in the interval represented by the gap.

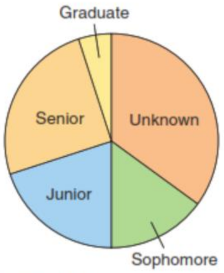
WHAT IS IT? ▶ A graphical summary for categorical data.

WHAT DOES IT DO? ▶ Shows a picture of the distribution of a categorical variable.

HOW DOES IT DO IT? ▶ Each category is represented by a bar. The height of the bar is proportional to the number of times that category occurs in the data set.

HOW IS IT USED? ▶ To see patterns of variation in categorical data. The categories can be presented in order of most frequent to least frequent, or they can be arranged in another meaningful order.

▪ Pie Charts



▲ **FIGURE 2.24** Pie chart showing the distribution of the categorical variable *Class* in a statistics course.

Pie charts are another popular format for displaying relative frequencies of data. A **pie chart** looks, as you would expect, like a pie. The pie is sliced into several pieces, and each piece represents a category of the variable. The area of the piece is proportional to the relative frequency of that category. The largest piece in the pie in Figure 2.24 belongs to the category "Unknown" and takes up about 35% of the total pie.

Some software will label each slice of the pie with the percentage occupied. This isn't always necessary, however, because a primary purpose of the pie chart is to help us judge how frequently categories occur relative to one another. For example, the pie chart in Figure 2.24 shows us that "Unknown" occupies a fairly substantial portion of the whole data set. Also, labeling each slice gets cumbersome and produces cluttered graphs if there are many categories.

WHAT IS IT? ▶ A graphical summary for categorical data.

WHAT DOES IT DO? ▶ Shows the proportion of observations that belong to each category.

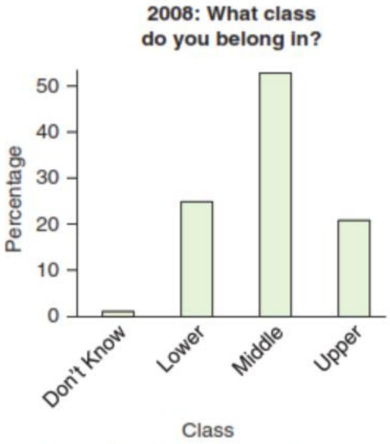
HOW DOES IT DO IT? ▶ Each category is represented by a wedge in the pie. The area of the wedge is proportional to the relative frequency of that category.

HOW IS IT USED? ▶ To understand which categories are most frequent and which are least frequent. Sometimes it is useful to label each wedge with the proportion of occurrence.

2.4 Summarizing Categorical Distributions

The concepts of shape, center, and spread that we used to summarize numerical distributions sometimes don't make sense for categorical distributions, because we can often order the categories any way we please. The center and shape would be different for every ordering of categories. However, we can still talk about typical outcomes and the variability in the sample.

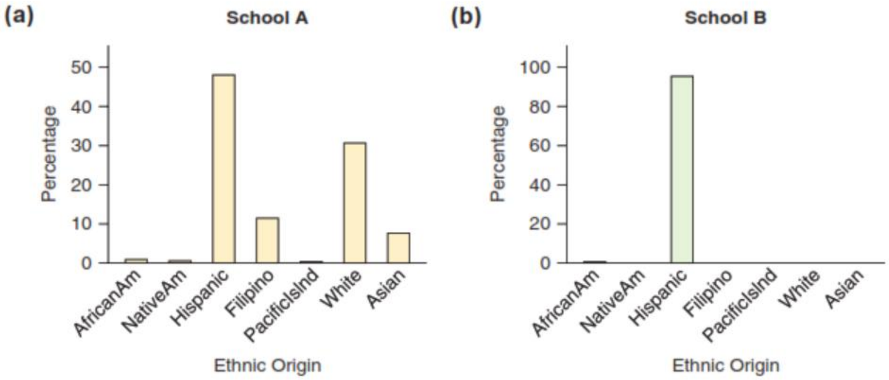
- ### The Mode



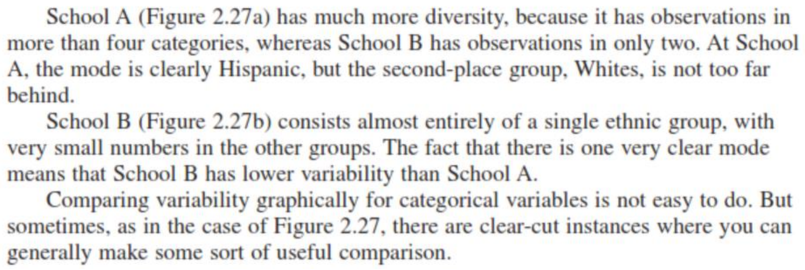
▲ **FIGURE 2.25** Percents of respondents who, in 2008, reported the economic class they felt they belonged to.

- ### Variability

When thinking about the variability of a categorical distribution, it is sometimes useful to think of the word diversity.



◀ **FIGURE 2.27** Percents of students at two Los Angeles schools who are identified with several ethnic groups. Which school has more ethnic variability?

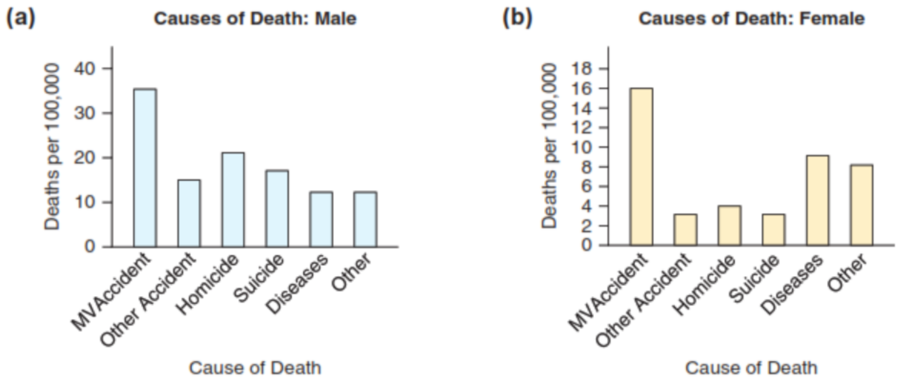


School A (Figure 2.27a) has much more diversity, because it has observations in more than four categories, whereas School B has observations in only two. At School A, the mode is clearly Hispanic, but the second-place group, Whites, is not too far behind.

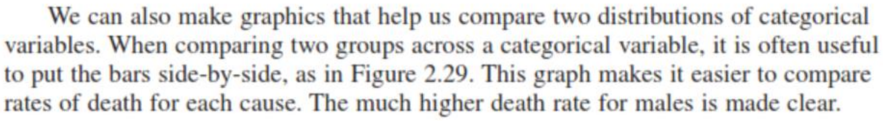
School B (Figure 2.27b) consists almost entirely of a single ethnic group, with very small numbers in the other groups. The fact that there is one very clear mode means that School B has lower variability than School A.

Comparing variability graphically for categorical variables is not easy to do. But sometimes, as in the case of Figure 2.27, there are clear-cut instances where you can generally make some sort of useful comparison.

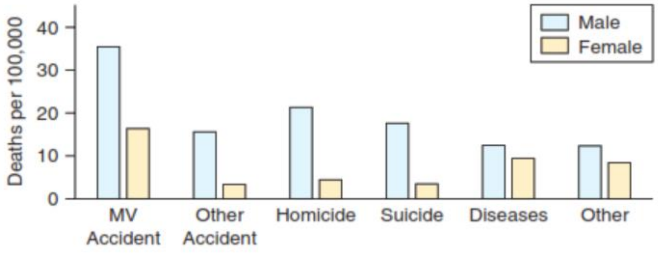
- When summarizing graphs of categorical data, report the mode or modes and describe the variability (diversity).



◀ **FIGURE 2.28** The number of deaths per 100,000 males **(a)** and females **(b)** for people 15–24 years old in a one-year period.



We can also make graphics that help us compare two distributions of categorical variables. When comparing two groups across a categorical variable, it is often useful to put the bars side-by-side, as in Figure 2.29. This graph makes it easier to compare rates of death for each cause. The much higher death rate for males is made clear.



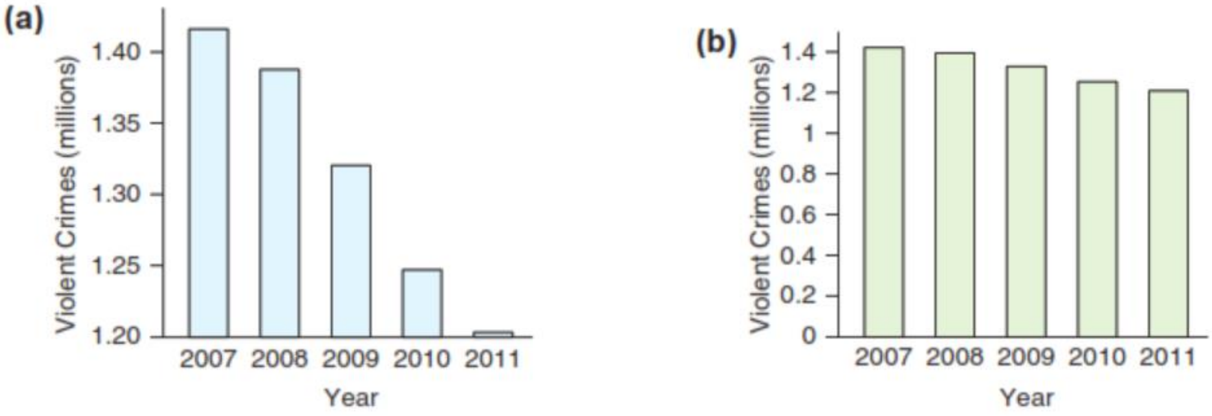
◀ **FIGURE 2.29** Death rates of males and females, graphed side by side.

2.5 Interpreting Graphs

- The first step in every investigation of data is to make an appropriate graph.
- Many analyses of data begin with visualizing the distribution of a variable, and this requires that you know whether the variable is numerical or categorical.
- When interpreting these graphics, you should pay attention to the center, spread, and shape.

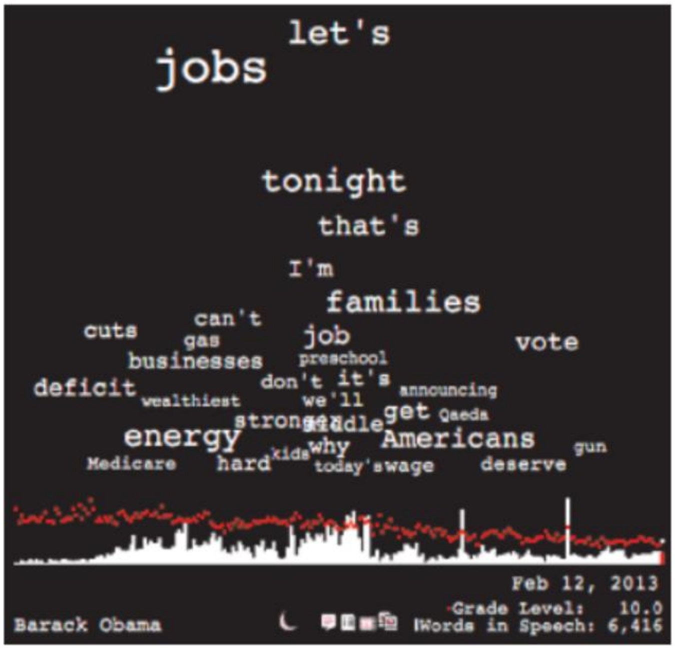
- **Misleading Graphs**

  A well-designed statistical graphic can help us discover patterns and trends and can communicate these patterns clearly to others. However, our eyes can play tricks on us, and manipulative people can take advantage of this to use graphs to give false impressions.



▲ **FIGURE 2.30** **(a)** This bar chart apparently shows a dramatic decline in the number of violent crimes since 2007. The origin for the vertical axis begins at about 1.20 million, not at 0. **(b)** This bar chart reports the same data as part (a), but here the vertical axis begins at the origin (0).

- **The Future of Statistical Graphics**
  - The Internet allows for a great variety of graphical displays of data that take us beyond simple visualizations of distributions.
  - Many statisticians, computer scientists, and designers are experimenting with new ways to visualize data. Most exciting is the rise of interactive displays.
  - The State of the Union Visualization, for example (http:// stateoftheunion.onetwothree.net), makes it possible to compare the content of State of the Union speeches.
  - Every U.S. president delivers a State of the Union address to Congress near the beginning of each year.
  - This interactive graphic enables users to compare words from different speeches and "drill down" to learn details about particular words or speeches.



◀ **FIGURE 2.32** President Obama's 2013 State of the Union speech, visualized as a "word cloud." This array shows us the most commonly used words, approximately where these words occurred in the speech, how often they occurred, and how unusual they are compared to the content of other State of the Union speeches. (Courtesy of Brad Borevitz, http://onetwothree.net. Used with permission.)

SUMMARY

- The first step in any statistical investigation is to make plots of the distributions of the data in your data set.
- You should identify whether the variables are numerical or categorical so that you can choose an appropriate graphical representation.
- If the variable is numerical, you can make a dotplot, histogram, or stemplot. Pay attention to the shape (Is it skewed or symmetric? Is it unimodal or multimodal?), to the center (What is a typical outcome?), and to the spread (How much variability is present?).
- You should also look for unusual features, such as outliers. Be aware that many of these terms are deliberately vague. You might think a particular observation is an outlier, but another person might not agree. That's okay, because the purpose isn't to determine whether such points are "outliers" but to indicate whether further investigation is needed.
- An outlier might, for example, be caused by a typing error someone made when entering the data.
- If you see a bimodal or multimodal distribution, ask yourself whether the data might contain two or more groups combined into the single graph.
- If the variable is categorical, you can make a bar chart, a Pareto chart (a bar chart with categories ordered from most frequent to least frequent), or a pie chart. Pay attention to the mode (or modes) and to the variability.