Displaying and Describing Categorical Data

CHAPTER 4

Keen, Inc.

KEEN, Inc. was started to create a sandal designed for a variety of water activities. The sandals quickly became popular due to their unique patented toe protection—a black bumper to protect the toes when adventuring out on rivers and trails. Today the KEEN brand offers over 300 different outdoor performance and outdoor inspired casual footwear styles as well as bags and socks.

Few companies experience the kind of growth that KEEN did in less than seven years. Amazingly, they've done this with relatively little advertising and by selling primarily to specialty footwear and outdoor stores, in addition to online outlets.

After the 2004 Tsunami disaster, KEEN cut its advertising budget almost completely and donated over $1 million to help the victims and establish the KEEN Foundation to support environmental and social causes. Philanthropy and community projects continue to play an integral part of the KEEN brand values. In fact, KEEN has established a giving program with a philanthropic effort devoted to helping the environment, conservation, and social movements involving the outdoors.

WHO	Visits to the KEEN, Inc. website
WHAT	Search Engine that led to KEEN's website
WHEN	September 2006
WHERE	Worldwide
HOW	Data compiled via *Google® Analytics* from KEEN website
WHY	To understand customer use of the website and how they got there

KEEN, Inc., like most companies, collects data on visits to its website. Each visit to the site and each subsequent action the visitor takes (changing the page, entering data, etc.) is recorded in a file called a usage, or access weblog. These logs contain a lot of potentially worthwhile information, but they are not easy to use. Here's one line from a log:

```
245.240.221.71 -- [03/Jan/2007:15:20:06-0800]" GET
http://www.keenfootwear.com/pdp_page.cfm?productID=148"
200 8788 "http://www.google.com/" "Mozilla/3.0WebTV/1.2
(compatible; MSIE 2.0)"
```

Unless the company has the analytic resources to deal with these files, it must rely on a third party to summarize the data. KEEN, like many other small and midsized companies, uses *Google Analytics* to collect and summarize its log data.

Imagine a whole table of data like the one above—with a line corresponding to every visit. In September 2006 there were 93,173 visits to the KEEN site, which would be a table with 93,173 rows. The problem with a file like this—and in fact even with data tables—is that we can't see what's going on. And seeing is exactly what we want to do. We need ways to show the data so that we can see patterns, relationships, trends, and exceptions.

4.1 Summarizing a Categorical Variable

The Three Rules of Data Analysis

There are three things you should always do with data:

1. **Make a picture.** A display of your data will reveal things you are not likely to see in a table of numbers and will help you to *plan* your approach to the analysis and think clearly about the patterns and relationships that may be hiding in your data.
2. **Make a picture.** A well-designed display will *do* much of the work of analyzing your data. It can show the important features and patterns. A picture will also reveal things you did not expect to see: extraordinary (possibly wrong) data values or unexpected patterns.
3. **Make a picture.** The best way to *report* to others what you find in your data is with a well-chosen picture.

These are the three rules of data analysis. These days, technology makes drawing pictures of data easy, so there is no reason not to follow the three rules. Here are some displays showing various aspects of traffic on one of the authors' websites.

Some displays communicate information better than others. We'll discuss some general principles for displaying information honestly in this chapter.

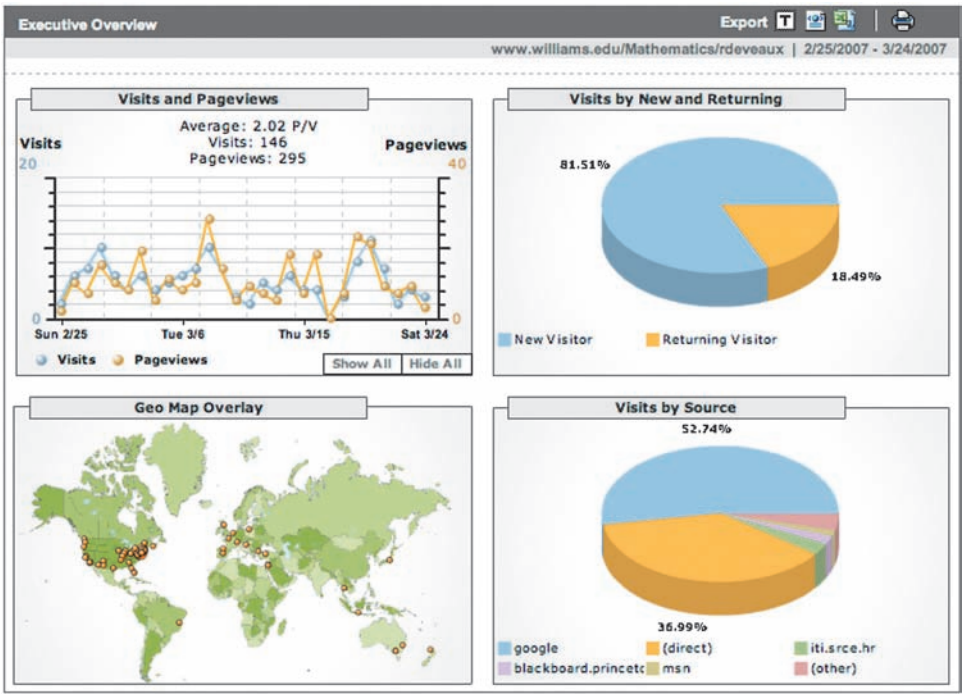


Figure 4.1 Part of the output from *Google Analytics* (www.google.com/analytics) for the period Feb. 25 to March 24, 2007 displaying website traffic.

Search Engine	Visits
Google	50,629
Direct	22,173
Yahoo	7272
MSN	3166
SnapLink	946
All Others	8987
Total	**93,167**

Table 4.1 A frequency table of the Search Engine used by visitors to the KEEN, Inc. website.

Search Engine	Visits by %
Google	54.34%
Direct	23.80%
Yahoo	7.80%
MSN	3.40%
SnapLink	1.02%
All Others	9.65%
Total	**100.00%**

Table 4.2 A relative frequency table for the same data.

Frequency Tables

KEEN might be interested to know how people find their website. They might use the information to allocate their advertising revenue to various search engines, putting ads where they'll be seen by the most potential customers. The variable *Search Engine* records, for each visit to KEEN's website, where the visit came from. The categories are all the search engines used, plus the label "Direct," which indicates that the customer typed in KEEN's web address (or URL) directly into the browser. In order to make sense of the 93,167 visits for which they have data, they'd like to summarize the variable and display the information in a way that can easily communicate the results to others.

In order to make a picture of any variable, we first need to organize its values. For a categorical variable, like *Search Engine*, this is easy—we just count the number of cases corresponding to each category. A **frequency table** (Table 4.1) records the counts for each of the categories of the variable and lists the counts under the category name. By ordering the categories by number of counts, we can easily see, for example, that the most popular source was Google.

The names of the categories label each row in the frequency table. For *Search Engine* these are "Google," "Direct," "Yahoo," and so on. Even with thousands of cases, a variable that doesn't have too many categories produces a frequency table that is easy to read. A frequency table with dozens or hundreds of categories would be much harder to read. Notice the label of the last line of the table—"All Others." When the number of categories gets too large, we often lump together values of the variable into "Other." When to do that is a judgment call, but it's a good idea to have fewer than about a dozen categories.

Counts are useful, but sometimes we want to know the fraction or **proportion** of the data in each category, so we divide the counts by the total number of cases. Usually we multiply by 100 to express these proportions as **percentages**. A **relative frequency table** (Table 4.2) displays the *percentages*, rather than the counts, of the

values in each category. Both types of table show how the cases are distributed across the categories. In this way, they describe the **distribution** of a categorical variable because they name the possible categories and tell how frequently each occurs.

For Example — Making frequency and relative frequency tables

The Super Bowl, the championship game of the National Football League of the United States, is an important annual social event for Americans, with tens of millions of viewers. The ads that air during the game are expensive: a 30-second ad during the 2010 Super Bowl cost about $3M. The high price of these commercials makes them high-profile and much anticipated, and so the advertisers feel pressure to be innovative, entertaining, and often humorous. Some people, in fact, watch the Super Bowl mainly for the commercials. Before the 2007 Super Bowl, the Gallup Poll asked 1008 U.S. adults whether they were more interested in watching the game or the commercials. Here are 40 of those responses (NA/Don't Know = No Answer or Don't Know):

Won't Watch	Game	Commercials	Won't Watch	Game
Game	Won't Watch	Commercials	Game	Game
Commercials	Commercials	Game	Won't Watch	Commercials
Game	NA/Don't Know	Commercials	Game	Game
Won't Watch	Game	Game	Won't Watch	Game
Game	Won't Watch	Won't Watch	Game	Won't Watch
Won't Watch	Commercials	Commercials	Game	Won't Watch
NA/Don't Know	Won't Watch	Game	Game	Game

Question: Make a frequency table for this variable. Include the percentages to display both a frequency and relative frequency table at the same time.

Answer: There were four different responses to the question about watching the Super Bowl. Counting the number of participants who responded to each of these gives the following table:

Response	Counts	Percentage
Commercials	8	20.0%
Game	18	45.0%
Won't Watch	12	30.0%
No Answer/Don't Know	2	5.0%
Total	**40**	**100.0%**

4.2 Displaying a Categorical Variable

The Area Principle

100.01%?
If you are careful to add the percentages in Table 4.2, you will notice the total is 100.01%. Of course the real total has to be 100.00%. The discrepancy is due to individual percentages being rounded. You'll often see this in tables of percents, sometimes with explanatory footnotes.

Now that we have a frequency table, we're ready to follow the three rules of data analysis. But we can't make just any display; a bad picture can distort our understanding rather than help it. For example, Figure 4.2 is a graph of the frequencies of Table 4.1. What impression do you get of the relative frequencies of visits from each source?

While it's true that the majority of people came to KEEN's website from Google, in Figure 4.2 it looks like nearly all did. That doesn't seem right. What's wrong? The lengths of the sandals *do* match the totals in the table. But our eyes tend to be more impressed by the *area* (or perhaps even the *volume*) than by other aspects of each sandal image, and it's that aspect of the image that we notice. Since there were about twice as many people who came from Google as those who typed

Figure 4.2 Although the length of each sandal corresponds to the correct number, the impression we get is all wrong because we perceive the entire area of the sandal. In fact, only a little more than 50% of all visitors used Google to get to the website.

the URL in directly, the sandal depicting the number of Google visitors is about two times longer than the sandal below it, but it occupies about four times the area. As you can see from the frequency table, that just isn't a correct impression.

The best data displays observe a fundamental principle of graphing data called the **area principle**, which says that the area occupied by a part of the graph should correspond to the magnitude of the value it represents.

Bar Charts

Figure 4.3 gives us a chart that obeys the area principle. It's not as visually entertaining as the sandals, but it does give a more *accurate* visual impression of the distribution. The height of each bar shows the count for its category. The bars are the

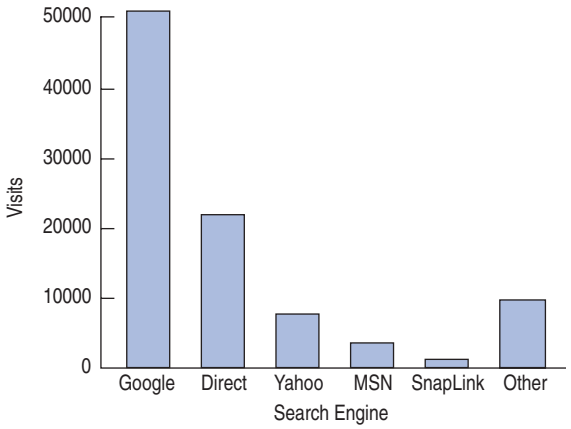
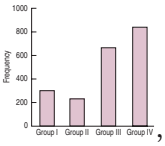
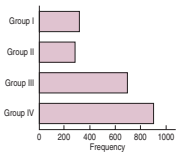


Figure 4.3 Visits to the KEEN, Inc. website by *Search Engine* choice. With the area principle satisfied, the true distribution is clear.

same width, so their heights determine their areas, and the areas are proportional to the counts in each class. Now it's easy to see that nearly half the site hits came from places other than Google—not the impression that the sandals in Figure 4.2 conveyed. We can also see that there were a little more than twice as many visits that originated with a Google search as there were visits that came directly. Bar charts make these kinds of comparisons easy and natural.

A **bar chart** displays the distribution of a categorical variable, showing the counts for each category next to each other for easy comparison. Bar charts should have small spaces between the bars to indicate that these are freestanding bars that could be rearranged into any order. The bars are lined up along a common base with labels for each category. The variable name is often used as a subtitle for the x-axis.

Bar charts are usually drawn vertically in columns, 

, but sometimes they are drawn with horizontal bars, like this.[1] 

If we want to draw attention to the relative *proportion* of visits from each *Search Engine*, we could replace the counts with percentages and use a **relative frequency bar chart**, like the one shown in Figure 4.4.

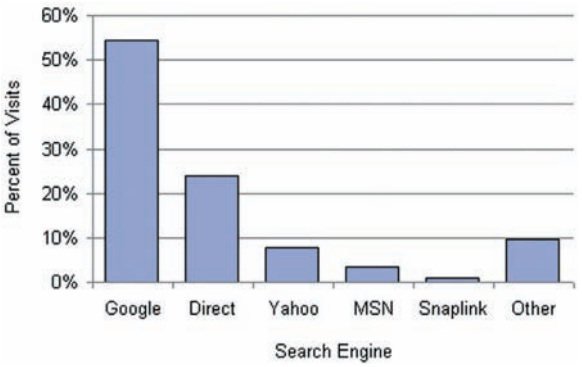


Figure 4.4 The relative frequency bar chart, created in **Excel**, looks the same as the bar chart (Figure 4.3) but shows the proportion of visits in each category rather than the counts.

Pie Charts

Another common display that shows how a whole group breaks into several categories is a pie chart. **Pie charts** show the whole group of cases as a circle. They slice the circle into pieces whose size is proportional to the fraction of the whole in each category.

[1]Excel refers to this display as a column chart when the bars are vertical and a bar chart when they are horizontal.

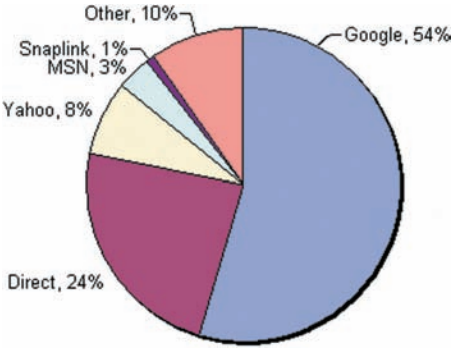


Figure 4.5 A pie chart shows the proportion of visits by Search Engine. Pie chart created in **Excel**.

Pie charts give a quick impression of how a whole group is partitioned into smaller groups. Because we're used to cutting up pies into 2, 4, or 8 pieces, pie charts are good for seeing relative frequencies near 1/2, 1/4, or 1/8. For example, in Figure 4.5, you can easily see that the slice representing Google is just slightly more than half the total. Unfortunately, other comparisons are harder to make with pie charts. Were there more visits from Yahoo, or from All Others? It's hard to tell since the two slices look about the same. Comparisons such as these are usually easier in a bar chart. (Compare to Figure 4.4.)

- **Think before you draw.** Our first rule of data analysis is *Make a picture*. But what kind of picture? We don't have a lot of options—yet. There's more to Statistics than pie charts and bar charts, and knowing when to use every type of display we'll discuss is a critical first step in data analysis. That decision depends in part on what type of data you have and on what you hope to communicate.

  We always have to check that the data are appropriate for whatever method of analysis we choose. Before you make a bar chart or a pie chart, always check the **Categorical Data Condition:** that the data are counts or percentages of individuals in categories.

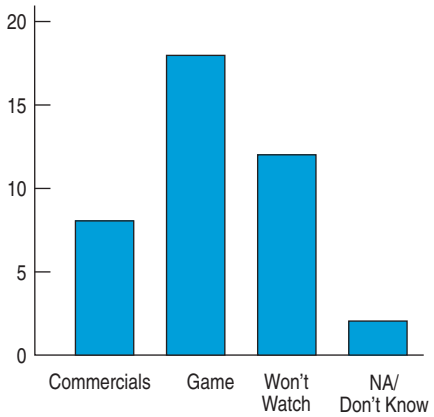
  If you want to make a pie chart or relative frequency bar chart, you'll need to also make sure that the categories don't overlap, so that no individual is counted in two categories. If the categories do overlap, it's misleading to make a pie chart, since the percentages won't add up to 100%. For the *Search Engine* data, either kind of display is appropriate because the categories don't overlap—each visit comes from a unique source.

  Throughout this course, you'll see that doing Statistics right means selecting the proper methods. That means you have to think about the situation at hand. An important first step is to check that the type of analysis you plan is appropriate. The Categorical Data Condition is just the first of many such checks.

For Example — Making a bar chart

Question: Make a bar chart for the 40 Super Bowl responses of the example on page 54.

Answer: Use the frequencies in the table in the example on page 54 to produce the heights of the bars:



4.3 Exploring Two Categorical Variables: Contingency Tables

WHO	Respondents in the GfK Roper Reports Worldwide Survey
WHAT	Responses to questions relating to perceptions of food and health
WHEN	Fall 2005; published in 2006
WHERE	Worldwide
HOW	Data collected by GfK Roper Consulting using a multistage design
WHY	To understand cultural differences in the perception of the food and beauty products we buy and how they affect our health

In Chapter 3 we saw how GfK Roper Consulting gathered information on consumers attitudes about health, food, and health care products. In order to effectively market food products across different cultures, it's essential to know how strongly people in different cultures feel about their food. One question in the Roper survey asked respondents whether they agreed with the following statement: "I have a strong preference for regional or traditional products and dishes from where I come from." Here is a frequency table (Table 4.3) of the responses.

Response to *Regional Food Preference Question*	Counts	Relative Frequency
Agree Completely	2346	30.51%
Agree Somewhat	2217	28.83%
Neither Disagree Nor Agree	1738	22.60%
Disagree Somewhat	811	10.55%
Disagree Completely	498	6.48%
Don't Know	80	1.04%
Total	**7690**	**100.00%**

Table 4.3 A combined frequency and relative frequency table for the responses (from all 5 countries represented: China, France, India, the U.K., and the U.S.) to the statement "I have a strong preference for regional or traditional products and dishes from where I come from."

The pie chart (Figure 4.6) shows clearly that more than half of all the respondents agreed with the statement.

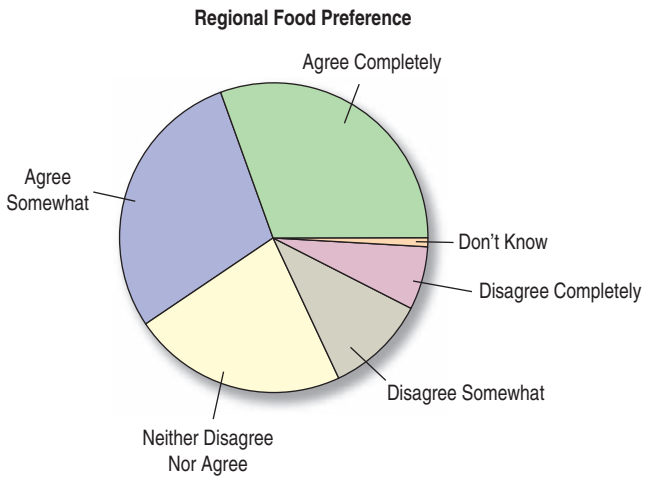


Figure 4.6 It's clear from the pie chart that the majority of respondents identify with their local foods.

But if we want to target our marketing differently in different countries, wouldn't it be more interesting to know how opinions vary from country to country?

To find out, we need to look at the two categorical variables *Regional Preference* and *Country* together, which we do by arranging the data in a two-way table. Table 4.4 is a two-way table of *Regional Preference* by *Country*. Because the table shows how the individuals are distributed along each variable, depending on, or *contingent on*, the value of the other variable, such a table is called a **contingency table**.

		Regional Preference						
		Agree Completely	Agree Somewhat	Neither Disagree Nor Agree	Disagree Somewhat	Disagree Completely	Don't Know	Total
Country	China	518	576	251	117	33	7	**1502**
	France	347	475	400	208	94	15	**1539**
	India	960	282	129	65	95	4	**1535**
	U.K.	214	407	504	229	175	28	**1557**
	U.S.	307	477	454	192	101	26	**1557**
	Total	**2346**	**2217**	**1738**	**811**	**498**	**80**	**7690**

Table 4.4 Contingency table of *Regional Preference* and *Country*. The bottom line "Totals" are the values that were in Table 4.3.

The margins of a contingency table give totals. In the case of Table 4.4, these are shown in both the right-hand column (in bold) and the bottom row (also in bold). The totals in the bottom row of the table show the frequency distribution of the variable *Regional Preference*. The totals in the right-hand column of the table show the frequency distribution of the variable *Country*. When presented like this, at the margins of a contingency table, the frequency distribution of either one of the variables is called its **marginal distribution**. The marginal distribution for a variable in a contingency table is the same frequency distribution we found by considering each variable separately.

Each **cell** of a contingency table (any intersection of a row and column of the table) gives the count for a combination of values of the two variables. For example, in Table 4.4 you can see that 504 people in the United Kingdom neither agreed nor disagreed. Looking down the Agree Completely column, you can see that the largest number of responses in that column (960) are from India. Are Britons less likely to agree with the statement than Indians or Chinese? Questions like this are more naturally addressed using percentages.

We know that 960 people from India agreed completely with the statement. We could display this number as a percentage, but as a percentage of what? The total number of people in the survey? (960 is 12.5% of the total.) The number of Indians in the survey? (960 is 62.5% of the row total.) The number of people who agree completely? (960 is 40.9% of the column total.) All of these are possibilities, and all are potentially useful or interesting. You'll probably wind up calculating (or letting your technology calculate) lots of percentages. Most statistics programs offer a choice of **total percent**, **row percent**, or **column percent** for contingency tables. Unfortunately, they often put them all together with several numbers in each cell of the table. The resulting table (Table 4.5) holds lots of information but is hard to understand.

Country		Regional Preference						
		Agree Completely	Agree Somewhat	Neither Disagree Nor Agree	Disagree Somewhat	Disagree Completely	Don't Know	Total
	China	518	576	251	117	33	7	**1502**
	% of Row	*34.49*	*38.35*	*16.71*	*7.79*	*2.20*	*0.47*	**100.00%**
	% of Column	*22.08*	*25.98*	*14.44*	*14.43*	*6.63*	*8.75*	**19.53%**
	% of Total	*6.74*	*7.49*	*3.26*	*1.52*	*0.43*	*0.09*	**19.53%**
	France	347	475	400	208	94	15	**1539**
	% of Row	*22.55*	*30.86*	*25.99*	*13.52*	*6.11*	*0.97*	**100.00%**
	% of Column	*14.79*	*21.43*	*23.01*	*25.65*	*18.88*	*18.75*	**20.01%**
	% of Total	*4.51*	*6.18*	*5.20*	*2.70*	*1.22*	*0.20*	**20.01%**
	India	960	282	129	65	95	4	**1535**
	% of Row	*62.54*	*18.37*	*8.40*	*4.23*	*6.19*	*0.26*	**100.00%**
	% of Column	*40.92*	*12.72*	*7.42*	*8.01*	*19.08*	*5.00*	**19.96%**
	% of Total	*12.48*	*3.67*	*1.68*	*0.85*	*1.24*	*0.05*	**19.96%**
	U.K.	214	407	504	229	175	28	**1557**
	% of Row	*13.74*	*26.14*	*32.37*	*14.71*	*11.24*	*1.80*	**100.00%**
	% of Column	*9.12*	*18.36*	*29.00*	*28.24*	*35.14*	*35.00*	**20.24%**
	% of Total	*2.78*	*5.29*	*6.55*	*2.98*	*2.28*	*0.36*	**20.24%**
	U.S.	307	477	454	192	101	26	**1557**
	% of Row	*19.72*	*30.64*	*29.16*	*12.33*	*6.49*	*1.67*	**100.00%**
	% of Column	*13.09*	*21.52*	*26.12*	*23.67*	*20.28*	*32.50*	**20.24%**
	% of Total	*3.99*	*6.20*	*5.90*	*2.50*	*1.31*	*0.34*	**20.24%**
	Total	**2346**	**2217**	**1738**	**811**	**498**	**80**	**7690**
	% of Row	**30.51%**	**28.83%**	**22.60%**	**10.55%**	**6.48%**	**1.04%**	**100.00%**
	% of Column	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	% of Total	**30.51%**	**28.83%**	**22.60%**	**10.55%**	**6.48%**	**1.04%**	**100.00%**

Table 4.5 Another contingency table of *Regional Preference* and *Country*. This time we see not only the counts for each combination of the two variables, but also the percentages these counts represent. For each count, there are three choices for the percentage: by row, by column, and by table total. There's probably too much information here for this table to be useful.

To simplify the table, let's pull out the values corresponding to the percentages of the total.

Country	Regional Preference						
	Agree Completely	Agree Somewhat	Neither Disagree Nor Agree	Disagree Somewhat	Disagree Completely	Don't Know	Total
China	*6.74*	*7.49*	*3.26*	*1.52*	*0.43*	*0.09*	**19.53**
France	*4.51*	*6.18*	*5.20*	*2.70*	*1.22*	*0.20*	**20.01**
India	*12.48*	*3.67*	*1.68*	*0.85*	*1.24*	*0.05*	**19.96**
U.K.	*2.78*	*5.29*	*6.55*	*2.98*	*2.28*	*0.36*	**20.25**
U.S.	*3.99*	*6.20*	*5.90*	*2.50*	*1.31*	*0.34*	**20.25**
Total	**30.51**	**28.83**	**22.60**	**10.55**	**6.48**	**1.04**	**100.00**

Table 4.6 A contingency table of *Regional Preference* and *Country* showing only the total percentages.

These percentages tell us what percent of *all* respondents belong to each combination of column and row category. For example, we see that 3.99% of the respondents were Americans who agreed completely with the question, which is slightly more than the percentage of Indians who agreed somewhat. Is this fact useful? Is that really what we want to know?

Always be sure to ask "percent of what?" That will help define the *who* and will help you decide whether you want *row*, *column*, or *table* percentages.

Percent of what?

The English language can be tricky when we talk about percentages. If asked, "What percent of those answering 'I Don't Know' were from India?" it's pretty clear that you should focus only on the *Don't Know* column. The question itself seems to restrict the *who* in the question to that column, so you should look at the number of those in each country among the 80 people who replied "I don't know." You'd find that in the column percentages, and the answer would be 4 out of 80 or 5.00%.

But if you're asked, "What percent were Indians who replied 'I don't know?' you'd have a different question. Be careful. The question really means "what percent of the entire sample were both from India and replied 'I don't know'?" So the *who* is all respondents. The denominator should be 7690, and the answer is the table percent 4/7690 = 0.05%.

Finally, if you're asked, "What percent of the Indians replied 'I don't know'?" you'd have a third question. Now the *who* is Indians. So the denominator is the 1535 Indians, and the answer is the row percent, 4/1535 = 0.26%.

Conditional Distributions

The more interesting questions are contingent on something. We'd like to know, for example, what percentage *of Indians* agreed completely with the statement and how that compares to the percentage *of Britons* who also agreed. Equivalently, we might ask whether the chance of agreeing with the statement depended on the *Country* of the respondent. We can look at this question in two ways. First, we could ask how the distribution of *Regional Preference* changes across *Country*. To do that we look at the row percentages.

		Regional Preference						
		Agree Completely	Agree Somewhat	Neither Disagree Nor Agree	Disagree Somewhat	Disagree Completely	Don't Know	Total
Country	India	960	282	129	65	95	4	**1535**
		62.54	18.37	8.40	4.23	6.19	0.26	**100%**
	U.K.	214	407	504	229	175	28	**1557**
		13.74	26.14	32.37	14.71	11.24	1.80	**100%**

Table 4.7 The conditional distribution of *Regional Preference* conditioned on two values of *Country:* India and the United Kingdom. This table shows the row percentages.

By focusing on each row separately, we see the distribution of *Regional Preference* under the condition of being in the selected *Country*. The sum of the percentages in each row is 100%, and we divide that up by the responses to the question. In effect, we can temporarily restrict the *who* first to Indians and look at how their responses are distributed. A distribution like this is called a **conditional distribution** because it shows the distribution of one variable for just those cases that satisfy a condition on another. We can compare the two conditional distributions with pie

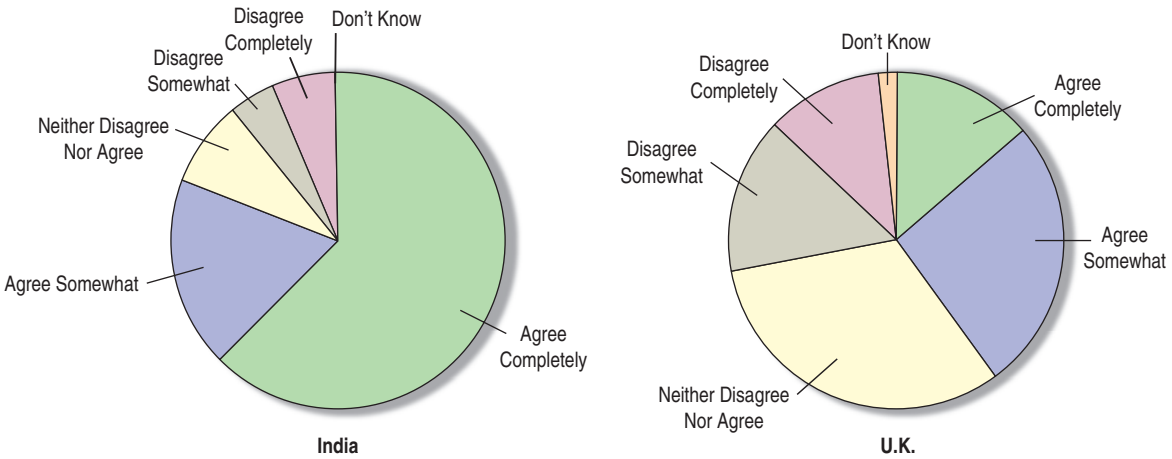


Figure 4.7 Pie charts of the conditional distributions of *Regional Food Preference* importance for India and the United Kingdom. The percentage of people who agree is much higher in India than in the United Kingdom.

charts (Figure 4.7). Of course, we could also turn the question around. We could look at the distribution of *Country* for each category of *Regional Preference*. To do this, we would look at the column percentages.

Looking at how the percentages change across each row, it sure looks like the distribution of responses to the question is different in each *Country*. To make the differences more vivid, we could also display the conditional distributions. Figure 4.8 shows an example of a side-by-side bar chart, displaying the responses to the questions for India and the United Kingdom.

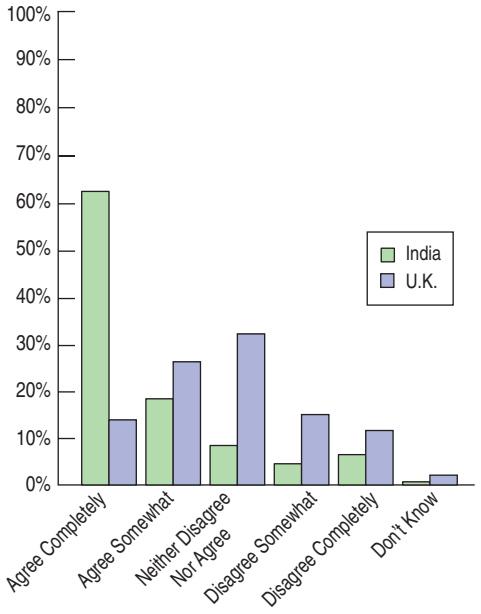


Figure 4.8 Side-by-side bar charts showing the conditional distribution of *Regional Food Preference* for both India and the United Kingdom. It's easier to compare percentages within each country with side-by-side bar charts than pie charts.

From Figure 4.8, it is clear that Indians have a stronger preference for their own cuisine than Britons have for theirs. For food companies, including GfK Roper's clients, that means Indians are less likely to accept a food product they perceive as foreign, and people in Great Britain are more accepting of "foreign" foods. This could be invaluable information for marketing products.

Variables can be associated in many ways and to different degrees. The best way to tell whether two variables are associated is to ask whether they are *not*.[2] In a contingency table, when the distribution of one variable is the same for all categories of another variable, we say that the two variables are **independent**. That tells us there's no association between these variables. We'll see a way to check for independence formally later in the book. For now, we'll just compare the distributions.

For Example — Contingency tables and side-by-side bar charts

Here is a contingency table of the responses to the question Gallup asked about the Super Bowl by sex:

	Sex		
	Female	Male	Total
Game	198	277	**475**
Commercials	154	79	**233**
NA/Don't Know	4	4	**8**
Won't Watch	160	132	**292**
Total	**516**	**492**	**1008**

Question: Does it seem that there is an association between what viewers are interested in watching and their sex?

Answer: First, find the conditional distributions of the four responses for each sex:

For Men:

Game = 277/492 = 56.3%

Commercials = 79/492 = 16.1%

Won't Watch = 132/492 = 26.8%

NA/Don't Know = 4/492 = 0.8%

For Women:

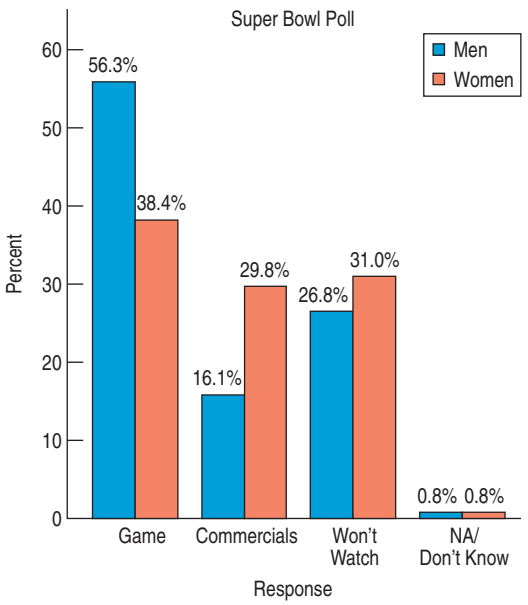
Game = 198/516 = 38.4%

Commercials = 154/516 = 29.8%

Won't Watch = 160/516 = 31.0%

NA/Don't Know = 4/516 = 0.8%

Now display the two distributions with side-by-side bar charts:



(continued)

[2]This kind of "backwards" reasoning shows up surprisingly often in science—and in statistics.

Based on this poll it appears that women were only slightly less interested than men in watching the Super Bowl telecast: 31% of the women said they didn't plan to watch, compared to just under 27% of men. Among those who planned to watch, however, there appears to be an association between the viewer's sex and what the viewer is most looking forward to. While more women are interested in the game (38%) than the commercials (30%), the margin among men is much wider: 56% of men said they were looking forward to seeing the game, compared to only 16% who cited the commercials.

Just Checking

So that they can balance their inventory, an optometry shop collects the following data for customers in the shop.

		Eye Condition			
		Nearsighted	Farsighted	Need Bifocals	Total
Sex	Males	6	20	6	32
	Females	4	16	12	32
	Total	10	36	18	64

1 What percent of females are farsighted?
2 What percent of nearsighted customers are female?
3 What percent of all customers are farsighted females?
4 What's the distribution of *Eye Condition*?
5 What's the conditional distribution of *Eye Condition* for males?
6 Compare the percent who are female among nearsighted customers to the percent of all customers who are female.
7 Does it seem that *Eye Condition* and *Sex* might be dependent? Explain.

Segmented Bar Charts

We could display the Roper survey information by dividing up bars rather than circles as we did when making pie charts. The resulting **segmented bar chart** treats each bar as the "whole" and divides it proportionally into segments corresponding to the percentage in each group. We can see that the distributions of responses to the question are very different in the two countries, indicating again that *Regional Preference* is not independent of *Country*.

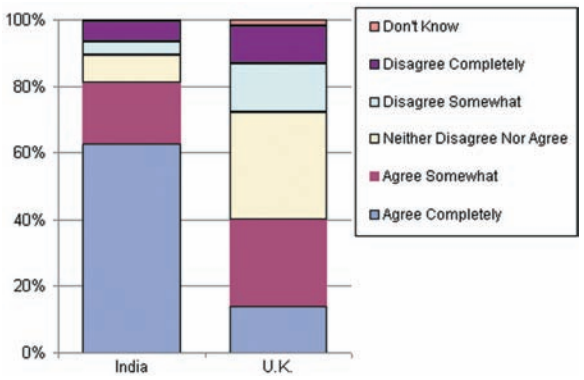


Figure 4.9 Although the totals for India and the United Kingdom are different, the bars are the same height because we have converted the numbers to percentages. Compare this display (created in **Excel**) with the side-by-side pie charts of the same data in Figure 4.7.

Guided Example Food Safety

Food storage and food safety are major issues for multinational food companies. A client wants to know if people of all age groups have the same degree of concern so GfK Roper Consulting asked 1500 people in five countries whether they agree with the following statement: "I worry about how safe the food I buy is." We might want to report to a client who was interested in how concerns about food safety were related to age.

PLAN

Setup

- State the objectives and goals of the study.
- Identify and define the variables.
- Provide the time frame of the data collection process.

Determine the appropriate analysis for data type.

The client wants to examine the distribution of responses to the food safety question and see whether they are related to the age of the respondent. GfK Roper Consulting collected data on this question in the fall of 2005 for their 2006 Worldwide report. We will use the data from that study.

The variable is *Food Safety*. The responses are in nonoverlapping categories of agreement, from Agree Completely to Disagree Completely (and Don't Know). There were originally 12 Age groups, which we can combine into five:

Teen	13–19
Young Adult	20–29
Adult	30–39
Middle Aged	40–49
Mature	50 and older

Both variables, *Food Safety* and *Age*, are ordered categorical variables. To examine any differences in responses across age groups, it is appropriate to create a contingency table and a side-by-side bar chart.
Here is a contingency table of "Food Safety" by "Age."

DO

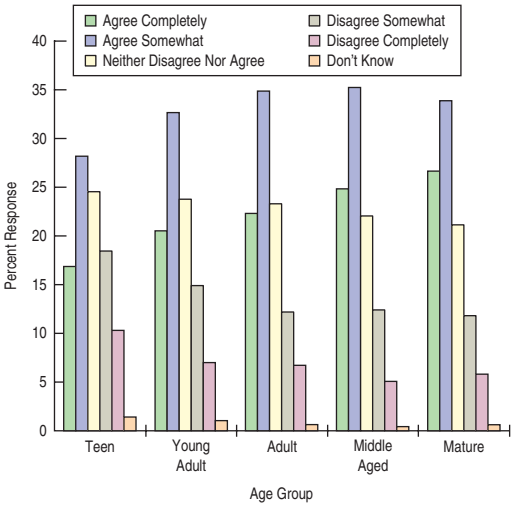
Mechanics For a large data set like this, we rely on technology to make table and displays.

		Food Safety						
		Agree Completely	Agree Somewhat	Neither Disagree Nor Agree	Disagree Somewhat	Disagree Completely	Don't Know	Total
Age	Teen	16.19	27.50	24.32	19.30	10.58	2.12	100%
	Young Adult	20.55	32.68	23.81	14.94	6.98	1.04	100%
	Adult	22.23	34.89	23.28	12.26	6.75	0.59	100%
	Middle Aged	24.79	35.31	22.02	12.43	5.06	0.39	100%
	Mature	26.60	33.85	21.21	11.89	5.82	0.63	100%

(continued)

A side-by-side bar chart is particularly helpful when comparing multiple groups.

A side-by-side bar chart shows the percent of each response to the question by Age group.



REPORT

Summary and Conclusions Summarize the charts and analysis in context. Make recommendations if possible and discuss further analysis that is needed.

MEMO

Re: Food safety concerns by age

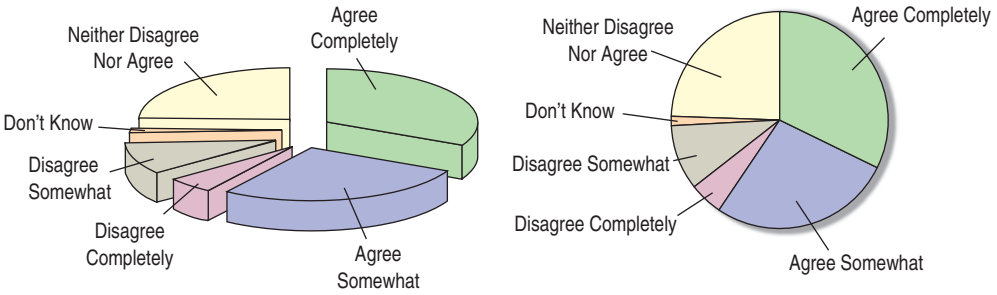
Our analysis of the GfK Roper Reports™ Worldwide survey data for 2006 shows a pattern of concern about food safety that generally increases from youngest to oldest.

Our analysis thus far has not considered whether this trend is consistent across countries. If it were of interest to your group, we could perform a similar analysis for each of the countries.

The enclosed tables and plots provide support for these conclusions.

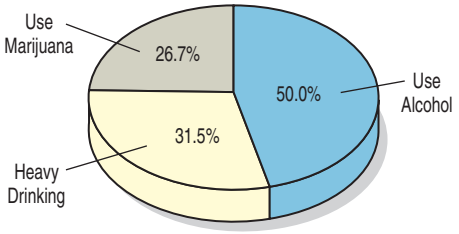
What Can Go Wrong?

- **Don't violate the area principle.** This is probably the most common mistake in a graphical display. Violations of the area principle are often made for the sake of artistic presentation. Here, for example, are two versions of the same pie chart for the *Regional Preference* data.



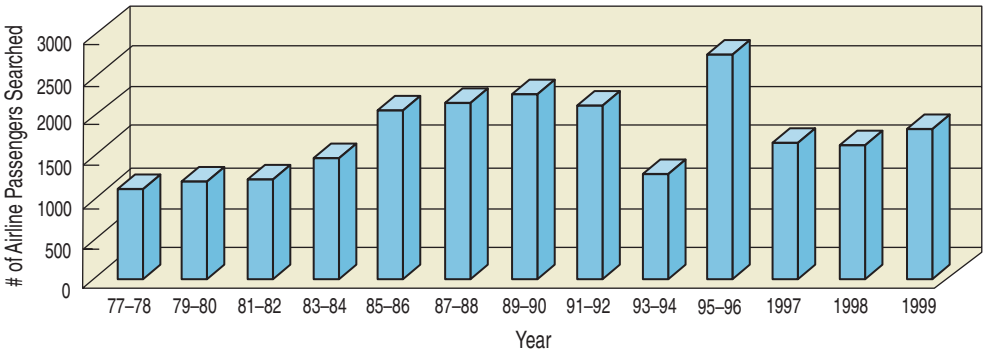
The one on the left looks interesting, doesn't it? But showing the pie three dimensionally on a slant violates the area principle and makes it much more difficult to compare fractions of the whole made up of each category of the response—the principal feature that a pie chart ought to show.

- **Keep it honest.** Here's a pie chart that displays data on the percentage of high school students who engage in specified dangerous behaviors as reported by the Centers for Disease Control. What's wrong with this plot?



Try adding up the percentages. Or look at the 50% slice. Does it look right? Then think: What are these percentages of? Is there a "whole" that has been sliced up? In a pie chart, the proportions shown by each slice of the pie must add up to 100%, and each individual must fall into only one category. Of course, showing the pie on a slant makes it even harder to detect the error.

Here's another example. This bar chart shows the number of airline passengers searched by security screening.



Looks like things didn't change much in the final years of the 20th century—until you read the bar labels and see that the last three bars represent single years, while all the others are for *pairs* of years. The false depth makes it even harder to see the problem.

- **Don't confuse percentages.** Many percentages based on a conditional and joint distributions sound similar, but are different (see Table 4.5):
  - The percentage of French who answered "Agree Completely": This is 347/1539 or 22.55%.
  - The percentage of those who answered "Don't Know" who were French: This is 15/80 or 18.75%.
  - The percentage of those who were French *and* answered "Agree Completely": This is 347/7690 or 4.51%.

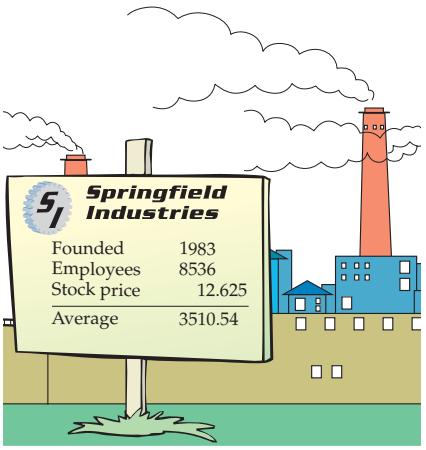
(continued)

In each instance, pay attention to the wording that makes a restriction to a smaller group (those who are French, those who answered "Don't Know," and all respondents, respectively) before a percentage is found. This restricts the *who* of the problem and the associated denominator for the percentage. Your discussion of results must make these differences clear.

- **Don't forget to look at the variables separately, too.** When you make a contingency table or display a conditional distribution, be sure to also examine the marginal distributions. It's important to know how many cases are in each category.
- **Be sure to use enough individuals.** When you consider percentages, take care that they are based on a large enough number of individuals (or cases). Take care not to make a report such as this one:

  We found that 66.67% of the companies surveyed improved their performance by hiring outside consultants. The other company went bankrupt.

- **Don't overstate your case.** Independence is an important concept, but it is rare for two variables to be *entirely* independent. We can't conclude that one variable has no effect whatsoever on another. Usually, all we know is that little effect was observed in our study. Other studies of other groups under other circumstances could find different results.
- **Don't use unfair or inappropriate percentages.** Sometimes percentages can be misleading. Sometimes they don't make sense at all. Be careful when finding percentages across different categories not to combine percentages inappropriately. The next section gives an example.



Simpson's Paradox

Here's an example showing that combining percentages across very different values or groups can give absurd results. Suppose there are two sales representatives, Peter and Katrina. Peter argues that he's the better salesperson, since he managed to close 83% of his last 120 prospects compared with Katrina's 78%. But let's look at the data a little more closely. Here (Table 4.8) are the results for each of their last 120 sales calls, broken down by the product they were selling.

	Product		
Sales Rep	**Printer Paper**	**USB Flash Drive**	**Overall**
Peter	90 out of 100 90%	10 out of 20 50%	100 out of 120 83%
Katrina	19 out of 20 95%	75 out of 100 75%	94 out of 120 78%

Table 4.8 Look at the percentages within each Product category. Who has a better success rate closing sales of paper? Who has the better success rate closing sales of Flash Drives? Who has the better performance overall?

Look at the sales of the two products separately. For printer paper sales, Katrina had a 95% success rate, and Peter only had a 90% rate. When selling flash drives, Katrina closed her sales 75% of the time, but Peter only 50%. So Peter has better "overall" performance, but Katrina is better selling each product. How can this be?

This problem is known as **Simpson's Paradox**, named for the statistician who described it in the 1960s. Although it is rare, there have been a few well-publicized cases of it. As we can see from the example, the problem results from inappropriately combining percentages of different groups. Katrina concentrates on selling flash drives, which is more difficult, so her *overall* percentage is heavily influenced by her flash drive average. Peter sells more printer paper, which appears to be easier to sell. With their different patterns of selling, taking an overall percentage is misleading. Their manager should be careful not to conclude rashly that Peter is the better salesperson.

The lesson of Simpson's Paradox is to be sure to combine comparable measurements for comparable individuals. Be especially careful when combining across different levels of a second variable. It's usually better to compare percentages *within* each level, rather than across levels.

Discrimination?

One famous example of Simpson's Paradox arose during an investigation of admission rates for men and women at the University of California at Berkeley's graduate schools. As reported in an article in *Science*, about 45% of male applicants were admitted, but only about 30% of female applicants got in. It looked like a clear case of discrimination. However, when the data were broken down by school (Engineering, Law, Medicine, etc.), it turned out that within each school, the women were admitted at nearly the same or, in some cases, much *higher* rates than the men. How could this be? Women applied in large numbers to schools with very low admission rates. (Law and Medicine, for example, admitted fewer than 10%.) Men tended to apply to Engineering and Science. Those schools have admission rates above 50%. When the total applicant pool was combined and the percentages were computed, the women had a much lower *overall* rate, but the combined percentage didn't really make sense.

Ethics in Action

Lyle Erhart has been working in sales for a leading vendor of Customer Relationship Management (CRM) software for the past three years. He was recently made aware of a published research study that examined factors related to the successful implementation of CRM projects among firms in the financial services industry. Lyle read the research report with interest and was excited to see that his company's CRM software product was included. Among the results were tables reporting the number of projects that were successful based on type of CRM implementation (Operational versus Analytical) for each of the top leading CRM products of 2006. Lyle quickly found the results for his company's product and their major competitor. He summarized the results into one table as follows:

	His Company	Major Competitor
Operational	16 successes out of 20	68 successes out of 80
Analytical	90 successes out of 100	19 successes out of 20

At first he was a bit disappointed, especially since most of their potential clients were interested in Operational CRM. He had hoped to be able to disseminate the findings of this report among the sales force so they could refer to it when visiting potential clients. After some thought, he realized that he could combine the results. His company's overall success rate was 106 out of 120 (over 88%) and was higher than that of its major competitor. Lyle was now happy that he found and read the report.

ETHICAL ISSUE *Lyle, intentionally or not, has benefited from Simpson's Paradox. By combining percentages, he can present the findings in a manner favorable to his company (related to item A, ASA Ethical Guidelines).*

ETHICAL SOLUTION *Lyle should not combine the percentages as the results are misleading. If he decides to disseminate the information to his sales force, he must do so without combining.*

What Have We Learned?

Learning Objectives

- Make and interpret a frequency table for a categorical variable.
  - We can summarize categorical data by counting the number of cases in each category, sometimes expressing the resulting distribution as percentages.
- Make and interpret a bar chart or pie chart.
  - We display categorical data using the area principle in either a **bar chart** or a **pie chart**.
- Make and interpret a contingency table.
  - When we want to see how two categorical variables are related, we put the counts (and/or percentages) in a two-way table called a **contingency table**.
- Make and interpret bar charts and pie charts of marginal distributions.
  - We look at the **marginal distribution** of each variable (found in the margins of the table). We also look at the **conditional distribution** of a variable within each category of the other variable.
  - Comparing conditional distributions of one variable across categories of another tells us about the association between variables. If the conditional distributions of one variable are (roughly) the same for every category of the other, the variables are **independent**.

Terms

Area principle
In a statistical display, each data value is represented by the same amount of area.

Bar chart (relative frequency bar chart)
A chart that represents the count (or percentage) of each category in a categorical variable as a bar, allowing easy visual comparisons across categories.

Cell
Each location in a contingency table, representing the values of two categorical variables, is called a cell.

Segmented bar chart
A segmented bar chart displays the conditional distribution of a categorical variable within each category of another variable.

Column percent
The proportion of each column contained in the cell of a frequency table.

Conditional distribution
The distribution of a variable restricting the *who* to consider only a smaller group of individuals.

Contingency table
A table displaying the frequencies (sometimes percentages) for each combination of two or more variables.

Distribution
The distribution of a variable is a list of:

- all the possible values of the variable
- the relative frequency of each value

Frequency table (relative frequency table)
A table that lists the categories in a categorical variable and gives the number (the percentage) of observations for each category. The row percent is the proportion of each row contained in the cell of a frequency table, while the column percent is the proportion of each column contained in the cell of a frequency table.

Independent variables
Variables for which the conditional distribution of one variable is the same for each category of the other.

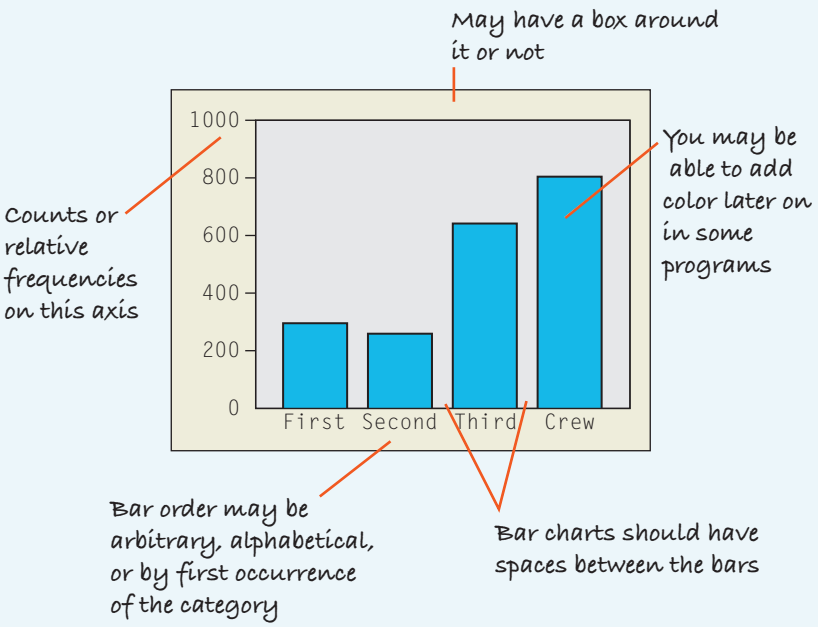
Marginal distribution
In a contingency table, the distribution of either variable alone. The counts or percentages are the totals found in the margins (usually the right-most column or bottom row) of the table.

Pie chart
Pie charts show how a "whole" divides into categories by showing a wedge of a circle whose area corresponds to the proportion in each category.

Row percent	The proportion of each row contained in the cell of a frequency table.
Simpson's paradox	A phenomenon that arises when averages, or percentages, are taken across different groups, and these group averages appear to contradict the overall averages.
Total percent	The proportion of the total contained in the cell of a frequency table.

Technology Help: Displaying Categorical Data on the Computer

Although every package makes a slightly different bar chart, they all have similar features:



Sometimes the count or a percentage is printed above or on top of each bar to give some additional information. You may find that your statistics package sorts category names in annoying orders by default. For example, many packages sort categories alphabetically or by the order the categories are seen in the data set. Often, neither of these is the best choice.

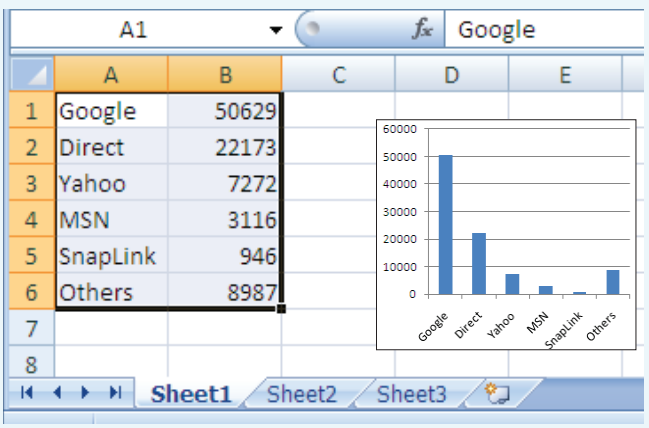
EXCEL 2007 XLSTAT

To make a bar chart:

- Select the variable in Excel you want to work with.
- Choose the **Column** command from the Insert tab in the Ribbon.
- Select the appropriate chart from the drop down dialog.

To change the bar chart into a pie chart:

- Right-click the chart and select **Change Chart Type**. . . from the menu. The Chart type dialog opens.
- Select a pie chart type.
- Click the **OK** button. Excel changes your bar chart into a pie chart.



(continued)

JMP

JMP makes a bar chart and frequency table together.

- From the **Analyze** menu, choose **Distribution.**
- In the Distribution dialog, drag the name of the variable into the empty variable window beside the label "Y, Columns"; click **OK.**
- To make a pie chart, choose **Chart** from the **Graph** menu.
- In the Chart dialog, select the variable name from the Columns list, click on the button labeled "Statistics," and select "N" from the drop-down menu.
- Click the "**Categories, X, Levels**" button to assign the same variable name to the X-axis.
- Under Options, click on the **second** button—labeled "**Bar Chart**"—and select "Pie" from the drop-down menu.

MINITAB

To make a bar chart,

- Choose **Bar Chart** from the **Graph** menu.
- Then select a Simple, Cluster, or Stack chart from the options and click **OK**.
- To make a **Simple** bar chart, enter the name of the variable to graph in the dialog box.
- To make a relative frequency chart, click **Chart Options**, and choose **Show Y as Percent**.
- In the Chart dialog, enter the name of the variable that you wish to display in the box labeled "Categorical variables."
- Click **OK**.

SPSS

To make a bar chart,

- Open the **Chart Builder** from the **Graphs** menu.
- Click the **Gallery** tab.
- Choose **Bar Chart** from the list of chart types.
- Drag the appropriate bar chart onto the canvas.
- Drag a categorical variable onto the x-axis drop zone.
- Click **OK**.

Comments

A similar path makes a pie chart by choosing **Pie chart** from the list of chart types.

KEEN

More of the data that KEEN, Inc. obtained from *Google Analytics* are in the file **KEEN**. Open the data file using a statistics package and find data on *Country of Origin*, *Top Keywords*, *Online Retailers*, *User Statistics*, and *Page Visits*. Create frequency tables, bar charts, and pie charts using your software. What might KEEN want to know about their Web traffic? Which of these tables and charts is most useful to address the question of where they should advertise and how they should position their products? Write a brief case report summarizing your analysis and results.

Exercises

SECTION 4.1

1. As part of the human resource group of your company you are asked to summarize the educational levels of the 512 employees in your division. From company records, you find that 164 have no college degree (None), 42 have an associate's degree (AA), 225 have a bachelor's degree (BA), 52 have a master's degree (MA), and 29 have PhDs. For the educational level of your division:

a) Make a frequency table.
b) Make a relative frequency table.

2. As part of the marketing group at Pixar, you are asked to find out the age distribution of the audience of Pixar's latest film. With the help of 10 of your colleagues, you conduct exit interviews by randomly selecting people to question at 20 different movie theatres. You ask them to tell you if they are younger than 6 years old, 6 to 9 years old, 10 to 14 years old, 15 to 21 years old, or older than 21. From 470 responses, you find out that 45 are younger than 6, 83 are 6 to 9 years old, 154 are 10 to 14, 18 are 15 to 21, and 170 are older than 21. For the age distribution:

a) Make a frequency table.
b) Make a relative frequency table.

SECTION 4.2

3. From the educational level data described in Exercise 1:

a) Make a bar chart using counts on the *y*-axis.
b) Make a relative frequency bar chart using percentages on the *y*-axis.
c) Make a pie chart.

4. From the age distribution data described in Exercise 2:

a) Make a bar chart using counts on the *y*-axis.
b) Make a relative frequency bar chart using percentages on the *y*-axis.
c) Make a pie chart.

5. For the educational levels described in Exercise 1:

a) Write two to four sentences summarizing the distribution.
b) What conclusions, if any, could you make about the educational level at other companies?

6. For the ages described in Exercise 2:

a) Write two to four sentences summarizing the distribution.
b) What possible problems do you see in concluding that the age distribution from these surveys accurately represents the ages of the national audience for this film?

SECTION 4.3

7. From Exercise 1, we also have data on how long each person has been with the company (tenure) categorized into three levels: less than 1 year, between 1 and 5 years, and more than 5 years. A table of the two variables together looks like:

	None	AA	BA	MA	PhD
<1 year	10	3	50	20	12
1–5 years	42	9	112	27	15
more than 5 years	112	30	63	5	2

a) Find the marginal distribution of the tenure. (*Hint*: find the row totals.)
b) Verify that the marginal distribution of the education level is the same as that given in Exercise 1.

8. In addition to their age levels, the movie audiences in Exercise 2 were also asked if they had seen the movie before (Never, Once, More than Once). Here is a table showing the responses by age group:

	Under 6	6 to 9	10 to 14	15 to 21	Over 21
Never	39	60	84	16	151
Once	3	20	38	2	15
More than Once	3	3	32	0	4

a) Find the marginal distribution of their previous viewing of the movie. (Hint: find the row totals.)
b) Verify that the marginal distribution of the ages is the same as that given in Exercise 2.

9. For the table in Exercise 7,

a) Find the column percentages.
b) Looking at the column percentages in part a, does the *tenure* distribution (how long the employee has been with the company) for each educational level look the same? Comment briefly.
c) Make a stacked bar chart showing the *tenure* distribution for each educational level.
d) Is it easier to see the differences in the distributions using the column percentages or the stacked bar chart?

10. For the table in Exercise 8,

a) Find the column percentages.
b) Looking at the column percentages in part a, does the distribution of how many times someone has seen the movie look the same for each age group? Comment briefly.
c) Make a stacked bar chart, showing the distribution of viewings for each age level.

d) Is it easier to see the differences in the distributions using the column percentages or the stacked bar chart?

CHAPTER EXERCISES

11. Graphs in the news. Find a bar graph of categorical data from a business publication (e.g., *Business Week*, *Fortune*, *The Wall Street Journal*, etc.).

a) Is the graph clearly labeled?
b) Does it violate the area principle?
c) Does the accompanying article tell the W's of the variable?
d) Do you think the article correctly interprets the data? Explain.

12. Graphs in the news, part 2. Find a pie chart of categorical data from a business publication (e.g., *Business Week*, *Fortune*, *The Wall Street Journal*, etc.).

a) Is the graph clearly labeled?
b) Does it violate the area principle?
c) Does the accompanying article tell the W's of the variable?
d) Do you think the article correctly interprets the data? Explain.

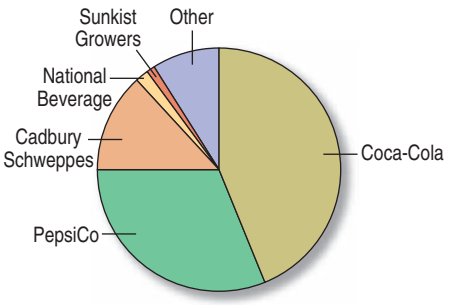
13. Tables in the news. Find a frequency table of categorical data from a business publication (e.g., *Business Week*, *Fortune*, *The Wall Street Journal*, etc.).

a) Is it clearly labeled?
b) Does it display percentages or counts?
c) Does the accompanying article tell the W's of the variable?
d) Do you think the article correctly interprets the data? Explain.

14. Tables in the news, part 2. Find a contingency table of categorical data from a business publication (e.g., *Business Week*, *Fortune*, *The Wall Street Journal*, etc.).

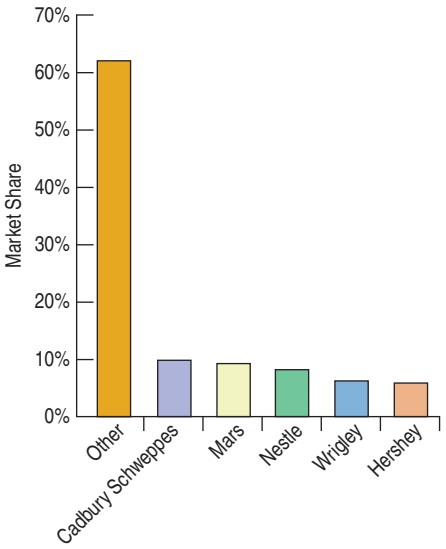
a) Is it clearly labeled?
b) Does it display percentages or counts?
c) Does the accompanying article tell the W's of the variable?
d) Do you think the article correctly interprets the data? Explain.

15. U.S. market share. An article in the *The Wall Street Journal* (March 16, 2007) reported the 2006 U.S. market share of leading sellers of carbonated drinks, summarized in the following pie chart:



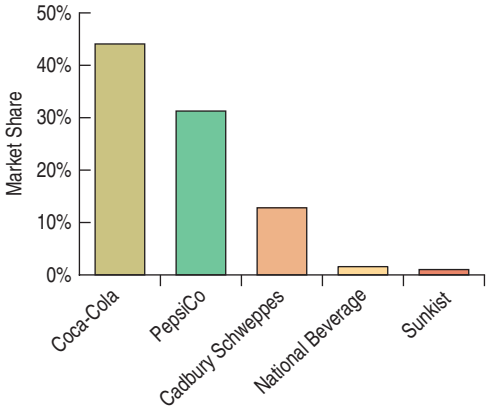
a) Is this an appropriate display for these data? Explain.
b) Which company had the largest share of the market?

16. World market share. *The Wall Street Journal* article described in Exercise 15 also indicated the 2005 world market share for leading distributors of total confectionery products. The following bar chart displays the values:



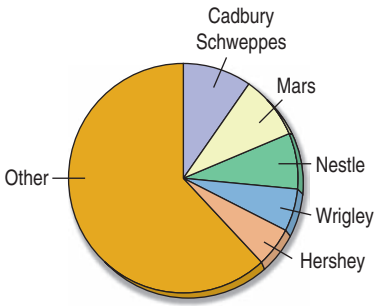
a) Is this an appropriate display for these data? Explain.
b) Which company had the largest share of the candy market?

17. Market share again. Here's a bar chart of the data in Exercise 15.



a) Compared to the pie chart in Exercise 15, which is better for displaying the relative portions of market share? Explain.
b) What is missing from this display that might make it misleading?

18. World market share again. Here's a pie chart of the data in Exercise 16.



a) Which display of these data is best for comparing the market shares of these companies? Explain.
b) Does Cadbury Schweppes or Mars have a bigger market share?

19. Insurance company. An insurance company is updating its payouts and cost structure for their insurance policies. Of particular interest to them is the risk analysis for customers currently on heart or blood pressure medication. The Centers for Disease Control lists causes of death in the United States during one year as follows.

Cause of Death	Percent
Heart disease	30.3
Cancer	23.0
Circulatory diseases and stroke	8.4
Respiratory diseases	7.9
Accidents	4.1

a) Is it reasonable to conclude that heart or respiratory diseases were the cause of approximately 38% of U.S. deaths during this year?
b) What percent of deaths were from causes not listed here?
c) Create an appropriate display for these data.

20. Revenue growth. A 2005 study by Babson College and The Commonwealth Institute surveyed the top women-led businesses in the state of Massachusetts in 2003 and 2004. The study reported the following results for continuing participants with a 9% response rate. (Does not add up to 100% due to rounding.)

2003–2004 Revenue Growth	
Decline	7%
Modest Decline	9%
Steady State	10%
Modest Growth	18%
Growth	54%

a) Describe the distribution of companies with respect to revenue growth.
b) Is it reasonable to conclude that 72% of all women-led businesses in the U.S. reported some level of revenue growth? Explain.

21. Web conferencing. Cisco Systems Inc. announced plans in March 2007 to buy WebEx Communications, Inc. for $3.2 billion, demonstrating their faith in the future of Web conferencing. The leaders in market share for the vendors in the area of Web conferencing in 2006 are as follows: WebEx 58.4% and Microsoft 26.3%. Create an appropriate graphical display of this information and write a sentence or two that might appear in a newspaper article about the market share.

22. Mattel. In their 2006 annual report, Mattel Inc. reported that their domestic market sales were broken down as follows: 44.1% Mattel Girls and Boys brand, 43.0% Fisher-Price brand and the rest of the nearly $3.5 billion revenues were due to their American Girl brand. Create an appropriate graphical display of this information and write a sentence or two that might appear in a newspaper article about their revenue breakdown.

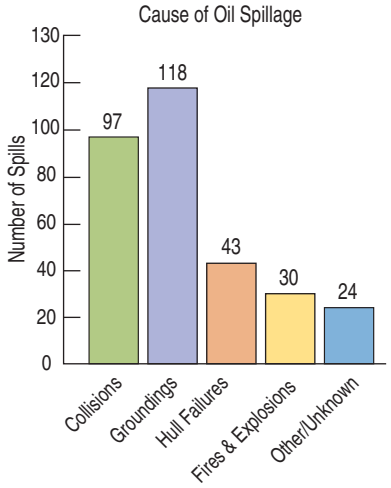
23. Small business productivity. The Wells Fargo/Gallup Small Business Index asked 592 small business owners in March 2004 what steps they had taken in the past year to increase productivity. They found that 60% of small business owners had updated their computers, 52% had made other (noncomputer) capital investments, 37% hired part-time instead of full-time workers, 24% had not replaced workers who left voluntarily, 15% had laid off workers, and 10% had lowered employee salaries.

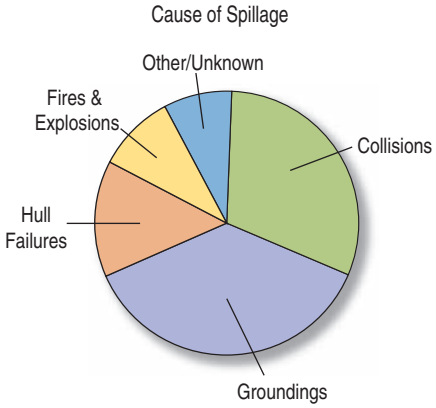
a) What do you notice about the percentages listed? How could this be?
b) Make a bar chart to display the results and label it clearly.
c) Would a pie chart be an effective way of communicating this information? Why or why not?
d) Write a couple of sentences on the steps taken by small businesses to increase productivity.

24. Small business hiring. In 2004, the Wells Fargo/Gallup Small Business Index found that 86% of the 592 small business owners they surveyed said their productivity for the previous year had stayed the same or increased and most had substituted productivity gains for labor. (See Exercise 23.) As a follow-up question, the survey gave them a list of possible economic outcomes and asked if that would make them hire more employees. Here are the percentages of owners saying that they would "definitely or probably hire more employees" for each scenario: a substantial increase in sales—79%, a major backlog of sales orders—71%, a general improvement in the economy—57%, a gain in productivity—50%, a reduction in overhead costs—43%, and more qualified employees available—39%.

a) What do you notice about the percentages listed?
b) Make a bar chart to display the results and label it clearly.
c) Would a pie chart be an effective way of communicating this information? Why or why not?
d) Write a couple of sentences on the responses to small business owners about hiring given the scenarios listed.

25. Environmental hazard. Data from the International Tanker Owners Pollution Federation Limited (www.itopf.com) give the cause of spillage for 312 large oil tanker accidents from 1974–2006. Here are the displays. Write a brief report interpreting what the displays show. Is a pie chart an appropriate display for these data? Why or why not?



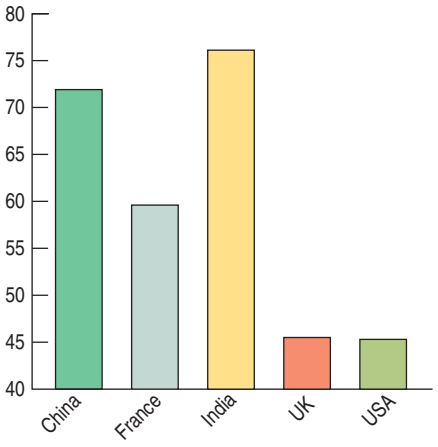


26. Winter Olympics. Twenty-six countries won medals in the 2010 Winter Olympics. The following table lists them, along with the total number of medals each won.

a) Try to make a display of these data. What problems do you encounter?
b) Can you find a way to organize the data so that the graph is more successful?

Country	Medals	Country	Medals
United States	37	Poland	6
Germany	30	Italy	5
Canada	26	Japan	5
Norway	23	Finland	5
Austria	16	Australia	3
Russia	15	Belarus	3
South Korea	14	Slovakia	3
China	11	Croatia	3
Sweden	11	Slovenia	3
France	11	Latvia	2
Switzerland	9	Great Britain	1
Netherlands	8	Estonia	1
Czech Republic	6	Kazakhstan	1

27. Importance of wealth. GfK Roper Reports Worldwide surveyed people in 2004, asking them "How important is acquiring wealth to you?" The percent who responded that it was of more than average importance were: 71.9% China, 59.6% France, 76.1% India, 45.5% UK, and 45.3% USA. There were about 1500 respondents per country. A report showed the following bar chart of these percentages.



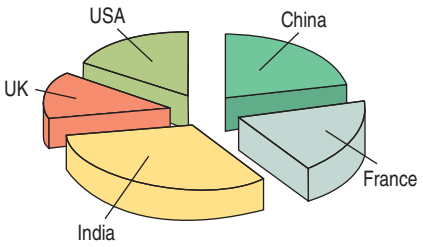
a) How much larger is the proportion of those who said acquiring wealth was important in India than in the United States?
b) Is that the impression given by the display? Explain.
c) How would you improve this display?
d) Make an appropriate display for the percentages.
e) Write a few sentences describing what you have learned about attitudes toward acquiring wealth.

28. Importance of power. In the same survey as that discussed in Exercise 27, GfK Roper Consulting also asked

"How important is having control over people and resources to you?" The percent who responded that it was of more than average importance are given in the following table:

China	49.1%
France	44.1%
India	74.2%
UK	27.8%
USA	36.0%

Here's a pie chart of the data:



a) List the errors you see in this display.
b) Make an appropriate display for the percentages.
c) Write a few sentences describing what you have learned about attitudes toward acquiring power.

29. **Google financials.** Google Inc. derives revenue from three major sources: advertising revenue from their websites, advertising revenue from the thousands of third-party websites that comprise the Google Network, and licensing and miscellaneous revenue. The following table shows the percentage of all revenue derived from these sources for the period 2002 to 2006.

Revenue Source	Year				
	2002	2003	2004	2005	2006
Google websites	70%	54%	50%	55%	60%
Google network websites	24%	43%	49%	44%	39%
Licensing & other revenue	6%	3%	1%	1%	1%

a) Are these row or column percentages?
b) Make an appropriate display of these data.
c) Write a brief summary of this information.

30. **Real estate pricing.** A study of a sample of 1057 houses in upstate New York reports the following percentages of houses falling into different Price and Size categories.

Size	Price			
	Low	Med Low	Med High	High
Small	61.5%	35.2%	5.2%	2.4%
Med Small	30.4%	45.3%	26.4%	4.7%
Med Large	5.4%	17.6%	47.6%	21.7%
Large	2.7%	1.9%	20.8%	71.2%

a) Are these column, row, or total percentages? How do you know?
b) What percent of the highest priced houses were small?
c) From this table, can you determine what percent of all houses were in the low price category?
d) Among the lowest prices houses, what percent were small or medium small?
e) Write a few sentences describing the association between *Price* and *Size*.

31. **Stock performance.** The following table displays information for 40 widely held U.S. stocks, on how their one-day change on March 15, 2007 compared with their previous 52-week change.

MARCH 15, 2007	Over prior 52 weeks	
	Positive Change	Negative Change
Positive Change	14	9
Negative Change	11	6

a) What percent of the companies reported a positive change in their stock price over the prior 52 weeks?
b) What percent of the companies reported a positive change in their stock price over both time periods?
c) What percent of the companies reported a negative change in their stock price over both time periods?
d) What percent of the companies reported a positive change in their stock price over one period and then a negative change in the other period?
e) Among those companies reporting a positive change in their stock price over the prior day what percentage also reported a positive change over the prior year?
f) Among those companies reporting a negative change in their stock price over the prior day what percentage also reported a positive change over the prior year?
g) What relationship, if any, do you see between the performance of a stock on a single day and its 52-week performance?

32. **New product.** A company started and managed by business students is selling campus calendars. The students have conducted a market survey with the various campus constituents to determine sales potential and identify which market segments should be targeted. (Should they advertise in the Alumni Magazine and/or the local newspaper?) The following table shows the results of the market survey.

		Buying Likelihood			
		Unlikely	Moderately Likely	Very Likely	Total
Campus Group	Students	197	388	320	**905**
	Faculty/Staff	103	137	98	**338**
	Alumni	20	18	18	**56**
	Town Residents	13	58	45	**116**
	Total	**333**	**601**	**481**	**1415**

a) What percent of all these respondents are alumni?
b) What percent of these respondents are very likely to buy the calendar?
c) What percent of the respondents who are very likely to buy the calendar are alumni?
d) Of the alumni, what percent are very likely to buy the calendar?
e) What is the marginal distribution of the campus constituents?
f) What is the conditional distribution of the campus constituents among those very likely to buy the calendar?
g) Does this study present any evidence that this company should focus on selling to certain campus constituents?

33. **Real estate.** *The Wall Street Journal* reported in March 2007 that the real estate market in Nashville, Tennessee, slowed slightly from 2006 to 2007. The supporting data are summarized in the following table.

		Type of Sale				
		Condos	Farms/ Land	Residential	Multi-family	Total
Year	2006	266	177	2119	48	**2610**
	2007	341	190	2006	38	**2575**
	Total	**607**	**367**	**4125**	**86**	**5185**

a) What percent of all sales in February 2006 were condominiums (condos)? In February 2007?
b) What percent of the sales in February 2006 were multifamily? In February 2007?
c) What was the change in the percent of residential real estate sales in Nashville, Tennessee, from February 2006 to February 2007?

34. **Google financials, part 2.** Google Inc. divides their total costs and expenses into five categories: cost of revenues, research and development, sales and marketing, general administrative, and miscellaneous. See the table at the bottom of the page.

a) What percent of all costs and expenses were cost of revenues in 2005? In 2006?
b) What percent of all costs and expenses were due to research and development in 2005? In 2006?
c) Have general administrative costs grown as a percentage of all costs and expenses over this time period?

35. **Movie ratings.** The movie ratings system is a voluntary system operated jointly by the Motion Picture Association of America (MPAA) and the National Association of Theatre Owners (NATO). The ratings themselves are given by a board of parents who are members of the Classification and Ratings Administration (CARA). The board was created in response to outcries from parents in the 1960s for some kind of regulation of film content, and the first ratings were introduced in 1968. Here is information on

		2002	2003	2004	2005	2006
Cost and Expenses	Cost of revenues	$132,575	$634,411	$1,468,967	$2,577,088	$4,225,027
	Research and development	$40,494	$229,605	$385,164	$599,510	$1,228,589
	Sales and marketing	$48,783	$164,935	$295,749	$468,152	$849,518
	General administrative	$31,190	$94,519	$188,151	$386,532	$751,787
	Miscellaneous	$0	$0	$201,000	$90,000	$0
	Total Costs and Expenses	**$253,042**	**$1,123,470**	**$2,539,031**	**$4,121,282**	**$7,054,921**

Table for Exercise 34

the ratings of 120 movies that came out in 2005, also classified by their genre.

		Rating				
		G	PG	PG-13	R	Total
Genre	Action/Adventure	4	5	17	9	35
	Comedy	2	12	20	4	38
	Drama	0	3	8	17	28
	Thriller/Horror	0	0	11	8	19
	Total	6	20	56	38	120

a) Find the conditional distribution (in percentages) of movie ratings for action/adventure films.
b) Find the conditional distribution (in percentages) of movie ratings for thriller/horror films.
c) Create a graph comparing the ratings for the four genres.
d) Are *Genre* and *Rating* independent? Write a brief summary of what these data show about movie ratings and the relationship to the genre of the film.

36. Wireless access. The Pew Internet and American Life Project has monitored access to the Internet since the 1990s. Here is an income breakdown of 798 Internet users surveyed in December 2006, asking whether they have logged on to the Internet using a wireless device or not.

		Wireless Users	Other Internet Users	Total
Income	Under $30K	34	128	162
	$30K–50K	31	133	164
	$50K–$75K	44	72	116
	Over $75K	83	111	194
	Don't know/ refused	51	111	162
	Total	243	555	798

a) Find the conditional distribution (in percentages) of income distribution for the wireless users.
b) Find the conditional distribution (in percentages) of income distribution for other Internet users.
c) Create a graph comparing the income distributions of the two groups.
d) Do you see any differences between the conditional distributions? Write a brief summary of what these data show about wireless use and its relationship to income.

37. MBAs. A survey of the entering MBA students at a university in the United States classified the country of origin of the students, as seen in the table.

		MBA Program		
		Two-Year MBA	Evening MBA	Total
Origin	Asia/Pacific Rim	31	33	64
	Europe	5	0	5
	Latin America	20	1	21
	Middle East/Africa	5	5	10
	North America	103	65	168
	Total	164	104	268

a) What percent of all MBA students were from North America?
b) What percent of the Two-Year MBAs were from North America?
c) What percent of the Evening MBAs were from North America?
d) What is the marginal distribution of origin?
e) Obtain the column percentages and show the conditional distributions of origin by MBA Program.
f) Do you think that origin of the MBA student is independent of the MBA program? Explain.

38. MBAs, part 2. The same university as in Exercise 37 reported the following data on the gender of their students in their two MBA programs.

		Type		
		Two-Year	Evening	Total
Sex	Men	116	66	182
	Women	48	38	86
	Total	164	104	268

a) What percent of all MBA students are women?
b) What percent of Two-Year MBAs are women?
c) What percent of Evening MBAs are women?
d) Do you see evidence of an association between the *Type* of MBA program and the percentage of women students? If so, why do you believe this might be true?

T **39.** Top producing movies. The following table shows the Motion Picture Association of America (MPA) (www.mpaa.org) ratings for the top 20 grossing films in the United States for each of the 10 years from 1999 to 2008. (Data are number of films.)

a) What percent of all these top 20 films are G rated?
b) What percent of all top 20 films in 2005 were G rated?
c) What percent of all top 20 films were PG-13 and came out in 1999?
d) What percent of all top 20 films produced in 2006 or later were PG-13?

Year	Rating				
	G	PG	PG-13	R	Total
2008	2	4	10	4	20
2007	1	5	11	3	20
2006	1	4	13	2	20
2005	1	4	13	2	20
2004	1	6	10	3	20
2003	1	3	11	5	20
2002	1	6	13	0	20
2001	2	4	10	4	20
2000	0	3	12	5	20
1999	2	3	7	8	20
Total	12	42	110	36	200

e) What percent of all top 20 films produced from 1999 to 2002 were rated PG-13 or R?
f) Compare the conditional distributions of the ratings for films produced in 2004 or later to those produced from 1999 to 2003. Write a couple of sentences summarizing what you see.

40. Movie admissions. The following table shows attendance data collected by the Motion Picture Association of America during the period 2002 to 2006. Figures are in millions of movie admissions.

Year	Patron Age						
	12 to 24	25 to 29	30 to 39	40 to 49	50 to 59	60 and over	Total
2006	485	136	246	219	124	124	1334
2005	489	135	194	216	125	122	1281
2004	567	132	265	236	145	132	1477
2003	567	124	269	193	152	118	1423
2002	551	158	237	211	119	130	1406
Total	2659	685	1211	1075	665	626	6921

a) What percent of all admissions during this period were bought by people between the ages of 12 and 24?
b) What percent of admissions in 2003 were bought by people between the ages of 12 and 24?
c) What percent of the admission were bought by people between the ages of 12 and 24 in 2006?
d) What percent of admissions in 2006 were bought by people 60 years old and older?
e) What percent of the admissions bought by people 60 and over were in 2002?
f) Compare the conditional distributions of the age groups across years. Write a couple of sentences summarizing what you see.

41. Tattoos. A study by the University of Texas Southwestern Medical Center examined 626 people to see if there was an increased risk of contracting hepatitis C associated with having a tattoo. If the subject had a tattoo, researchers asked whether it had been done in a commercial tattoo parlor or elsewhere. Write a brief description of the association between tattooing and hepatitis C, including an appropriate graphical display.

	Tatto done in commercial parlor	Tattoo done elsewhere	No tattoo
Has hepatitis C	17	8	18
No hepatitis C	35	53	495

42. Working parents. In July 1991 and again in April 2001, the Gallup Poll asked random samples of 1015 adults about their opinions on working parents. The following table summarizes responses to this question: "*Considering the needs of both parents and children, which of the following do you see as the ideal family in today's society?*" Based upon these results, do you think there was a change in people's attitudes during the 10 years between these polls? Explain.

Response	Year	
	1991	2001
Both work full-time	142	131
One works full-time, other part-time	274	244
One works, other works at home	152	173
One works, other stays home for kids	396	416
No opinion	51	51

43. Revenue growth, last one. The study completed in 2005 and described in Exercise 20 also reported on education levels of the women chief executives. The column percentages for CEO education for each level of revenue are summarized in the following table. (Revenue is in $ million.)

	Graduate Education and Firm Revenue Size		
	< $10 M revenue	$10–$49.999 M revenue	≥ $50 M revenue
% with High School Education only	8%	4%	8%
% with College Education, but no Graduate Education	48%	42%	33%
% with Graduate Education	44%	54%	59%
Total	100%	100%	100%

a) What percent of these CEOs in the highest revenue category had only a high school education?
b) From this table, can you determine what percent of all these CEOs had graduate education? Explain.
c) Among the CEOs in the lowest revenue category, what percent had more than a high school education?
d) Write a few sentences describing the association between *Revenue* and *Education*.

44. Minimum wage workers. The U.S. Department of Labor (www.bls.gov) collects data on the number of U.S. workers who are employed at or below the minimum wage. Here is a table showing the number of hourly workers by *Age* and *Sex* and the number who were paid at or below the prevailing minimum wage:

Age	Hourly Workers (in thousands)		At or Below Minimum Wage (in thousands)	
	Men	Women	Men	Women
16–24	7978	7701	384	738
25–34	9029	7864	150	332
35–44	7696	7783	71	170
45–54	7365	8260	68	134
55–64	4092	4895	35	72
65+	1174	1469	22	50

a) What percent of the women were ages 16 to 24?
b) Using side-by-side bar graphs, compare the proportions of the men and women who worked at or below minimum wage at each *Age* group. Write a couple of sentences summarizing what you see.

45. Moviegoers and ethnicity. The Motion Picture Association of America studies the ethnicity of moviegoers to understand changes in the demographics of moviegoers over time. Here are the numbers of moviegoers (in millions) classified as to whether they were Hispanic, African-American, or Caucasian for the years 2002 to 2006.

		Year					
		2002	2003	2004	2005	2006	Total
Ethnicity	Hispanic	21	23	25	25	26	**120**
	African-American	21	20	22	21	20	**104**
	Caucasian	118	127	127	113	120	**605**
	Total	**160**	**170**	**174**	**159**	**166**	**829**

a) Find the marginal distribution *Ethnicity* of moviegoers.
b) Find the conditional distribution of *Ethnicity* for the year 2006.
c) Compare the conditional distribution of *Ethnicity* for all 5 years with a segmented bar graph.
d) Write a brief description of the association between *Year* and *Ethnicity* among these respondents.

46. Department store. A department store is planning its next advertising campaign. Since different publications are read by different market segments, they would like to know if they should be targeting specific age segments. The results of a marketing survey are summarized in the following table by *Age* and *Shopping Frequency* at their store.

		Age			
	Shopping	Under 30	30–49	50 and Over	Total
Frequency	Low	27	37	31	**95**
	Moderate	48	91	93	**232**
	High	23	51	73	**147**
	Total	**98**	**179**	**197**	**474**

a) Find the marginal distribution of *Shopping Frequency*.
b) Find the conditional distribution of *Shopping Frequency* within each age group.
c) Compare these distributions with a segmented bar graph.
d) Write a brief description of the association between *Age* and *Shopping Frequency* among these respondents.
e) Does this prove that customers ages 50 and over are more likely to shop at this department store? Explain.

47. Women's business centers. A study conducted in 2002 by Babson College and the Association of Women's Centers surveyed women's business centers in the United States. The data showing the location of established centers (at least 5 years old) and less established centers are summarized in the following table.

	Location	
	Urban	Nonurban
Less Established	74%	26%
Established	80%	20%

a) Are these percentages column percentages, row percentages, or table percentages?
b) Use graphical displays to compare these percentages of women's business centers by location.

48. Advertising. A company that distributes a variety of pet foods is planning their next advertising campaign. Since

different publications are read by different market segments, they would like to know how pet ownership is distributed across different income segments. The U.S. Census Bureau reports the number of households owning various types of pets. Specifically, they keep track of dogs, cats, birds, and horses.

Income Distribution of Households Owing Pets (percent)

Income	Pet: Dog	Cat	Bird	Horse
Under $12,500	14	15	16	9
$12,500 to $24,999	20	20	21	21
$25,000 to $39,999	24	23	24	25
$40,000 to $59,999	22	22	21	22
$60,000 and over	20	20	18	23
Total	100	100	100	100

a) Do you think the income distributions of the households who own these different animals would be roughly the same? Why or why not?
b) The table shows the percentages of income levels for each type of animal owned. Are these row percentages, column percentages, or table percentages?
c) Do the data support that the pet food company should not target specific market segments based on household income? Explain.

49. **Worldwide toy sales.** Around the world, toys are sold through different channels. For example, in some parts of the world toys are sold primarily through large toy store chains, while in other countries department stores sell more toys. The following table shows the percentages by region of the distribution of toys sold through various channels in Europe and America in 2003, accumulated by the International Council of Toy Industries (www.toy-icti.org).

a) Are these row percentages, column percentages, or table percentages?
b) Can you tell what percent of toys sold by mail order in both Europe and America are sold in Europe? Why or why not?
c) Use a graphical display to compare the distribution of channels between Europe and America.
d) Summarize the distribution of toy sales by channel in a few sentences. What are the biggest differences between these two continents?

Location	Channel: General Merchandise	Toy Specialists	Department Stores	Mass Merchant Discounters & Food Hypermarkets	Mail Order	Other
America	9%	25%	3%	51%	4%	8%
Europe	13%	36%	7%	24%	5%	15%

Table for Exercise 49

50. **Internet piracy.** Illegal downloading of copyrighted movies is an international problem estimated to have cost the international movie industry more than $18 billion in 2005. The typical pirate worldwide is a 16 to 24-year old male living in an urban area, according to a study by the international strategy consulting firm LEK (www.mpaa.org/researchStatistics.asp). The following table compares the age distribution of the U.S. pirate to the rest of the world.

Region	Age: 16–24	25–29	30–39	Over 40
United States	71	11	7	11
Rest of World	58	15	18	9

a) Are these row percentages, column percentages, or table percentages?
b) Can you tell what percent of pirates worldwide are in the 16 to 24 age group?
c) Use a graphical display to compare the age distribution of pirates in the United States to the distribution in the rest of the world.
d) Summarize the distribution of *Age* by *Region* in a few sentences. What are the biggest differences between these two regions?

51. **Insurance company, part 2.** An insurance company that provides medical insurance is concerned with recent data. They suspect that patients who undergo surgery at large hospitals have their discharges delayed for various reasons—which results in increased medical costs to the insurance company. The recent data for area hospitals and two types of surgery (major and minor) are shown in the following table.

Procedure	Discharge Delayed: Large Hospital	Small Hospital
Major surgery	120 of 800	10 of 50
Minor surgery	10 of 200	20 of 250

a) Overall, for what percent of patients was discharge delayed?
b) Were the percentages different for major and minor surgery?
c) Overall, what were the discharge delay rates at each hospital?
d) What were the delay rates at each hospital for each kind of surgery?
e) The insurance company is considering advising their clients to use large hospitals for surgery to avoid postsurgical complications. Do you think they should do this?
f) Explain, in your own words, why this confusion occurs.

52. **Delivery service.** A company must decide which of two delivery services they will contract with. During a recent trial period, they shipped numerous packages with each service and have kept track of how often deliveries did not arrive on time. Here are the data.

Delivery Service	Type of Service	Number of Deliveries	Number of Late Packages
Pack Rats	Regular	400	12
	Overnight	100	16
Boxes R Us	Regular	100	2
	Overnight	400	28

a) Compare the two services' overall percentage of late deliveries.
b) Based on the results in part a, the company has decided to hire Pack Rats. Do you agree they deliver on time more often? Why or why not? Be specific.
c) The results here are an instance of what phenomenon?

53. **Graduate admissions.** A 1975 article in the magazine *Science* examined the graduate admissions process at Berkeley for evidence of gender bias. The following table shows the number of applicants accepted to each of four graduate programs.

Program	Males Accepted (of Applicants)	Females Accepted (of Applicants)
1	511 of 825	89 of 108
2	352 of 560	17 of 25
3	137 of 407	132 of 375
4	22 of 373	24 of 341
Total	**1022 of 2165**	**262 of 849**

a) What percent of total applicants were admitted?
b) Overall, were a higher percentage of males or females admitted?
c) Compare the percentage of males and females admitted in each program.
d) Which of the comparisons you made do you consider to be the most valid? Why?

54. **Simpson's Paradox.** Develop your own table of data that is a business example of Simpson's Paradox. Explain the conflict between the conclusions made from the conditional and marginal distributions.

Just Checking Answers

1 50.0%
2 40.0%
3 25.0%
4 15.6% Nearsighted, 56.3% Farsighted, 28.1% Need Bifocals
5 18.8% Nearsighted, 62.5% Farsighted, 18.8% Need Bifocals
6 40% of the nearsighted customers are female, while 50% of customers are female.
7 Since nearsighted customers appear less likely to be female, it seems that they may not be independent. (But the numbers are small.)