To: AP Statistics Students

Although summer is here, the math teachers at Notre Dame High School are already anticipating the next school year. We want to do everything possible to ensure that it will be very successful for you. It is important that the foundation for statistics is thoroughly understood. Therefore, the attached summer packet has been designed to introduce you to the course. It is required that you take the time to read the packet and complete all the exercises specified at the bottom of this page. **Your completed exercises will be collected the first day of school**, and will be graded as your first assignment. You will be tested on the material during the first week of classes.

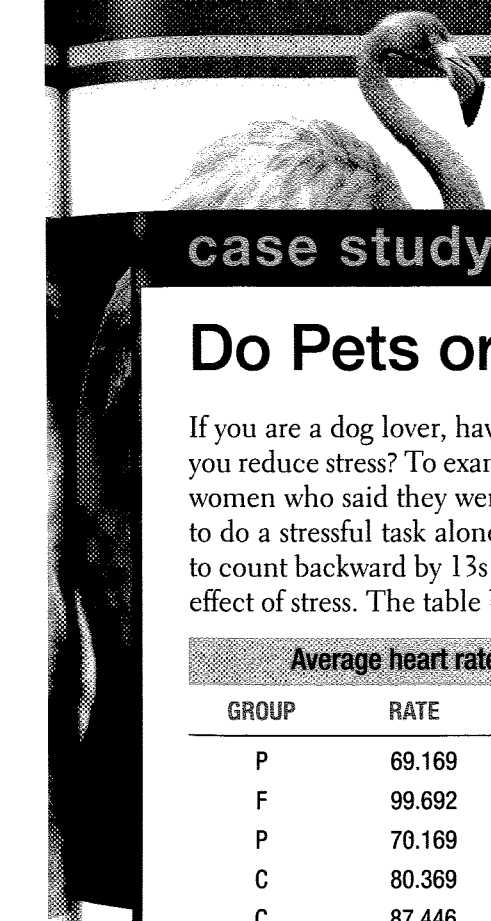
It is also required that you have your own graphing calculator that you bring to class on a daily basis. We have found that the TI-84+ calculator is the most helpful since it comes with a statistical package as powerful as most of the computer programs used in college statistics courses.

We hope you have an enjoyable and relaxing summer. We are looking forward to seeing you during the next school year. Bring your completed summer packet with you on the first day of class so we can immediately begin a fun and challenging semester of AP Statistics!

Exercises:

Pages 6-7, #1-7 odd; Pages 20-24, #9-25 odd, 26-34 all

Exploring Data



case study

Do Pets or Friends Help Reduce Stress?

If you are a dog lover, having your dog with you may reduce your stress level. Does having a friend with you reduce stress? To examine the effect of pets and friends in stressful situations, researchers recruited 45 women who said they were dog lovers. Fifteen women were assigned at random to each of three groups: to do a stressful task alone, with a good friend present, or with their dogs present. The stressful task was to count backward by 13s or 17s. The woman's average heart rate during the task was one measure of the effect of stress. The table below shows the data.[1]

Average heart rates during stress with a pet (P), with a friend (F), and for the control group (C)

GROUP	RATE	GROUP	RATE	GROUP	RATE	GROUP	RATE
P	69.169	P	68.862	C	84.738	C	75.477
F	99.692	C	87.231	C	84.877	C	62.646
P	70.169	P	64.169	P	58.692	P	70.077
C	80.369	C	91.754	P	79.662	F	88.015
C	87.446	C	87.785	P	69.231	F	81.600
P	75.985	F	91.354	C	73.277	F	86.985
F	83.400	F	100.877	C	84.523	F	92.492
F	102.154	C	77.800	C	70.877	P	72.262
P	86.446	P	97.538	F	89.815	P	65.446
F	80.277	P	85.000	F	98.200		
C	90.015	F	101.062	F	76.908		
C	99.046	F	97.046	P	69.538		

Based on the data, does it appear that the presence of a pet or friend reduces heart rate during a stressful task? In this chapter, you'll develop the tools to help answer this question.

Introduction Data Analysis: Making Sense of Data

WHAT YOU WILL LEARN **By the end of the section, you should be able to:**

- Identify the individuals and variables in a set of data.
- Classify variables as categorical or quantitative.

Statistics is the science of data. The volume of data available to us is overwhelming. For example, the Census Bureau's American Community Survey collects data from 3,000,000 housing units each year. Astronomers work with data on tens of millions of galaxies. The checkout scanners at Walmart's 10,000 stores in 27 countries record hundreds of millions of transactions every week.

In all these cases, the data are trying to tell us a story—about U.S. households, objects in space, or Walmart shoppers. To hear what the data are saying, we need to help them speak by organizing, displaying, summarizing, and asking questions. That's **data analysis**.

Individuals and Variables

Any set of data contains information about some group of **individuals**. The characteristics we measure on each individual are called **variables**.

DEFINITION: Individuals and variables

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

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

A high school'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, grade point average, homeroom, and grade level. In practice, any set of data is accompanied by background information that helps us understand the data. When you first meet a new data set, ask yourself the following questions:

1. *Who* are the individuals described by the data? How many individuals are there?
2. *What* are the variables? In what *units* are the variables recorded? Weights, for example, might be recorded in grams, pounds, thousands of pounds, or kilograms.

We could follow a newspaper reporter's lead and extend our list of questions to include *Why*, *When*, *Where*, and *How* were the data produced? For now, we'll focus on the first two questions.

Some variables, like gender and grade level, assign labels to individuals that place them 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 group of students, but it doesn't make sense to give an "average" gender.

DEFINITION: Categorical variable and quantitative variable

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

A **quantitative variable** takes numerical values for which it makes sense to find an average.

AP® EXAM TIP If you learn to distinguish categorical from quantitative variables now, it will pay big rewards later. You will be expected to analyze categorical and quantitative variables correctly on the AP® exam.

Not every variable that takes number values is quantitative. Zip code is one example. Although zip codes are numbers, it doesn't make sense to talk about the average zip code. In fact, zip codes place individuals (people or dwellings) into categories based on location. Some variables—such as gender, race, and occupation—are categorical by nature. Other categorical variables are created by grouping values of a quantitative variable into classes. For instance, we could classify people in a data set by age: 0–9, 10–19, 20–29, and so on.

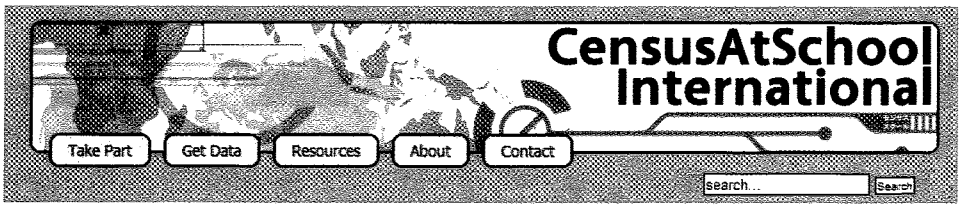
The proper method of analysis for a variable depends on whether it is categorical or quantitative. As a result, it is important to be able to distinguish these two types of variables. The type of data determines what kinds of graphs and which numerical summaries are appropriate.

EXAMPLE

Census at School

Data, individuals, and variables

CensusAtSchool is an international project that collects data about primary and secondary school students using surveys. Hundreds of thousands of students from Australia, Canada, New Zealand, South Africa, and the United Kingdom have taken part in the project since 2000. Data from the surveys are available at the project's Web site (www.censusatschool.com). We used the site's "Random Data Selector" to choose 10 Canadian students who completed the survey in a recent year. The table below displays the data.



Province	Gender	Language spoken	Handed	Height (cm)	Wrist circum. (mm)	Preferred communication
Saskatchewan	Male	1	Right	175	180	In person
Ontario	Female	1	Right	162.5	160	In person
Alberta	Male	1	Right	178	174	Facebook
Ontario	Male	2	Right	169	160	Cell phone
Ontario	Female	2	Right	166	65	In person
Nunavut	Male	1	Right	168.5	160	Text messaging
Ontario	Female	1	Right	166	165	Cell phone
Ontario	Male	4	Left	157.5	147	Text Messaging
Ontario	Female	2	Right	150.5	187	Text Messaging
Ontario	Female	1	Right	171	180	Text Messaging

There is at least one suspicious value in the data table. We doubt that the girl who is 166 cm tall really has a wrist circumference of 65 mm (about 2.6 inches). Always look to be sure the values make sense!

PROBLEM:

(a) Who are the individuals in this data set?

(b) What variables were measured? Identify each as categorical or quantitative.

(c) Describe the individual in the highlighted row.

We'll see in Chapter 4 why choosing at random, as we did in this example, is a good idea.

SOLUTION:

(a) The individuals are the 10 randomly selected Canadian students who participated in the CensusAtSchool survey.

(b) The seven variables measured are the province where the student lives (categorical), gender (categorical), number of languages spoken (quantitative), dominant hand (categorical), height (quantitative), wrist circumference (quantitative), and preferred communication method (categorical).

(c) This student lives in Ontario, is male, speaks four languages, is left-handed, is 157.5 cm tall (about 62 inches), has a wrist circumference of 147 mm (about 5.8 inches), and prefers to communicate via text messaging.

For Practice *Try Exercise* 3

Most data tables follow the format shown in the example—each row is an individual, and each column is a variable. Sometimes the individuals are called *cases*.

To make life simpler, we sometimes refer to "categorical data" or "quantitative data" instead of identifying the variable as categorical or quantitative.

A variable generally takes values that vary (hence the name "variable"!). Categorical variables sometimes have similar counts in each category and sometimes don't. For instance, we might have expected similar numbers of males and females in the CensusAtSchool data set. But we aren't surprised to see that most students are right-handed. Quantitative variables may take values that are very close together or values that are quite spread out. We call the pattern of variation of a variable its **distribution**.

DEFINITION: Distribution

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

Section 1.1 begins by looking at how to describe the distribution of a single categorical variable and then examines relationships between categorical variables. Sections 1.2 and 1.3 and all of Chapter 2 focus on describing the distribution of a quantitative variable. Chapter 3 investigates relationships between two quantitative variables. In each case, we begin with graphical displays, then add numerical summaries for a more complete description.

HOW TO EXPLORE DATA

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

CHECK YOUR UNDERSTANDING

Jake is a car buff who wants to find out more about the vehicles that students at his school drive. He gets permission to go to the student parking lot and record some data. Later, he does some research about each model of car on the Internet. Finally, Jake

makes a spreadsheet that includes each car's model, year, color, number of cylinders, gas mileage, weight, and whether it has a navigation system.

1. Who are the individuals in Jake's study?
2. What variables did Jake measure? Identify each as categorical or quantitative.

From Data Analysis to Inference

Sometimes, we're interested in drawing conclusions that go beyond the data at hand. That's the idea of **inference**. In the CensusAtSchool example, 9 of the 10 randomly selected Canadian students are right-handed. That's 90% of the *sample*. Can we conclude that 90% of the *population* of Canadian students who participated in CensusAtSchool are right-handed? No.

If another random sample of 10 students was selected, the percent who are right-handed might not be exactly 90%. Can we at least say that the actual population value is "close" to 90%? That depends on what we mean by "close."

The following Activity gives you an idea of how statistical inference works.

ACTIVITY Hiring discrimination—it just won't fly!

MATERIALS:

Bag with 25 beads (15 of one color and 10 of another) or 25 identical slips of paper (15 labeled "M" and 10 labeled "F") for each student or pair of students

An airline has just finished training 25 pilots—15 male and 10 female—to become captains. Unfortunately, only eight captain positions are available right now. Airline managers announce that they will use a lottery to determine which pilots will fill the available positions. The names of all 25 pilots will be written on identical slips of paper. The slips will be placed in a hat, mixed thoroughly, and drawn out one at a time until all eight captains have been identified.

A day later, managers announce the results of the lottery. Of the 8 captains chosen, 5 are female and 3 are male. Some of the male pilots who weren't selected suspect that the lottery was not carried out fairly. One of these pilots asks your statistics class for advice about whether to file a grievance with the pilots' union.

The key question in this possible discrimination case seems to be: *Is it plausible (believable) that these results happened just by chance?* To find out, you and your classmates will *simulate* the lottery process that airline managers said they used.

1. Mix the beads/slips thoroughly. Without looking, remove 8 beads/slips from the bag. Count the number of female pilots selected. Then return the beads/slips to the bag.
2. Your teacher will draw and label a number line for a class *dotplot*. On the graph, plot the number of females you got in Step 1.
3. Repeat Steps 1 and 2 if needed to get a total of at least 40 simulated lottery results for your class.
4. Discuss the results with your classmates. Does it seem believable that airline managers carried out a fair lottery? What advice would you give the male pilot who contacted you?
5. Would your advice change if the lottery had chosen 6 female (and 2 male) pilots? What about 7 female pilots? Explain.

Our ability to do inference is determined by how the data are produced. Chapter 4 discusses the two main methods of data production—sampling and experiments—and the types of conclusions that can be drawn from each. As the Activity illustrates, the logic of inference rests on asking, "What are the chances?" *Probability*, the study of chance behavior, is the topic of Chapters 5 through 7. We'll introduce the most common inference techniques in Chapters 8 through 12.

Introduction Summary

- A data set contains information about a number of **individuals.** Individuals may be people, animals, or things. For each individual, the data give values for one or more **variables.** A variable describes some characteristic of an individual, such as a person's height, gender, or salary.
- Some variables are **categorical** and others are **quantitative.** A categorical variable assigns a label that places each individual into one of several groups, such as male or female. A quantitative variable has numerical values that measure some characteristic of each individual, such as height in centimeters or salary in dollars.
- The **distribution** of a variable describes what values the variable takes and how often it takes them.

Introduction Exercises

The solutions to all exercises numbered in red are found in the Solutions Appendix, starting on page S-1.

1. **Protecting wood** How can we help wood surfaces resist weathering, especially when restoring historic wooden buildings? In a study of this question, researchers prepared wooden panels and then exposed them to the weather. Here are some of the variables recorded: type of wood (yellow poplar, pine, cedar); type of water repellent (solvent-based, water-based); paint thickness (millimeters); paint color (white, gray, light blue); weathering time (months). Identify each variable as categorical or quantitative.

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. Here are some of the variables recorded: 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). Identify each as categorical or quantitative.

3. **A class survey** Here is a small part of the data set that describes the students in an AP® Statistics class. The data come from anonymous responses to a questionnaire filled out on the first day of class.

pg 3

Gender	Hand	Height (in.)	Homework time (min)	Favorite music	Pocket change (cents)
F	L	65	200	Hip-hop	50
M	L	72	30	Country	35
M	R	62	95	Rock	35
F	L	64	120	Alternative	0
M	R	63	220	Hip-hop	0
F	R	58	60	Alternative	76
F	R	67	150	Rock	215

(a) What individuals does this data set describe?

(b) What variables were measured? Identify each as categorical or quantitative.

(c) Describe the individual in the highlighted row.

4. **Coaster craze** Many people like to ride roller coasters. Amusement parks try to increase attendance by building exciting new coasters. The following table displays data on several roller coasters that were opened in a recent year.[2]

Roller coaster	Type	Height (ft)	Design	Speed (mph)	Duration (s)
Wild Mouse	Steel	49.3	Sit down	28	70
Terminator	Wood	95	Sit down	50.1	180
Manta	Steel	140	Flying	56	155
Prowler	Wood	102.3	Sit down	51.2	150
Diamondback	Steel	230	Sit down	80	180

(a) What individuals does this data set describe?

(b) What variables were measured? Identify each as categorical or quantitative.

(c) Describe the individual in the highlighted row.

5. **Ranking colleges** Popular magazines rank colleges and universities on their "academic quality" in serving undergraduate students. Describe two categorical variables and two quantitative variables that you might record for each institution.

6. **Students and TV** You are preparing to study the television-viewing habits of high school students. Describe two categorical variables and two quantitative variables that you might record for each student.

Multiple choice: Select the best answer.
Exercises 7 and 8 refer to the following setting. At the Census Bureau Web site www.census.gov, you can view detailed data collected by the American Community Survey. The following table includes data for 10 people chosen at random from the more than 1 million people in households contacted by the survey. "School" gives the highest level of education completed.

Weight (lb)	Age (yr)	Travel to work (min)	School	Gender	Income last year ($)
187	66	0	Ninth grade	1	24,000
158	66	n/a	High school grad	2	0
176	54	10	Assoc. degree	2	11,900
339	37	10	Assoc. degree	1	6000
91	27	10	Some college	2	30,000
155	18	n/a	High school grad	2	0
213	38	15	Master's degree	2	125,000
194	40	0	High school grad	1	800
221	18	20	High school grad	1	2500
193	11	n/a	Fifth grade	1	0

7. The individuals in this data set are

(a) households.

(b) people.

(c) adults.

(d) 120 variables.

(e) columns.

8. This data set contains

(a) 7 variables, 2 of which are categorical.

(b) 7 variables, 1 of which is categorical.

(c) 6 variables, 2 of which are categorical.

(d) 6 variables, 1 of which is categorical.

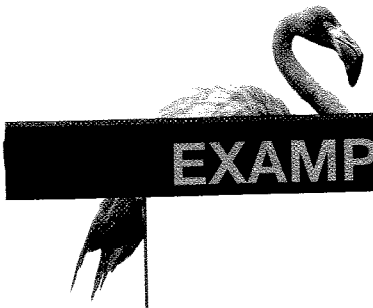
(e) None of these.

1.1 Analyzing Categorical Data

WHAT YOU WILL LEARN **By the end of the section, you should be able to:**

- Display categorical data with a bar graph. Decide if it would be appropriate to make a pie chart.
- Identify what makes some graphs of categorical data deceptive.
- Calculate and display the marginal distribution of a categorical variable from a two-way table.
- Calculate and display the conditional distribution of a categorical variable for a particular value of the other categorical variable in a two-way table.
- Describe the association between two categorical variables by comparing appropriate conditional distributions.

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 within each category. Here's an example.



EXAMPLE

Radio Station Formats

Distribution of a categorical variable

The radio audience rating service Arbitron places U.S radio stations into categories that describe the kinds of programs they broadcast. Here are two different tables showing the distribution of station formats in a recent year:[3]

Frequency table

Format	Count of stations
Adult contemporary	1556
Adult standards	1196
Contemporary hit	569
Country	2066
News/Talk/Information	2179
Oldies	1060
Religious	2014
Rock	869
Spanish language	750
Other formats	1579
Total	**13,838**

Relative frequency table

Format	Percent of stations
Adult contemporary	11.2
Adult standards	8.6
Contemporary hit	4.1
Country	14.9
News/Talk/Information	15.7
Oldies	7.7
Religious	14.6
Rock	6.3
Spanish language	5.4
Other formats	11.4
Total	**99.9**

In this case, the *individuals* are the radio stations and the *variable* being measured is the kind of programming that each station broadcasts. The table on the left, which we call a **frequency table**, displays the counts (*frequencies*) of stations in each format category. On the right, we see a **relative frequency table** of the data that shows the percents (*relative frequencies*) of stations in each format category.

It's a good idea to check data for consistency. The counts should add to 13,838, the total number of stations. They do. The percents should add to 100%. In fact, they add to 99.9%. What happened? Each percent is rounded to the nearest tenth. The exact percents would add to 100, but the rounded percents only come close. This is **roundoff error.** Roundoff errors don't point to mistakes in our work, just to the effect of rounding off results.

Bar Graphs and Pie Charts

Columns of numbers take time to read. You can use a **pie chart** or a **bar graph** to display the distribution of a categorical variable more vividly. Figure 1.1 illustrates both displays for the distribution of radio stations by format.

Pie charts show the distribution of a categorical variable as a "pie" whose slices are sized by the counts or percents for the categories. A pie chart must include all the categories that make up a whole. In the radio station example, we needed the "Other formats" category to complete the whole (all radio stations) and allow us to make a pie chart. Use a pie chart only when you want to emphasize each

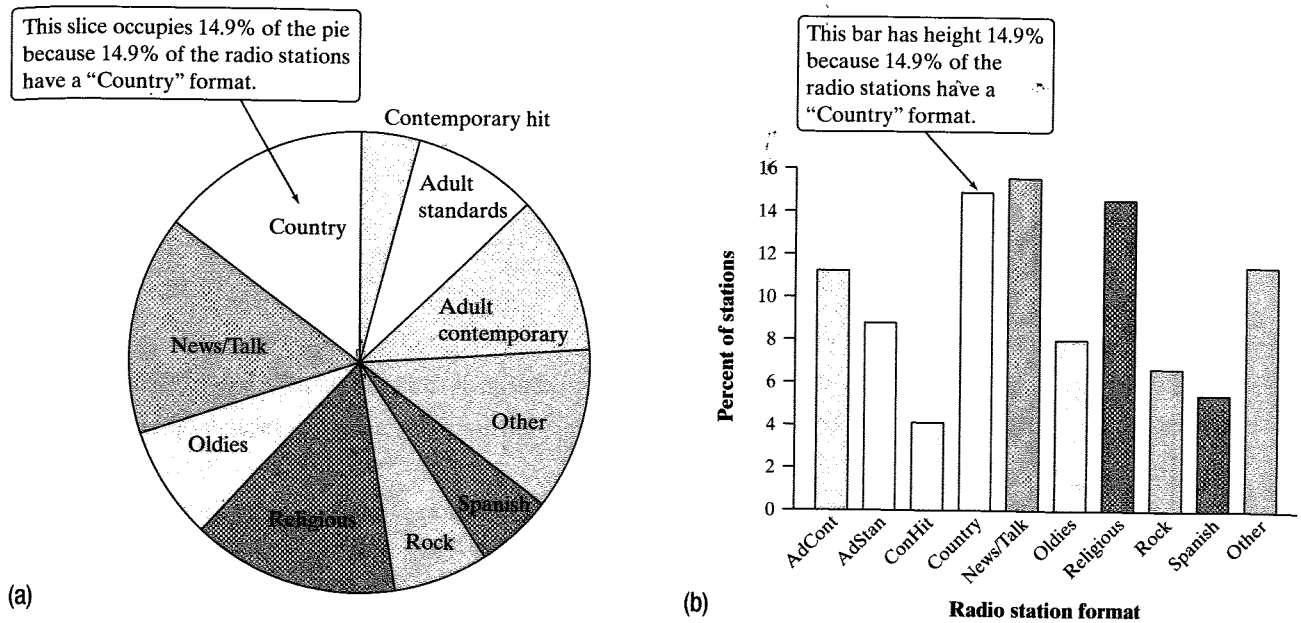
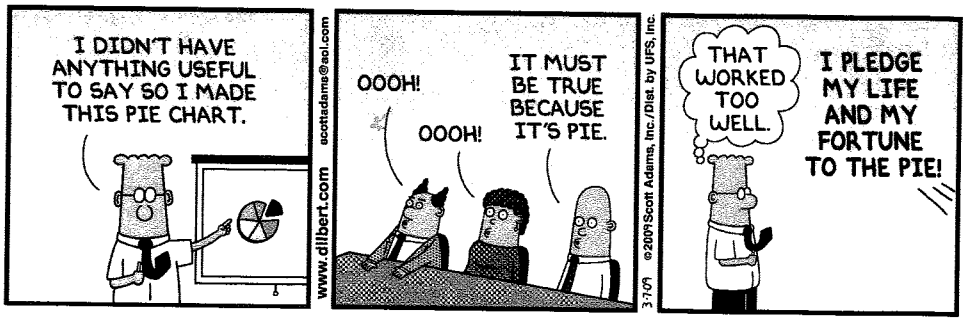


FIGURE 1.1 (a) Pie chart and (b) bar graph of U.S. radio stations by format.

category's relation to the whole. Pie charts are awkward to make by hand, but technology will do the job for you.



Bar graphs are also called *bar charts*.

Bar graphs represent each category as a bar. The bar heights show the category counts or percents. Bar graphs are easier to make than pie charts and are also easier to read. To convince yourself, try to use the pie chart in Figure 1.1 to estimate the percent of radio stations that have an "Oldies" format. Now look at the bar graph—it's easy to see that the answer is about 8%.

Bar graphs are also more flexible than pie charts. Both graphs can display the distribution of a categorical variable, but a bar graph can also compare any set of quantities that are measured in the same units.

EXAMPLE

Who Owns an MP3 Player?

Choosing the best graph to display the data

Portable MP3 music players, such as the Apple iPod, are popular—but not equally popular with people of all ages. Here are the percents of people in various age groups who own a portable MP3 player, according to an Arbitron survey of 1112 randomly selected people.[4]

Age group (years)	Percent owning an MP3 player
12 to 17	54
18 to 24	30
25 to 34	30
35 to 54	13
55 and older	5

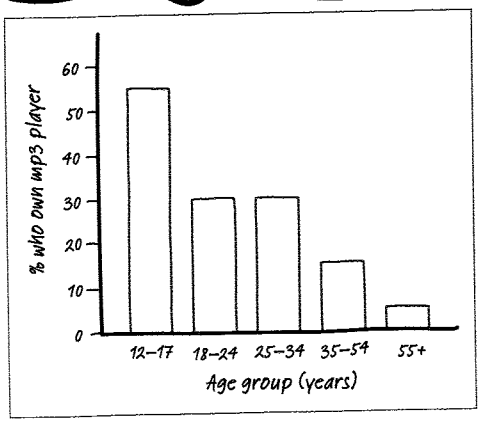


FIGURE 1.2 Bar graph comparing the percents of several age groups who own portable MP3 players.

PROBLEM:

(a) Make a well-labeled bar graph to display the data. Describe what you see.

(b) Would it be appropriate to make a pie chart for these data? Explain.

SOLUTION:

(a) We start by labeling the axes: age group goes on the horizontal axis, and percent who own an MP3 player goes on the vertical axis. For the vertical scale, which is measured in percents, we'll start at 0 and go up to 60, with tick marks for every 10. Then for each age category, we draw a bar with height corresponding to the percent of survey respondents who said they have an MP3 player. Figure 1.2 shows the completed bar graph. It appears that MP3 players are more popular among young people and that their popularity generally decreases as the age category increases.

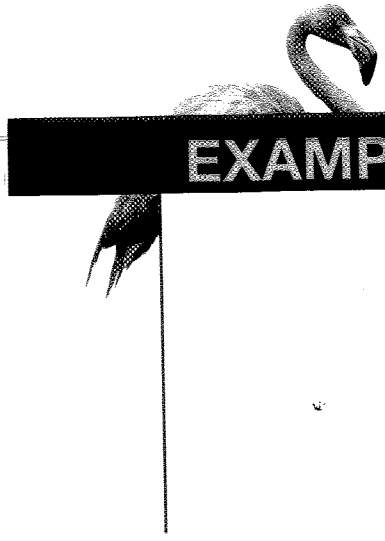
(b) Making a pie chart to display these data is not appropriate because each percent in the table refers to a different age group, not to parts of a single whole.

For Practice *Try Exercise* 15

Graphs: Good and Bad

Bar graphs compare several quantities by comparing the heights of bars that represent the quantities. Our eyes, however, react to the *area* of the bars as well as to their height. When all bars have the same width, the area (width × height) varies in proportion to the height, and our eyes receive the right impression. When you draw a bar graph, make the bars equally wide.

Artistically speaking, bar graphs are a bit dull. It is tempting to replace the bars with pictures for greater eye appeal. Don't do it! The following example shows why.

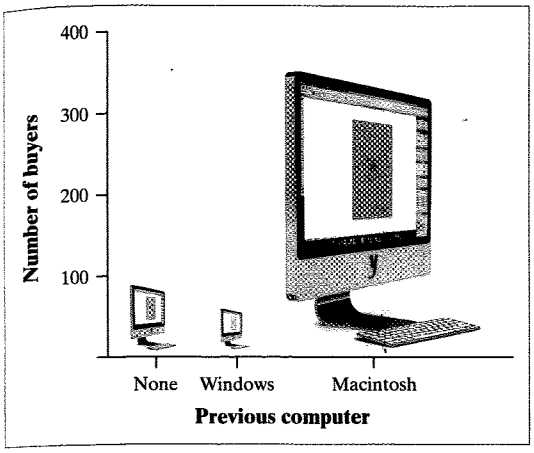


EXAMPLE

Who Buys iMacs?

Beware the pictograph!

When Apple, Inc., introduced the iMac, the company wanted to know whether this new computer was expanding Apple's market share. Was the iMac mainly being bought by previous Macintosh owners, or was it being purchased by first-time computer buyers and by previous PC users who were switching over? To find out, Apple hired a firm to conduct a survey of 500 iMac customers. Each customer was categorized as a new computer purchaser, a previous PC owner, or a previous Macintosh owner. The table summarizes the survey results.[5]

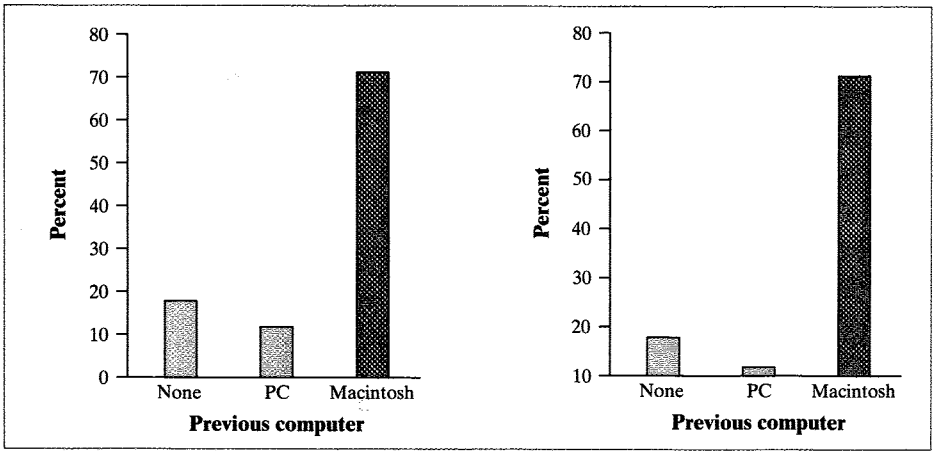


Previous ownership	Count	Percent (%)
None	85	17.0
PC	60	12.0
Macintosh	355	71.0
Total	**500**	**100.0**

PROBLEM:

(a) Here's a clever graph of the data that uses pictures instead of the more traditional bars. How is this graph misleading?

(b) Two possible bar graphs of the data are shown below. Which one could be considered deceptive? Why?



SOLUTION:

(a) Although the heights of the pictures are accurate, our eyes respond to the area of the pictures. The pictograph makes it seem like the percent of iMac buyers who are former Mac owners is at least ten times higher than either of the other two categories, which isn't the case.

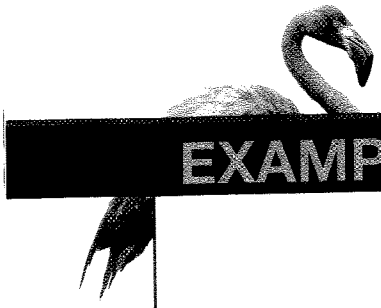
(b) The bar graph on the right is misleading. By starting the vertical scale at 10 instead of 0, it looks like the percent of iMac buyers who previously owned a PC is less than half the percent who are first-time computer buyers. We get a distorted impression of the relative percents in the three categories.

For Practice *Try Exercise* 17

There are two important lessons to be learned from this example: *(1) beware the pictograph, and (2) watch those scales.*

Two-Way Tables and Marginal Distributions

We have learned some techniques for analyzing the distribution of a single categorical variable. What do we do when a data set involves two categorical variables? We begin by examining the counts or percents in various categories for one of the variables. Here's an example to show what we mean.



EXAMPLE I'm Gonna Be Rich!

Two-way tables

A survey of 4826 randomly selected young adults (aged 19 to 25) asked, "What do you think the chances are you will have much more than a middle-class income at age 30?" The table below shows the responses.[6]

Young adults by gender and chance of getting rich

	Gender		
Opinion	**Female**	**Male**	**Total**
Almost no chance	96	98	**194**
Some chance but probably not	426	286	**712**
A 50-50 chance	696	720	**1416**
A good chance	663	758	**1421**
Almost certain	486	597	**1083**
Total	**2367**	**2459**	**4826**

This is a **two-way table** because it describes two categorical variables, gender and opinion about becoming rich. Opinion is the *row variable* because each row in the table describes young adults who held one of the five opinions about their chances. Because the opinions have a natural order from "Almost no chance" to "Almost certain," the rows are also in this order. Gender is the *column variable*. The entries in the table are the counts of individuals in each opinion-by-gender class.

How can we best grasp the information contained in the two-way table above? First, *look at the distribution of each variable separately.* The distribution of a categorical variable says how often each outcome occurred. The "Total" column at the right of the table contains the totals for each of the rows. These row totals give the distribution of opinions about becoming rich in the entire group of 4826 young adults: 194 thought that they had almost no chance, 712 thought they had just some chance, and so on. (If the row and column totals are missing, the first thing to do in studying a two-way table is to calculate them.) The distributions of opinion alone and gender alone are called **marginal distributions** because they appear at the right and bottom margins of the two-way table.

DEFINITION: Marginal distribution

The **marginal distribution** of one of the categorical variables in a two-way table of counts is the distribution of values of that variable among all individuals described by the table.

Percents are often more informative than counts, especially when we are comparing groups of different sizes. We can display the marginal distribution of opinions in percents by dividing each row total by the table total and converting to a percent. For instance, the percent of these young adults who think they are almost certain to be rich by age 30 is

$$\frac{\text{almost certain total}}{\text{table total}} = \frac{1083}{4826} = 0.224 = 22.4\%$$

EXAMPLE

I'm Gonna Be Rich!

Examining a marginal distribution

PROBLEM:

(a) Use the data in the two-way table to calculate the marginal distribution (in percents) of opinions.

(b) Make a graph to display the marginal distribution. Describe what you see.

SOLUTION:

(a) We can do four more calculations like the one shown above to obtain the marginal distribution of opinions in percents. Here is the complete distribution.

Response	Percent
Almost no chance	$\frac{194}{4826} = 4.0\%$
Some chance	$\frac{712}{4826} = 14.8\%$
A 50–50 chance	$\frac{1416}{4826} = 29.3\%$
A good chance	$\frac{1421}{4826} = 29.4\%$
Almost certain	$\frac{1083}{4826} = 22.4\%$

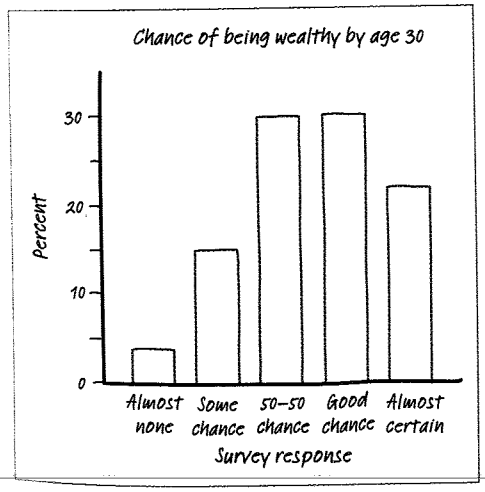


FIGURE 1.3 Bar graph showing the marginal distribution of opinion about chance of being rich by age 30.

(b) Figure 1.3 is a bar graph of the distribution of opinion among these young adults. It seems that many young adults are optimistic about their future income. Over 50% of those who responded to the survey felt that they had "a good chance" or were "almost certain" to be rich by age 30.

For Practice *Try Exercise* 19

Each marginal distribution from a two-way table is a distribution for a single categorical variable. As we saw earlier, we can use a bar graph or a pie chart to display such a distribution.

CHECK YOUR UNDERSTANDING

A random sample of 415 children aged 9 to 17 from the United Kingdom and the United States who completed a CensusAtSchool survey in a recent year was selected. Each student's country of origin was recorded along with which superpower they would most like to have: the ability to fly, ability to freeze time, invisibility, superstrength, or telepathy (ability to read minds). The data are summarized in the table.[7]

	Country	
Superpower	**U.K.**	**U.S.**
Fly	54	45
Freeze time	52	44
Invisibility	30	37
Superstrength	20	23
Telepathy	44	66

1. Use the two-way table to calculate the marginal distribution (in percents) of superpower preferences.
2. Make a graph to display the marginal distribution. Describe what you see.

Relationships between Categorical Variables: Conditional Distributions

The two-way table contains much more information than the two marginal distributions of opinion alone and gender alone. *Marginal distributions tell us nothing about the relationship between two variables.* To describe a relationship between two categorical variables, we must calculate some well-chosen percents from the counts given in the body of the table.

Young adults by gender and chance of getting rich

	Gender		
Opinion	**Female**	**Male**	**Total**
Almost no chance	96	98	**194**
Some chance but probably not	426	286	**712**
A 50-50 chance	696	720	**1416**
A good chance	663	758	**1421**
Almost certain	486	597	**1083**
Total	**2367**	**2459**	**4826**

Conditional distribution of opinion among women

Response	Percent
Almost no chance	$\frac{96}{2367} = 4.1\%$
Some chance	$\frac{426}{2367} = 18.0\%$
A 50-50 chance	$\frac{696}{2367} = 29.4\%$
A good chance	$\frac{663}{2367} = 28.0\%$
Almost certain	$\frac{486}{2367} = 20.5\%$

We can study the opinions of women alone by looking only at the "Female" column in the two-way table. To find the percent of *young women* who think they are almost certain to be rich by age 30, divide the count of such women by the total number of women, the column total:

$$\frac{\text{women who are almost certain}}{\text{column total}} = \frac{486}{2367} = 0.205 = 20.5\%$$

Doing this for all five entries in the "Female" column gives the **conditional distribution** of opinion among women. See the table in the margin. We use the term "conditional" because this distribution describes only young adults who satisfy the condition that they are female.

DEFINITION: Conditional distribution

A **conditional distribution** of a variable describes the values of that variable among individuals who have a specific value of another variable. There is a separate conditional distribution for each value of the other variable.

Now let's examine the men's opinions.

EXAMPLE

I'm Gonna Be Rich!

Calculating a conditional distribution

PROBLEM: Calculate the conditional distribution of opinion among the young men.

SOLUTION: To find the percent of *young men* who think they are almost certain to be rich by age 30, divide the count of such men by the total number of men, the column total:

$$\frac{\text{men who are almost certain}}{\text{column total}} = \frac{597}{2459} = 24.3\%$$

If we do this for all five entries in the "Male" column, we get the conditional distribution shown in the table.

Conditional distribution of opinion among men

Response	Percent
Almost no chance	$\frac{98}{2459} = 4.0\%$
Some chance	$\frac{286}{2459} = 11.6\%$
A 50-50 chance	$\frac{720}{2459} = 29.3\%$
A good chance	$\frac{758}{2459} = 30.8\%$
Almost certain	$\frac{597}{2459} = 24.3\%$

For Practice *Try Exercise* 21

There are *two sets* of conditional distributions for any two-way table: one for the column variable and one for the row variable. So far, we have looked at the conditional distributions of opinion for the two genders. We could also examine the five conditional distributions of gender, one for each of the five opinions, by looking separately at the rows in the original two-way table. For instance, the conditional distribution of gender among those who responded "Almost certain" is

Female	Male
$\frac{486}{1083} = 44.9\%$	$\frac{597}{1083} = 55.1\%$

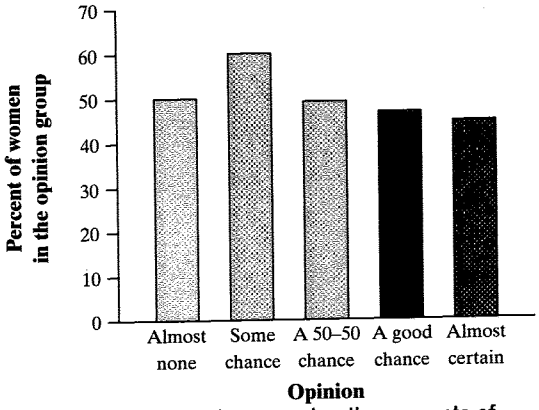


FIGURE 1.4 Bar graph comparing the percents of females among those who hold each opinion about their chance of being rich by age 30.

That is, of the young adults who said they were almost certain to be rich by age 30, 44.9% were female and 55.1% were male.

Because the variable "gender" has only two categories, comparing the five conditional distributions amounts to comparing the percents of women among young adults who hold each opinion. Figure 1.4 makes this comparison in a bar graph. The bar heights do *not* add to 100%, because each bar represents a different group of people.

THINK ABOUT IT

Which conditional distributions should we compare? Our goal all along has been to analyze the relationship between gender and opinion about chances of becoming rich for these young adults. We started by examining the conditional distributions of opinion for males and females. Then we looked at the conditional distributions of gender for each of the five opinion categories. Which of these two gives us the information we want? Here's a hint: think about whether changes in one variable might help explain changes in the other. In this case, it seems reasonable to think that gender might influence young adults' opinions about their chances of getting rich. To see whether the data support this idea, we should compare the conditional distributions of opinion for women and men.

Software will calculate conditional distributions for you. Most programs allow you to choose which conditional distributions you want to compute.

1. TECHNOLOGY CORNER ANALYZING TWO-WAY TABLES

Figure 1.5 presents the two conditional distributions of opinion, for women and for men, and also the marginal distribution of opinion for all of the young adults. The distributions agree (up to rounding) with the results in the last two examples.

Session

	Female	Male	All
A: Almost no chance	96	98	194
	4.06	3.99	4.02
B: Some chance but probably not	426	286	712
	18.00	11.63	14.75
C: A 50-50 chance	696	720	1416
	29.40	29.28	29.34
D: A good chance	663	758	1421
	28.01	30.83	29.44
E: Almost certain	486	597	1083
	20.53	24.28	22.44
All	2367	2459	4826
	100.00	100.00	100.00

Cell Contents: Count
% of Column

FIGURE 1.5 Minitab output for the two-way table of young adults by gender and chance of being rich, along with each entry as a percent of its column total. The "Female" and "Male" columns give the conditional distributions of opinion for women and men, and the "All" column shows the marginal distribution of opinion for all these young adults.

Putting It All Together: Relationships Between Categorical Variables

Now it's time to complete our analysis of the relationship between gender and opinion about chances of becoming rich later in life.

EXAMPLE Women's and Men's Opinions

Conditional distributions and relationships

PROBLEM: **Based on the survey data, can we conclude that young men and women differ in their opinions about the likelihood of future wealth? Give appropriate evidence to support your answer.**

SOLUTION: **We suspect that gender might influence a young adult's opinion about the chance of getting rich. So we'll compare the conditional distributions of response for men alone and for women alone.**

Response	Percent of Females	Percent of Males
Almost no chance	$\frac{96}{2367} = 4.1\%$	$\frac{98}{2459} = 4.0\%$
Some chance	$\frac{426}{2367} = 18.0\%$	$\frac{286}{2459} = 11.6\%$
A 50-50 chance	$\frac{696}{2367} = 29.4\%$	$\frac{720}{2459} = 29.3\%$
A good chance	$\frac{663}{2367} = 28.0\%$	$\frac{758}{2459} = 30.8\%$
Almost certain	$\frac{486}{2367} = 20.5\%$	$\frac{597}{2459} = 24.3\%$

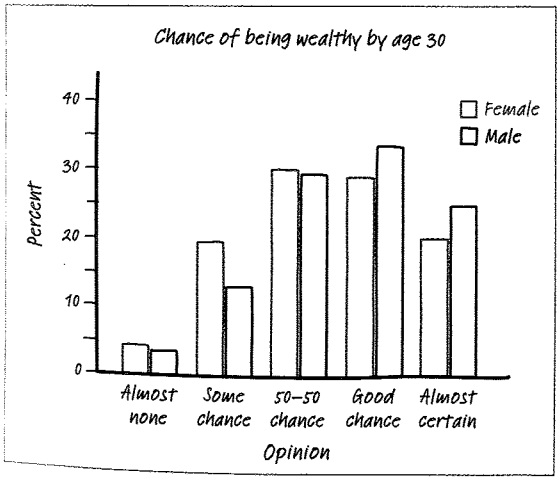


FIGURE 1.6 Side-by-side bar graph comparing the opinions of males and females.

We'll make a side-by-side bar graph to compare the opinions of males and females. Figure 1.6 displays the completed graph.

Based on the sample data, men seem somewhat more optimistic about their future income than women. Men were less likely to say that they have "some chance but probably not" than women (11.6% vs. 18.0%). Men were more likely to say that they have "a good chance" (30.8% vs. 28.0%) or are "almost certain" (24.3% vs. 20.5%) to have much more than a middle-class income by age 30 than women were.

For Practice *Try Exercise* 25

We could have used a **segmented bar graph** to compare the distributions of male and female responses in the previous example. Figure 1.7 shows the completed graph. Each bar has five segments—one for each of the opinion categories. It's fairly difficult to compare the percents of males and females in each category because the "middle" segments in the two bars start at different locations on the vertical axis. The side-by-side bar graph in Figure 1.6 makes comparison easier.

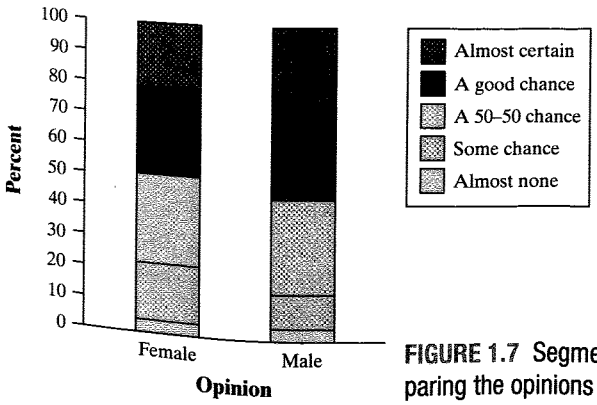


FIGURE 1.7 Segmented bar graph comparing the opinions of males and females.

Both graphs provide evidence of an **association** between gender and opinion about future wealth in this sample of young adults. Men more often rated their chances of becoming rich in the two highest categories; women said "some chance but probably not" much more frequently.

DEFINITION: Association

We say that there is an association between two variables if knowing the value of one variable helps predict the value of the other. If knowing the value of one variable does not help you predict the value of the other, then there is no association between the variables.

Can we say that there is an association between gender and opinion in the *population* of young adults? Making this determination requires formal inference, which will have to wait a few chapters.

THINK ABOUT IT

What does "no association" mean? Figure 1.6 (page 17) suggests an association between gender and opinion about future wealth for young adults. Knowing that a young adult is male helps us predict his opinion: he is more likely than a female to say "a good chance" or "almost certain." What would the graph look like if there was *no* association between the two variables? In that case, knowing that a young adult is male would not help us predict his opinion. He would be no more or less likely than a female to say "a good chance" or "almost certain" or any of the other possible responses. That is, the conditional distributions of opinion about becoming rich would be the *same* for males and females. The segmented bar graphs for the two genders would look the same, too.

CHECK YOUR UNDERSTANDING

Let's complete our analysis of the data on superpower preferences from the previous Check Your Understanding (page 14). Here is the two-way table of counts once again.

	Country	
Superpower	**U.K.**	**U.S.**
Fly	54	45
Freeze time	52	44
Invisibility	30	37
Superstrength	20	23
Telepathy	44	66

1. Find the conditional distributions of superpower preference among students from the United Kingdom and the United States.
2. Make an appropriate graph to compare the conditional distributions.
3. Is there an association between country of origin and superpower preference? Give appropriate evidence to support your answer.

There's one caution that we need to offer: *even a strong association between two categorical variables can be influenced by other variables lurking in the background.* The Data Exploration that follows gives you a chance to explore this idea using a famous (or infamous) data set.

DATA EXPLORATION A Titanic disaster

In 1912 the luxury liner *Titanic*, on its first voyage across the Atlantic, struck an iceberg and sank. Some passengers got off the ship in lifeboats, but many died. The two-way table below gives information about adult passengers who lived and who died, by class of travel.

	Class of Travel		
Survival status	**First class**	**Second class**	**Third class**
Lived	197	94	151
Died	122	167	476

Here's another table that displays data on survival status by gender and class of travel.

	Class of Travel					
	First class		**Second class**		**Third class**	
Survival status	**Female**	**Male**	**Female**	**Male**	**Female**	**Male**
Lived	140	57	80	14	76	75
Died	4	118	13	154	89	387

The movie *Titanic*, starring Leonardo DiCaprio and Kate Winslet, suggested the following:

- First-class passengers received special treatment in boarding the lifeboats, while some other passengers were prevented from doing so (especially third-class passengers).
- Women and children boarded the lifeboats first, followed by the men.

1. What do the data tell us about these two suggestions? Give appropriate graphical and numerical evidence to support your answer.

2. How does gender affect the relationship between class of travel and survival status? Explain.

Section 1.1 Summary

- The distribution of a categorical variable lists the categories and gives the count (**frequency**) or percent (**relative frequency**) of individuals that fall within each category.
- **Pie charts** and **bar graphs** display the distribution of a categorical variable. Bar graphs can also compare any set of quantities measured in the same units. When examining any graph, ask yourself, "What do I see?"
- A **two-way table** of counts organizes data about two categorical variables measured for the same set of individuals. Two-way tables are often used to summarize large amounts of information by grouping outcomes into categories.

- The row totals and column totals in a two-way table give the **marginal distributions** of the two individual variables. It is clearer to present these distributions as percents of the table total. Marginal distributions tell us nothing about the relationship between the variables.
- There are two sets of **conditional distributions** for a two-way table: the distributions of the row variable for each value of the column variable, and the distributions of the column variable for each value of the row variable. You may want to use a **side-by-side bar graph** (or possibly a **segmented bar graph**) to display conditional distributions.
- There is an **association** between two variables if knowing the value of one variable helps predict the value of the other. To see whether there is an association between two categorical variables, compare an appropriate set of conditional distributions. Remember that even a strong association between two categorical variables can be influenced by other variables.

1.1 TECHNOLOGY CORNER

1. Analyzing two-way tables page 16

Section 1.1 Exercises

9. **Cool car colors** The most popular colors for cars and light trucks change over time. Silver passed green in 2000 to become the most popular color worldwide, then gave way to shades of white in 2007. Here is the distribution of colors for vehicles sold in North America in 2011.[8]

Color	Percent of vehicles
White	23
Black	18
Silver	16
Gray	13
Red	10
Blue	9
Brown/beige	5
Yellow/gold	3
Green	2

(a) What percent of vehicles had colors other than those listed?

(b) Display these data in a bar graph. Be sure to label your axes.

(c) Would it be appropriate to make a pie chart of these data? Explain.

10. **Spam** Email spam is the curse of the Internet. Here is a compilation of the most common types of spam:[9]

Type of spam	Percent
Adult	19
Financial	20
Health	7
Internet	7
Leisure	6
Products	25
Scams	9
Other	??

(a) What percent of spam would fall in the "Other" category?

(b) Display these data in a bar graph. Be sure to label your axes.

(c) Would it be appropriate to make a pie chart of these data? Explain.

11. **Birth days** Births are not evenly distributed across the days of the week. Here are the average numbers of babies born on each day of the week in the United States in a recent year:[10]

Day	Births
Sunday	7374
Monday	11,704
Tuesday	13,169
Wednesday	13,038
Thursday	13,013
Friday	12,664
Saturday	8459

(a) Present these data in a well-labeled bar graph. Would it also be correct to make a pie chart?

(b) Suggest some possible reasons why there are fewer births on weekends.

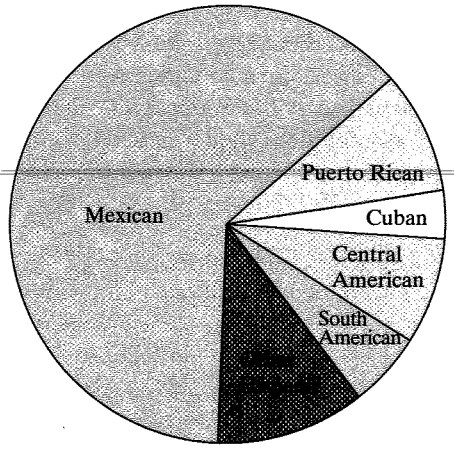
12. **Deaths among young people** Among persons aged 15 to 24 years in the United States, the leading causes of death and number of deaths in a recent year were as follows: accidents, 12,015; homicide, 4651; suicide, 4559; cancer, 1594; heart disease, 984; congenital defects, 401.[11]

(a) Make a bar graph to display these data.

(b) To make a pie chart, you need one additional piece of information. What is it?

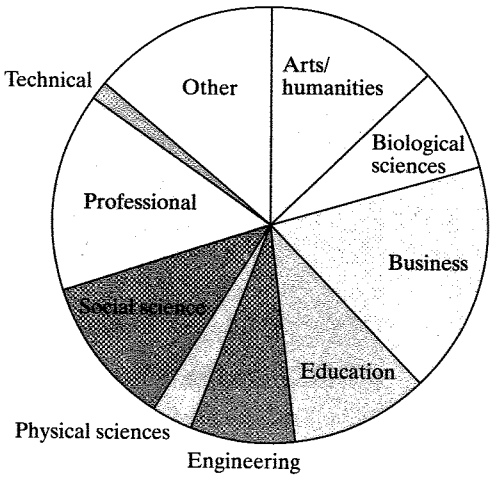
13. **Hispanic origins** Below is a pie chart prepared by the Census Bureau to show the origin of the more than 50 million Hispanics in the United States in 2010.[12] About what percent of Hispanics are Mexican? Puerto Rican?

Percent Distribution of Hispanics by Type: 2010



Comment: You see that it is hard to determine numbers from a pie chart. Bar graphs are much easier to use. (The Census Bureau did include the percents in its pie chart.)

14. **Which major?** About 1.6 million first-year students enroll in colleges and universities each year. What do they plan to study? The pie chart displays data on the percents of first-year students who plan to major in several discipline areas.[13] About what percent of first-year students plan to major in business? In social science?



15. **Buying music online** Young people are more likely than older folk to buy music online. Here are the percents of people in several age groups who bought music online in a recent year:[14]

Age group	Bought music online
12 to 17 years	24%
18 to 24 years	21%
25 to 34 years	20%
35 to 44 years	16%
45 to 54 years	10%
55 to 64 years	3%
65 years and over	1%

(a) Explain why it is *not* correct to use a pie chart to display these data.

(b) Make a bar graph of the data. Be sure to label your axes.

16. **The audience for movies** Here are data on the percent of people in several age groups who attended a movie in the past 12 months:[15]

Age group	Movie attendance
18 to 24 years	83%
25 to 34 years	73%
35 to 44 years	68%
45 to 54 years	60%
55 to 64 years	47%
65 to 74 years	32%
75 years and over	20%

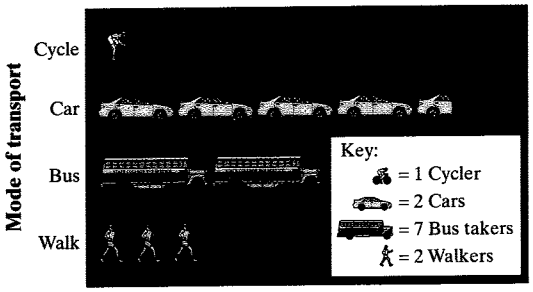
(a) Display these data in a bar graph. Describe what you see.

(b) Would it be correct to make a pie chart of these data? Why or why not?

(c) A movie studio wants to know what percent of the total audience for movies is 18 to 24 years old. Explain why these data do not answer this question.

17. **Going to school** Students in a high school statistics class were given data about the main method of transportation to school for a group of 30 students. They produced the pictograph shown.

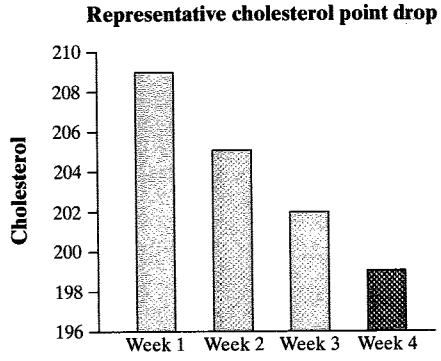
pg 10



(a) How is this graph misleading?

(b) Make a new graph that isn't misleading.

18. **Oatmeal and cholesterol** Does eating oatmeal reduce cholesterol? An advertisement included the following graph as evidence that the answer is "Yes."



(a) How is this graph misleading?

(b) Make a new graph that isn't misleading. What do you conclude about the relationship between eating oatmeal and cholesterol reduction?

19. **Attitudes toward recycled products** Recycling is supposed to save resources. Some people think recycled products are lower in quality than other products, a fact that makes recycling less practical. People who use a recycled product may have different opinions from those who don't use it. Here are data on attitudes toward coffee filters made of recycled paper from a sample of people who do and don't buy these filters:[16]

pg 13

	Buy recycled filters?	
Think quality is	Yes	No
Higher	20	29
The same	7	25
Lower	9	43

(a) How many people does this table describe? How many of these were buyers of coffee filters made of recycled paper?

(b) Give the marginal distribution (in percents) of opinion about the quality of recycled filters. What percent of the people in the sample think the quality of the recycled product is the same or higher than the quality of other filters?

20. **Smoking by students and parents** Here are data from a survey conducted at eight high schools on smoking among students and their parents:[17]

	Neither parent smokes	One parent smokes	Both parents smoke
Student does not smoke	1168	1823	1380
Student smokes	188	416	400

(a) How many students are described in the two-way table? What percent of these students smoke?

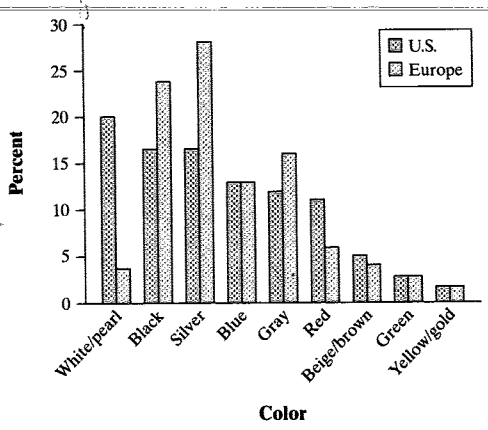
(b) Give the marginal distribution (in percents) of parents' smoking behavior, both in counts and in percents.

21. **Attitudes toward recycled products** Exercise 19 gives data on the opinions of people who have and have not bought coffee filters made from recycled paper. To see the relationship between opinion and experience with the product, find the conditional distributions of opinion (the response variable) for buyers and nonbuyers. What do you conclude?

pg 15

22. **Smoking by students and parents** Refer to Exercise 20. Calculate three conditional distributions of students' smoking behavior: one for each of the three parental smoking categories. Describe the relationship between the smoking behaviors of students and their parents in a few sentences.

23. **Popular colors—here and there** Favorite vehicle colors may differ among countries. The side-by-side bar graph shows data on the most popular colors of cars in a recent year for the United States and Europe. Write a few sentences comparing the two distributions.



24. **Comparing car colors** Favorite vehicle colors may differ among types of vehicle. Here are data on the most popular colors in a recent year for luxury cars and for SUVs, trucks, and vans.

Color	Luxury cars (%)	SUVs, trucks, vans (%)
Black	22	13
Silver	16	16
White pearl	14	1
Gray	12	13
White	11	25
Blue	7	10
Red	7	11
Yellow/gold	6	1
Green	3	4
Beige/brown	2	6

(a) Make a graph to compare colors by vehicle type.
(b) Write a few sentences describing what you see.

25. **Snowmobiles in the park** Yellowstone National Park surveyed a random sample of 1526 winter visitors to the park. They asked each person whether they owned, rented, or had never used a snowmobile. Respondents were also asked whether they belonged to an environmental organization (like the Sierra Club). The two-way table summarizes the survey responses.

	Environmental Club		
	No	Yes	Total
Never used	445	212	**657**
Snowmobile renter	497	77	**574**
Snowmobile owner	279	16	**295**
Total	**1221**	**305**	**1526**

Do these data suggest that there is an association between environmental club membership and snowmobile use among visitors to Yellowstone National Park? Give appropriate evidence to support your answer.

26. **Angry people and heart disease** People who get angry easily tend to have more heart disease. That's the conclusion of a study that followed a random sample of 12,986 people from three locations for about four years. All subjects were free of heart disease at the beginning of the study. The subjects took the Spielberger Trait Anger Scale test, which measures how prone a person is to sudden anger. Here are data for the 8474 people in the sample who had normal blood pressure. CHD stands for "coronary heart disease." This includes people who had heart attacks and those who needed medical treatment for heart disease.

	Low anger	Moderate anger	High anger	Total
CHD	53	110	27	**190**
No CHD	3057	4621	606	**8284**
Total	**3110**	**4731**	**633**	**8474**

Do these data support the study's conclusion about the relationship between anger and heart disease? Give appropriate evidence to support your answer.

Multiple choice: Select the best answer for Exercises 27 to 34.

Exercises 27 to 30 refer to the following setting. The National Survey of Adolescent Health interviewed several thousand teens (grades 7 to 12). One question asked was "What do you think are the chances you will be married in the next ten years?" Here is a two-way table of the responses by gender:[18]

	Female	Male
Almost no chance	119	103
Some chance, but probably not	150	171
A 50-50 chance	447	512
A good chance	735	710
Almost certain	1174	756

27. The percent of females among the respondents was
(a) 2625. (c) about 46%. (e) None of these.
(b) 4877. (d) about 54%.

28. Your percent from the previous exercise is part of
(a) the marginal distribution of females.
(b) the marginal distribution of gender.
(c) the marginal distribution of opinion about marriage.
(d) the conditional distribution of gender among adolescents with a given opinion.
(e) the conditional distribution of opinion among adolescents of a given gender.

29. What percent of females thought that they were almost certain to be married in the next ten years?
(a) About 16% (c) About 40% (e) About 61%
(b) About 24% (d) About 45%

30. Your percent from the previous exercise is part of
(a) the marginal distribution of gender.
(b) the marginal distribution of opinion about marriage.
(c) the conditional distribution of gender among adolescents with a given opinion.
(d) the conditional distribution of opinion among adolescents of a given gender.
(e) the conditional distribution of "Almost certain" among females.

31. For which of the following would it be inappropriate to display the data with a single pie chart?

(a) The distribution of car colors for vehicles purchased in the last month.

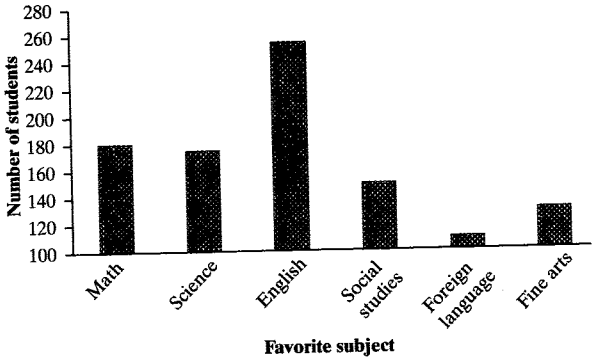
(b) The distribution of unemployment percentages for each of the 50 states.

(c) The distribution of favorite sport for a sample of 30 middle school students.

(d) The distribution of shoe type worn by shoppers at a local mall.

(e) The distribution of presidential candidate preference for voters in a state.

32. The following bar graph shows the distribution of favorite subject for a sample of 1000 students. What is the most serious problem with the graph?



(a) The subjects are not listed in the correct order.

(b) This distribution should be displayed with a pie chart.

(c) The vertical axis should show the percent of students.

(d) The vertical axis should start at 0 rather than 100.

(e) The foreign language bar should be broken up by language.

33. In the 2010–2011 season, the Dallas Mavericks won the NBA championship. The two-way table below displays the relationship between the outcome of each game in the regular season and whether the Mavericks scored at least 100 points.

	100 or more points	Fewer than 100 points	Total
Win	43	14	**57**
Loss	4	21	**25**
Total	**47**	**35**	**82**

Which of the following is the best evidence that there is an association between the outcome of a game and whether or not the Mavericks scored at least 100 points?

(a) The Mavericks won 57 games and lost only 25 games.

(b) The Mavericks scored at least 100 points in 47 games and fewer than 100 points in only 35 games.

(c) The Mavericks won 43 games when scoring at least 100 points and only 14 games when scoring fewer than 100 points.

(d) The Mavericks won a higher proportion of games when scoring at least 100 points (43/47) than when they scored fewer than 100 points (14/35).

(e) The combination of scoring 100 or more points and winning the game occurred more often (43 times) than any other combination of outcomes.

34. The following partially complete two-way table shows the marginal distributions of gender and handedness for a sample of 100 high school students.

	Male	Female	Total
Right	x		**90**
Left			**10**
Total	**40**	**60**	**100**

If there is no association between gender and handedness for the members of the sample, which of the following is the correct value of x?

(a) 20.

(b) 30.

(c) 36.

(d) 45.

(e) Impossible to determine without more information.

35. **Marginal distributions aren't the whole story** Here are the row and column totals for a two-way table with two rows and two columns:

a	b	50
c	d	50
60	40	100

Find *two different* sets of counts a, b, c, and d for the body of the table that give these same totals. This shows that the relationship between two variables can not be obtained from the two individual distributions of the variables.

36. **Fuel economy** (Introduction) Here is a small part of a data set that describes the fuel economy (in miles per gallon) of model year 2012 motor vehicles:

Make and model	Vehicle type	Transmission type	Number of cylinders	City mpg	Highway mpg
Aston Martin Vantage	Two-seater	Manual	8	14	20
Honda Civic Hybrid	Subcompact	Automatic	4	44	44
Toyota Prius	Midsize	Automatic	4	51	48
Chevrolet Impala	Large	Automatic	6	18	30

(a) What are the individuals in this data set?

(b) What variables were measured? Identify each as categorical or quantitative.