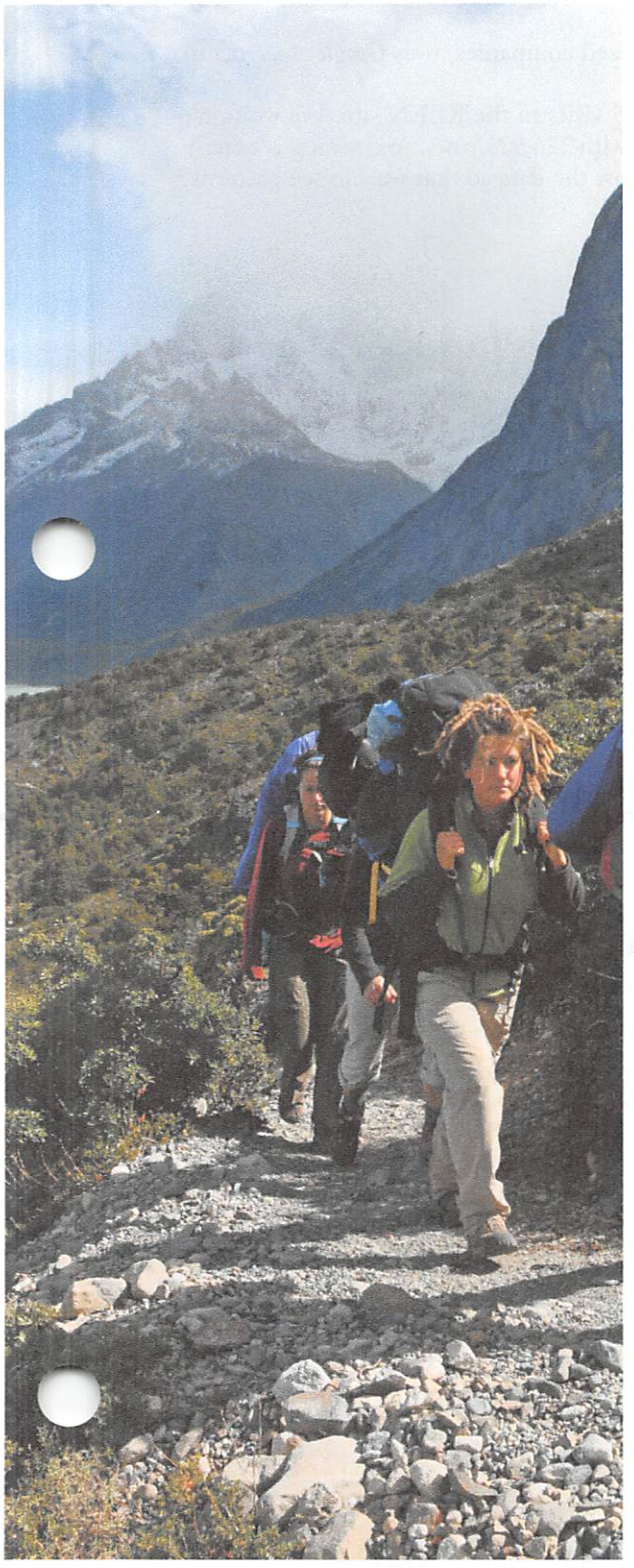
2

Displaying and Describing Categorical Data



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 has in its first nine 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 in Japan, 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	Source (search engine or other) that led to KEEN's website
WHEN	February 2013
WHERE	Worldwide
HOW	Data compiled by KEEN
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 -- [1/Apr/2013:13:15:08-0800]" GET http://
www.keenfootwear.com/us/en/product/shoes/men/cnx/clear
water%20cnx/forest%20night!rust "http://www.google.com/"
"Mozilla/5.0WebTV/1.2 (compatible; MSIE 2.0)"
```

KEEN, like many other small and mid-sized companies, uses *Google Analytics* to collect and summarize its log data.

In February 2013, there were 226,925 visits to the KEEN site. We wouldn't be able to see any patterns in a data table with 226,925 rows. 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.

2.1 Summarizing a Categorical Variable

KEEN might want to know how people reach 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 *Source* records, for each visit to KEEN's website, where the visitor 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. To make sense of the 226,925 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.

Frequency Tables

A **frequency table** records the counts for each of the categories of the variable. Some tables report percentages, and many report both. For example, Table 2.1 shows the ways that customers found their way to the KEEN website.

Source	Visits	Visits by %
Google	130,158	57.36
Direct	52,969	23.34
E-mail	16,084	7.09
Bing	9,581	4.22
Yahoo	7,439	3.28
Facebook	2,253	0.99
Mobile	1,701	0.75
Other	6,740	2.97
Total	**226,925**	**100.00**

Table 2.1 A frequency table of the *Source* used by visitors to the KEEN, Inc. website. Notice the label "Other." 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. (Source: KEEN, Inc., personal communication.)

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 2013 Super Bowl cost about $4M. 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. Polls often ask whether respondents are more interested in the game

or the commercials. Here are 40 responses from one such poll. (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**

100.01%?

Sometimes if you carefully add the percentages of all categories, you will notice the total isn't exactly 100.00% even though we know that that's what the total has to be. The discrepancy is due to individual percentages being rounded. You'll often see this in tables of percents, sometimes with explanatory footnotes.

2.2 Displaying 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.

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

Data visualization has become a special discipline in its own right. A well-designed display can show features of even a large, complex data set. Figure 2.1 on the next page is a specially designed visualization showing the connections between two categorical variables, *College major* and *Career choice*, for 15,600 alumni of Williams College. Innovative visualizations such as this one—many of them interactive or animated—are becoming more common as Big Data is mined for unanticipated patterns and relationships.

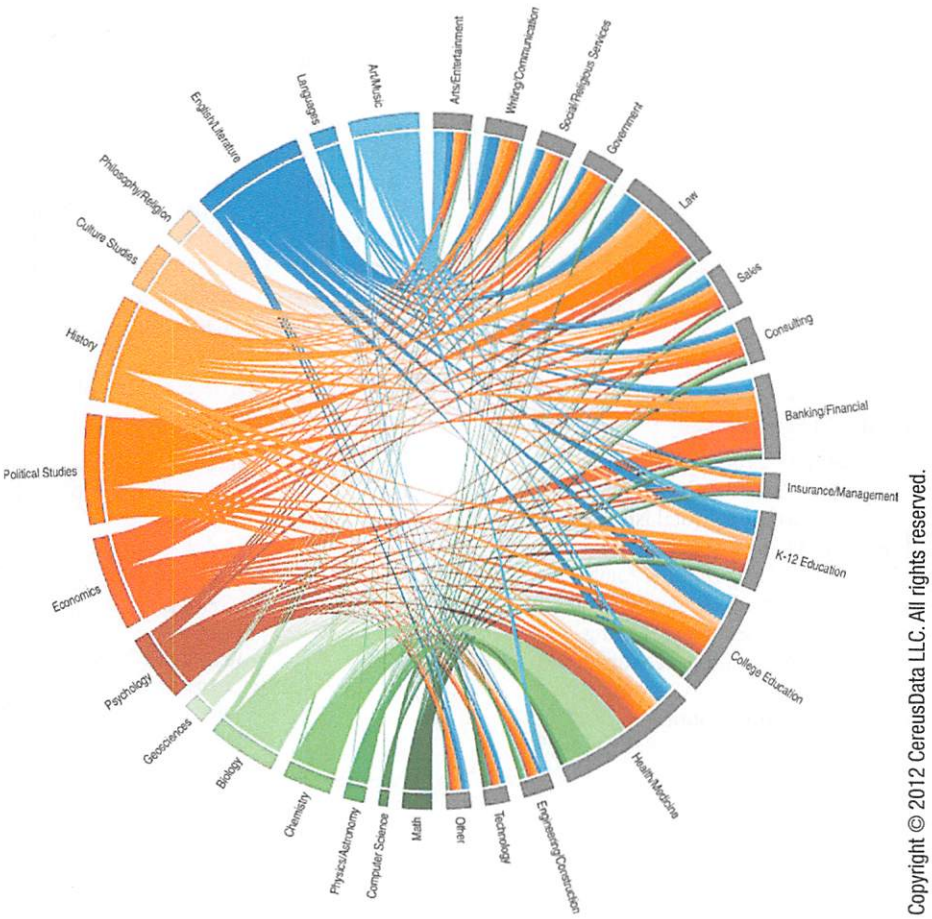


Figure 2.1 Visualization of the link between major in college and career of Williams College alumni. Each individual is graphed as an arc connecting his or her major on the left with a career area on the right. Each major is assigned a color: Humanities in the blue range, Social Sciences in the reds and oranges, and Sciences in greens. It is easy to see the expected large arc connecting Biology and Health/Medicine and the spread of Math majors to many careers. Possibly less expected is that Economics majors choose a wide range of careers. Banking/Finance draws many from Economics, but also quite a few from History, Political Science, and the Humanities. (This image was created by Satyan Devadoss, Hayley Brooks, and Kaison Tanabe using the CIRCOS software; an interactive version of this graph can be found at http://cereusdata.com.)

The Area Principle

We can't make just any display; a bad picture can distort our understanding rather than help it. For example, Figure 2.2 is a graph of the frequencies of Table 2.1. What impression do you get of the relative frequencies of visits from each source? You can easily see from both the table and the figure that the most popular source was

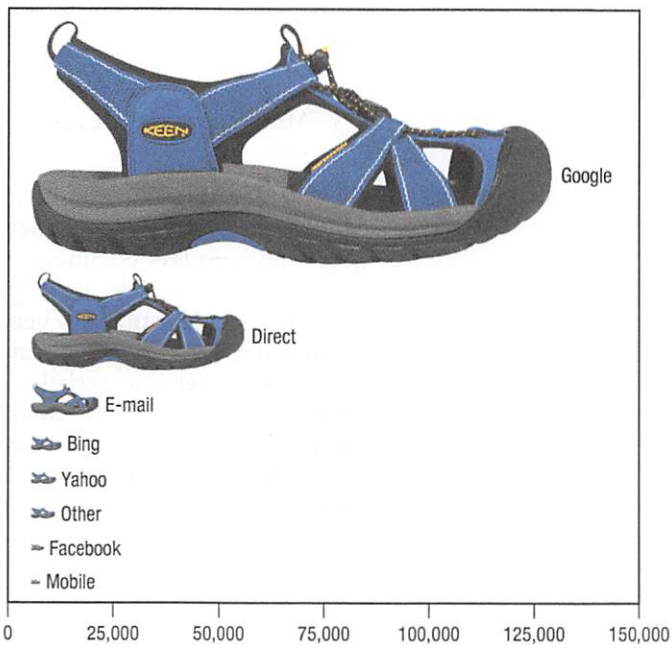


Figure 2.2 Although the length of each sandal corresponds to the correct number, the impression we get is all wrong because we perceive the entire area of the sandal. In fact, only about 57% of all visitors used Google to get to the website.

Google. But the impression given by Figure 2.2 doesn't seem to correspond well to the numbers in the table.

Although it's true that the majority of people came to KEEN's website from Google, in Figure 2.2 it looks like nearly all did. That doesn't seem right. What's wrong? The lengths of the sandals *do* match the frequencies 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 two and a half as many people who came from Google as those who typed in the URL directly, the sandal depicting the number of Google visitors is about two and a half times longer than the sandal below it, but it occupies more than six 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 2.3 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 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. We can also see that there were about two and a half times 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.

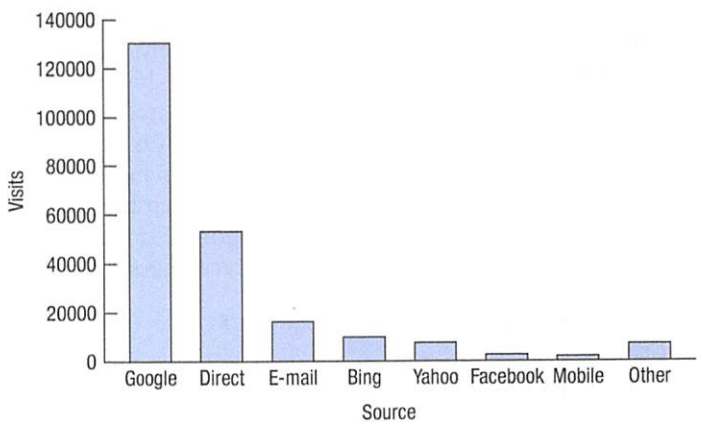
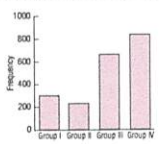
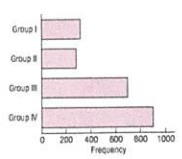


Figure 2.3 Visits to the KEEN, Inc. website by *Source*. With the area principle satisfied, the true distribution is clear.

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 horizontal axis.

Bar charts are usually drawn vertically in columns,





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

[1] Excel refers to this display as a column chart when the bars are vertical and a bar chart when they are horizontal, but that's not standard statistics terminology.

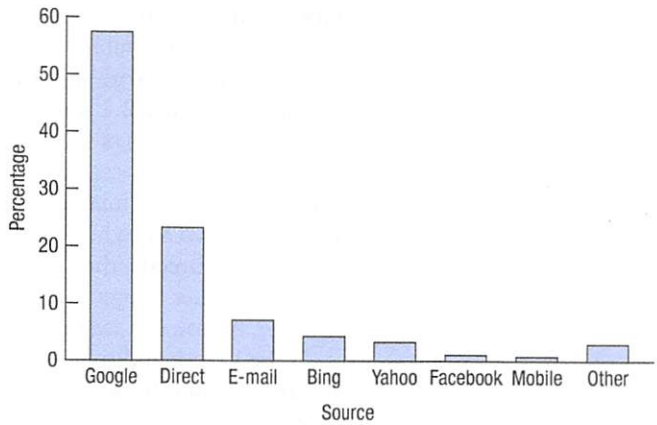


Figure 2.4 The relative frequency bar chart looks the same as the bar chart (Figure 2.3) but shows the proportion of visits in each category rather than the counts.

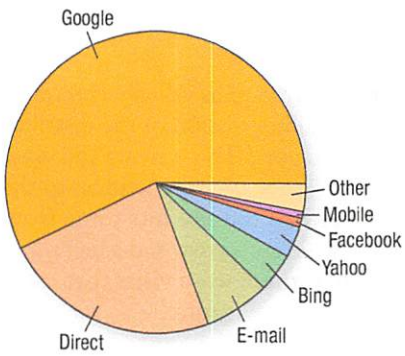


Figure 2.5 A pie chart shows the proportion of visits by *Source*.

If we want to draw attention to the relative *proportion* of visits from each *Source*, we could replace the counts with percentages and use a **relative frequency bar chart**, like the one shown in Figure 2.4.

Pie Charts

A **pie chart** shows how a whole group breaks into several categories. Pie charts show all the cases as a circle sliced into pieces whose areas are proportional to the fraction of cases in each category.

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 2.5, you can easily see that the slice representing Google is just a bit more than half the total. Unfortunately, other comparisons are harder to make with pie charts.

For example, Figure 2.6 shows three pie charts that look pretty much alike along with bar charts of the same data. The bar charts show three distinctly different patterns, but it is almost impossible to see those in the pie charts.

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%.

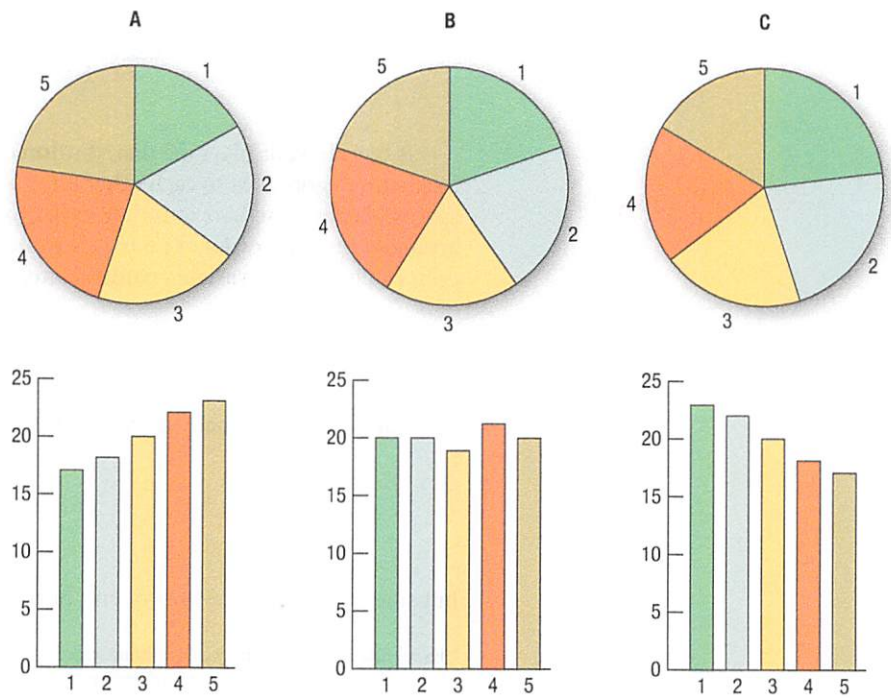
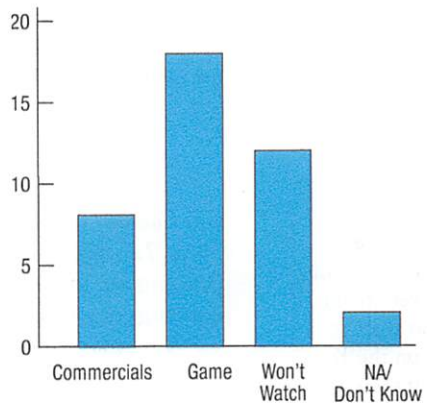


Figure 2.6 Patterns that are easy to see in the bar charts are often hard to see in the corresponding pie charts.

FOR EXAMPLE Making a bar chart

QUESTION Make a bar chart for the 40 Super Bowl responses of the example on pages 20 and 21.

ANSWER Use the frequencies in the table in the example on page 21 to produce the heights of the bars:



2.3 Exploring Two Categorical Variables: Contingency Tables

WHO	Respondents in the Pew Research Worldwide Survey
WHAT	Responses to question about social networking
WHEN	2012
WHERE	Worldwide
HOW	Data collected by Pew Research using a multistage design. For details see www.pewglobal.org/2012/12/12/survey-methods-43/
WHY	To understand penetration of social networking worldwide

In December 2012 Pew Research conducted surveys in countries across the world (www.pewglobal.org/2012/12/12/social-networking-popular-across-globe/). One question of interest to business decision makers is how common it is for citizens of different countries to use social networking and whether they have it available to them. Table 2.2 gives a table of responses for several of the surveyed countries. Note that N/A means "not available" because respondents lacked internet access—a situation that marketers planning for the future might expect to see change.

The pie chart (Figure 2.7) shows clearly that fewer than half of respondents said that they had access to social networking and used it.

But if we want to target our online customer relations with social networks differently in different countries, wouldn't it be more interesting to know how social networking use varies from country to country?

Social Networking	Count	Relative frequency
No	1249	24.787
Yes	2175	43.163
N/A	1615	32.050

Table 2.2 A combined frequency and relative frequency table for the responses from 5 countries (Britain, Egypt, Germany, Russia, and the U.S.) to the question "Do you use social networking sites?" N/A means "Not Available."

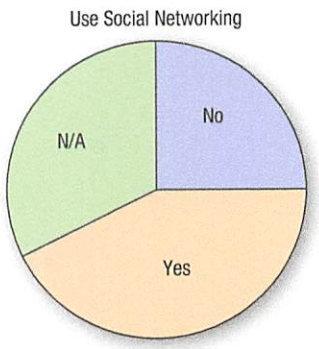


Figure 2.7 Responses to the question "Do you use social networking sites?" N/A means "No Internet Available."

	Britain	Egypt	Germany	Russia	U.S.	Total
No	336	70	460	90	293	**1249**
Yes	529	300	340	500	506	**2175**
N/A	153	630	200	420	212	**1615**
Total	**1018**	**1000**	**1000**	**1010**	**1011**	**5039**

Table 2.3 Contingency table of *Social Networking* and *Country*. The right margin "Totals" are the values that were in Table 2.2.

Percent of What?

The English language can be tricky. If asked, "What percent of those answering 'Yes' were from Russia?" it's pretty clear that you should focus only on the *Yes* row. The question itself seems to restrict the who in the question to that row, so you should look at the number of those in each country among the 2175 people who replied "Yes." You'd find that in the row percentages.

But if you're asked, "What percent were Russians who replied 'yes'?" you'd have a different question. Be careful. That question really means "what percent of the entire sample were both from Russia and replying 'Yes'?", so the *who* is all respondents. The denominator should be 5039, and the answer is the table percent.

Finally, if you're asked, "What percent of the Russians replied 'yes'?" you'd have a third question. Now the *who* is Russians. So the denominator is the 1010 Russians, and the answer is the column percent.

To find out, we need to look at the two categorical variables *Social Networking* and *Country* together, which we do by arranging the data in a two-way table such as Table 2.3. Because they show how individuals are distributed along each variable depending on, or *contingent on*, the value of the other variable, tables like this are called **contingency tables.**

The margins of a contingency table give totals. The totals in the right-hand column of Table 2.3 show the frequency distribution of the variable *Social Networking*. We can see, for example, that Internet access is certainly not yet universal. The totals in the bottom row of the table show the frequency distribution of the variable *Country*—how many respondents Pew obtained in each 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 as its frequency distribution.

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 2.3 we can see that 153 respondents did not have internet access in Britain. Looking across the *Yes* row, you can see that the largest number of responses in that row (529) is from Britain. Are Egyptians less likely to use social media than Britons? Questions like this are more naturally addressed using percentages.

We know that 300 Egyptians report that they use social networking. We could display this count as a percentage, but as a percentage of what? The total number of people in the survey? (300 is 5.95% of the total.) The number of Egyptians surveyed? (300 is 30% of the 1000 Egyptians surveyed.) The number of respondents who use social networking? (300 is 13.8% of social networking users.) 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 2.4) holds lots of information but is hard to understand.

Conditional Distributions

The more interesting questions are contingent on something. We'd like to know, for example, whether these countries are similar in use and availability of social networking. That's the kind of information that could inform a business decision. Table 2.5 shows the distribution of social networking conditional on country.

By comparing the frequencies conditional on *Country*, we can see interesting patterns. For example, Germany stands out as the country in which the largest percentage (46%) have Internet access but don't use social networking ("No"). Russia and Egypt may have more respondents with no Internet access, but those who have

	Britain	Egypt	Germany	Russia	U.S.	Total
No	336	70	460	90	293	1249
	26.9	5.6	36.8	7.2	23.5	100
	33.0	7.0	46.0	8.9	29.0	24.8
	6.7	1.4	9.1	1.8	5.8	24.8
Yes	529	300	340	500	506	2175
	24.3	13.8	15.6	23.0	23.3	100
	52.0	30.0	34.0	49.5	50.0	43.2
	10.5	6.0	6.8	9.9	10.0	43.2
N/A	153	630	200	420	212	1615
	9.5	39.0	12.4	26.0	13.1	100
	15.0	63.0	20.0	41.6	21.0	32.1
	3.0	12.5	4.0	8.3	4.2	32.1
Total	1018	1000	1000	1010	1011	5039
	20.2	19.8	19.8	20.0	20.1	100
	100	100	100	100	100	100
	20.2	19.8	19.8	20.0	20.1	100

Table contents:
Count
Percent of Row Total
Percent of Column Total
Percent of Table Total

Table 2.4 Another contingency table of *Social Networking* and *Country* showing the counts and 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.

Social Networking	Britain	Egypt	Germany	Russia	U.S.	Total
No	336	70	460	90	293	1249
	33.0	7.0	46.0	8.9	29.0	24.8
Yes	529	300	340	500	506	2175
	52.0	30.0	34.0	49.5	50.0	43.2
N/A	153	630	200	420	212	1615
	15.0	63.0	20.0	41.6	21.0	32.1
Total	1018	1000	1000	1010	1011	5039
	100	100	100	100	100	100

Table 2.5 The conditional distribution of *Social Networking* conditioned on 5 values of *Country*. This table shows the column percentages.

access are very likely to use social networking. 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. 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 for 1008 adult U.S. respondents to the question about watching the Super Bowl discussed in the previous For Example.

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

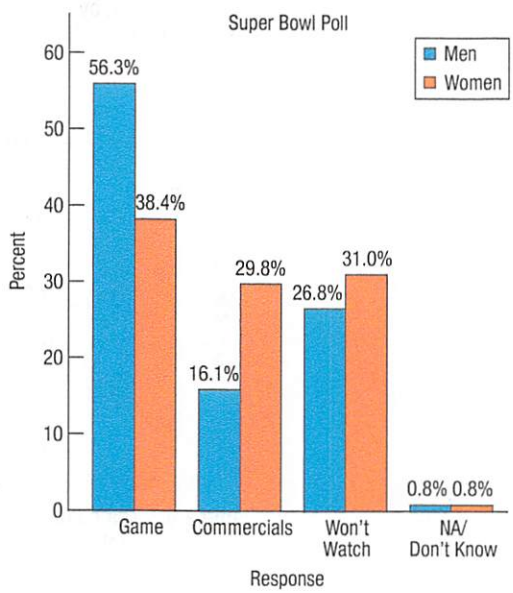
Game = 198/516 = 38.4%

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:



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.

2.4 Segmented Bar Charts and Mosaic Plots

Everyone knows what happened in the North Atlantic on the night of April 14, 1912 as the *Titanic*, thought by many to be unsinkable, sank, leaving almost 1500 passengers and crew members on board to meet their icy fate. Women and children first was the rule for those commanding the lifeboats, but how did the class of ticket held enter into the order?

Here is a contingency table of the 2201 people on board, categorized by *Survival* and *Ticket Class*.

			Class				
			First	Second	Third	Crew	Total
Survival	Alive	Count	203	118	178	212	711
		% of Column	62.5%	41.4%	25.2%	24.0%	32.3%
	Dead	Count	122	167	528	673	1490
		% of Column	37.5%	58.6%	74.8%	76.0%	67.7%
	Total	Count	325	285	706	885	2201
			100%	100%	100%	100%	100%

Table 2.6 A contingency table of *Class* by *Survival* with only counts and column percentages. Each column represents the conditional distribution of *Survival* for a given category of ticket *Class*.

Looking at how the percentages change across each row, it sure looks like ticket class mattered in whether a passenger survived. To make it more vivid, we could display the percentages for surviving and not for each *Class* in a side-by-side bar chart such as in Fig 2.8 on the next page.

Now it's easy to compare the risks. Among first-class passengers, 37.5% perished, compared to 58.6% for second-class ticket holders, 74.8% for those in third class, and 76.0% for crew members.

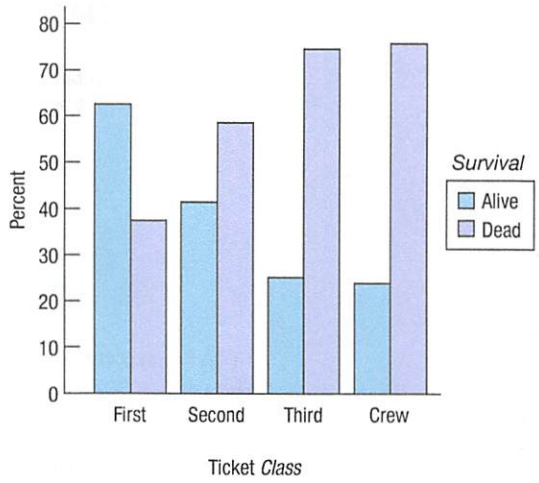


Figure 2.8 *Side-by-side bar chart* showing the conditional distribution of *Survival* for each category of ticket *Class*.

We could also display the *Titanic* information by dividing up bars rather than circles (as we did for pie charts). The resulting **segmented (or stacked) bar chart** treats each bar as the "whole" and divides it proportionally into segments corresponding to the percentage in each group. We can clearly see that the distributions of ticket Class are different, indicating again that survival was not independent of ticket Class.

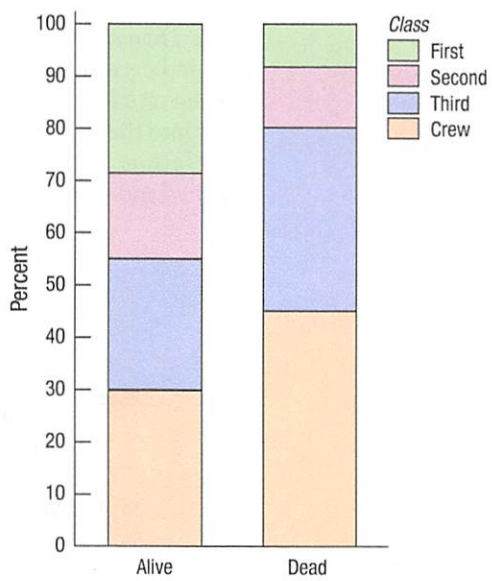


Figure 2.9 A segmented bar chart for *Class by Survival*. Notice that although the totals for survivors and nonsurvivors are quite different, the bars are the same height because we have converted the numbers to *percentages*. Compare this display with the bar chart in Figure 2.8.

A variant of the segmented bar chart, a **mosaic plot** (Figure 2.10), looks like a segmented bar chart, but obeys the area principle better by making the bars proportional to the sizes of the groups. Now, each rectangle is proportional to the number of cases in the data set. Mosaic plots are increasingly popular for displaying contingency tables and are found in many software packages.

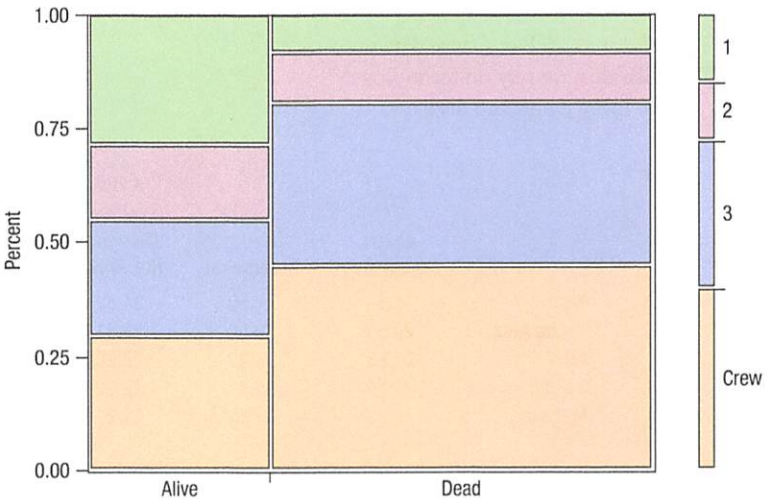


Figure 2.10 A mosaic plot for *Class by Survival*. The plot is just like the segmented bar chart in Figure 2.9 except that the space has been taken out between the categories on the *x*-axis and the rectangles are proportional to the number of cases of the *x* variable as well. We can easily see that the number of survivors was far less than the nonsurvivors, something that we can't in the bar charts.

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 would want to report to the client how concerns about food safety are 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".

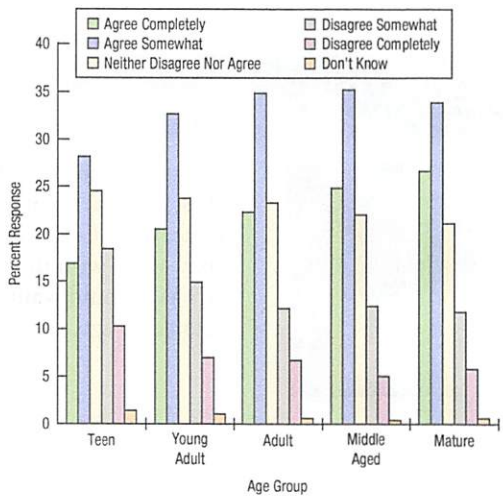
(*continued*)

DO **Mechanics** For a large data set like this, we rely on technology to make tables 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%

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

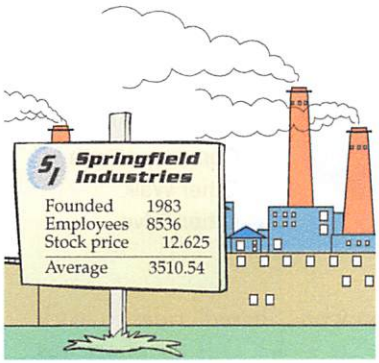
Our analysis of the GfK Roper Reports™ Worldwide survey data shows a weak pattern of concern about food safety that generally increases from youngest to oldest.

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.

2.5 Simpson's Paradox

Here's an example showing that combining percentages across very different values or groups can give confusing 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 2.7) are the results for each of their last 120 sales calls, broken down by the product they were selling.



		Product		
		Printer Paper	USB Flash Drive	Overall
Sales Rep	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 2.7 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.

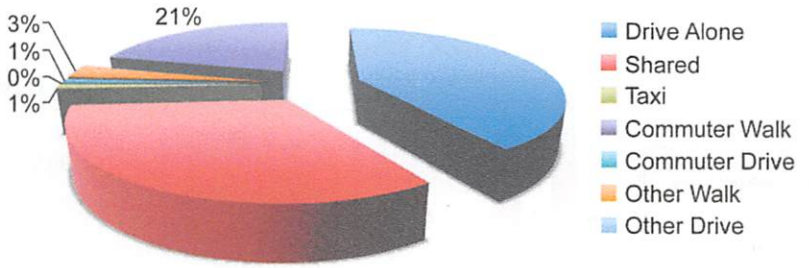
The lesson of Simpson's Paradox is to be sure to combine only comparable measurements for comparable individuals. Be especially careful when combining across different levels of a second variable. It's usually better to compare percentages *within* each level, rather than across levels.

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.

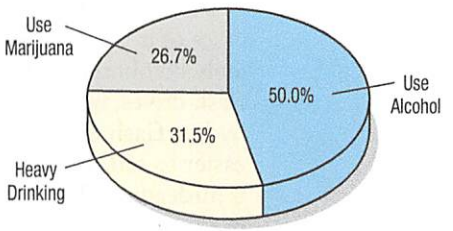
WHAT CAN GO WRONG?

- **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. Consider this pie chart of ways that respondents said they commute to work.



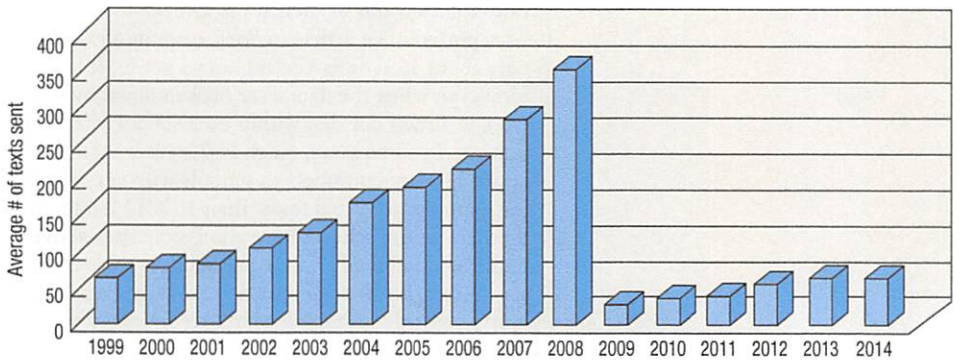
Would it surprise you to learn that the fraction who "Shared" rides to work is 33%, while the fraction who "Drive Alone" is 41%? This pie chart was made in Excel, but overuse of features that make it look interesting has hurt its ability to convey accurate information.

- **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 one. This chart shows the average number of texts sent by American cell phone customers in the period 1999 to 2014.



It may look as though text messaging decreased suddenly sometime around 2009, which probably doesn't seem right to you. In fact, this chart has several problems. First, it's not a bar chart. Bar charts display counts of categories. This bar chart is a plot of a quantitative variable (average number of texts) against year. Of course, the real problem is that starting in 2009, they reported the data for texts sent per day, not per month. Mistakes like this in graphics are more common than you think.

- **Don't confuse percentages.** Many percentages based on conditional and joint distributions sound similar, but are different (see Table 2.4):
  - The percentage of Russians who answered "Yes": This is 500/1010 or 49.5%.
  - The percentage of those who answered "Yes" who were Russian: This is 500/2175 or 23%.
  - The percentage of those who were Russian *and* answered "Yes": This is 500/5039 or 9.92%.

  In each instance, pay attention to the wording that makes a restriction to a smaller group (those who are Russian, those who answered "Yes," 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.

ETHICS IN ACTION

Mount Ashland Promotions Inc. is organizing one of its most popular events, the ZenNaturals Annual Trade Fest. At this trade show, producers, manufacturers, and distributors in the natural foods market display the latest trends in organic foods, herbal supplements, and natural body care products. The Trade Fest attracts a wide variety of participants, from large distributors who display a wide range of products to small, independent companies.

As in previous years, Nina Li and her team at Mount Ashland are in charge of managing the event, which includes all advertising and publicity as well as arranging spots for exhibitors. The success of this event depends on Nina's ability to attract large numbers of small independent retailers in the natural foods market who are looking to expand their product lines. She knows that these small retailers tend to be zealously committed to the principles of healthful lifestyle. Moreover, many are members of the Organic Trade Federation (OTF), an organization that advocates ethical consumerism.

The OTF has been known to boycott trade shows that include too many products with controversial ingredients such as ginkgo biloba, hemp, or kava kava. Nina is aware that some herbal diet teas have been receiving lots of negative attention lately in trade publications and the popular press. These teas claim to be "thermogenic" or fat burning, and typically contain ma huang (or ephedra). Ephedra is particularly controversial, not only because it can be unsafe for people with certain existing health conditions, but because this fast-acting stimulant commonly found in diet and energy products is contrary to the OTF's principles and values.

Worried that too many products at the ZenNaturals Trade Fest may be thermogenic teas, Nina decides to take a closer look at vendors already committed to participate in the event. Based on the data that her team pulled together, she finds that more than 33% of them do indeed include teas in their product lines. She was quite surprised to find that this percentage is so high. She decides to categorize the vendors into four groups: (1) those selling herbal supplements only; (2) those selling organic foods and herbal supplements; (3) those selling organic foods, herbal supplements, and natural body care products; and (4) all others. She finds that only 2% of groups 1, 2, and 4 include tea in their product lines, while 34% of the third group do. Even though group 3 contains most of the vendors, Nina instructs her team to use the average percentage 10% in its communications, especially with the OTF, about the upcoming ZenNaturals Annual Trade Fest.

- Identify the ethical dilemma in this scenario.
- What are the undesirable consequences?
- Propose an ethical solution that considers the welfare of all stakeholders.

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

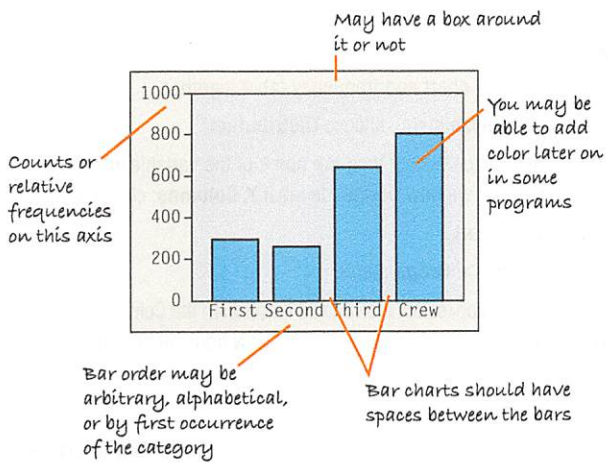
Area principle — In a statistical display, each data value is represented by the same amount of area.

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.
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.
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.
Mosaic plot	A mosaic plot is a graphical representation of a (usually two-way) contingency table. The plot is divided into rectangles so that the area of each rectangle is proportional to the number of cases in the corresponding cell.
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.
Segmented bar chart	A segmented bar chart displays the conditional distribution of a categorical variable within each category of another variable.
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

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

Excel offers a versatile and powerful tool it calls a *PivotTable*. A pivot table can summarize, organize, and present data from an Excel spreadsheet. Pivot tables can be used to create frequency distributions and contingency tables. They provide a starting point for several kinds of displays. Pivot tables are linked to data in your Excel spreadsheet so they will update when you make changes to your data. They can also be linked directly to a *PivotChart* to display the data graphically.

In a pivot table, all types of data are summarized into a row-by-column table format. Pivot table cells can hold counts, percentages, and descriptive statistics.

To create a pivot table:

- Open a data file in Excel. At least one of the variables in the dataset should be categorical.
- Choose **Insert > PivotTable** or **Data > PivotTable** (Mac). If you are using a PC, choose to put the pivot table in a new worksheet. Macintosh users should choose the option to create a custom pivot table.
- The PivotTable builder has five boxes:
  - *Field List* (top): variables from the data set linked to the *PivotTable*. (The *PivotTable* tool calls the variables "fields.") Fields can be selected using the checkbox or dragged and dropped into one of the areas below in the *PivotTable* builder.
  - *Report Filter* (middle left): Variables placed here filter the data in the pivot table. When selected, the filter variable name appears above the pivot table. Use the drop-down list to the right of the variable name to choose values to display.
  - *Row Labels* (bottom left): Values of variables placed here become row labels in the pivot table.
  - *Column Labels* (middle right): Values of variables placed here become column labels in the pivot table.
  - *Values* (bottom right): Variables placed here are summarized in the cells of the table. Change settings to display count, sum, minimum, maximum, average, and more or to display percentages and ranks.

To create a frequency distribution pivot table:

- Drag a categorical variable from the Field List into **Row Labels**.
- Choose another variable from the data set and drag it into **Values**.
- To change what fact or statistics about the **Values** variable is displayed, click the arrow next to the variable in the **Values** box and open the **Value Field Settings**. For a frequency distribution, select **count of [VARIABLE].** When changing **Value Field Settings**, note the tab **Show Values As,** which provides other display options (e.g., % of row, % of column).

The result will be a frequency table with a column for count.

To create a contingency table using *PivotTable:*

	Row Labels	Count of Cut
3	Row Labels	Count of Cut
4	Excellent	1276
5	Good	165
6	Ideal	185
7	Very Good	1064
8	Grand Total	2690

- Drag a categorical variable from the Field List into **Row Labels**.
- Drag a second categorical variable from the Field List into **Column Labels**.
- Choose another variable from the dataset and drag it into **Values**. The resulting pivot table is a row-by-column contingency table.

	Count of Cut	Column Labels							
4	Row Labels	IF	SI1	SI2	VS1	VS2	VVS1	VVS2	Grand Total
5	Excellent	92	278	196	184	220	160	146	1276
6	Good	4	39	47	33	26	4	12	165
7	Ideal	22	24	24	20	24	41	30	185
8	Very Good	26	283	263	155	190	64	83	1064
9	Grand Total	144	624	530	392	460	269	271	2690

NOTE: As with the frequency distribution, you can use the **Value Field Settings** to change the type of summary.

To create a chart from a pivot table frequency distribution or contingency table:

- Place the cursor anywhere on the pivot table.
- Click **PivotTable Tools > PivotChart**.
- Choose the type of chart: options include pie chart, bar chart, and segmented bar graph.
- Move the chart to a new worksheet by right-clicking the chart and selecting **Move chart.**
- In a bar chart created from a contingency table, by default, rows display on the x-axis and the columns are separate bars. To change this, place cursor in chart and choose **PivotChart Tools > Design > Switch Row/Column**.
- On Macs, choose the **Charts** tab and select your chart from the ribbon or choose a chart type from the **Chart** menu.

XLSTAT

To create a contingency table from unsummarized data:

- On the XLStat tab, choose **Preparing data**.
- From the menu, choose **Create a contingency table**.
- In the dialog box, enter your data range on the General tab. Your data should be in two columns, one of which is the row variable and the other is the column variable.
- On the Outputs tab, check the box next to **Contingency table** and optionally choose **Percentages/Row or Column** to see the conditional distributions.

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 > 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 Chart** 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.**

R

To make a bar chart or pie chart in **R**, you first need to create the frequency table for the desired variable:

- **table(X)** will give a frequency table for a single variable X.
- **barplot(table(X))** will give a bar chart for X.
- Similarly **pie(table(X))** will give a pie chart.

Comments

Stacked bar charts of two variables, X and Y, can be made using **barplot(xtabs(~X + Y))** or directly from a two-way table of counts or percentages. Legends and other options are available for all charts using various functions.

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.

Brief Case

Credit Card Bank

In Chapter 1, you identified the W's for the data in the file **Credit Card Bank.** For the categorical variables in the data set, create frequency tables, bar charts, and pie charts using your software. What might the bank want to know about these variables? Which of the tables and charts do you find most useful for communicating information about the bank's customers? Write a brief case report summarizing your analysis and results.

EXERCISES

SECTION 2.1

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 theaters. 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 2.2

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 2.3

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	6–9	10–14	15–21	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.

SECTION 2.4

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 segmented or 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 segmented bar chart?
e) How would a mosaic plot help to accurately display these data?

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 segmented 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 segmented bar chart?
e) How would a mosaic plot represent these data more appropriately?

CHAPTER EXERCISES

11. Graphs in the news. Find a bar graph of categorical data from a business publication (*Bloomberg Businessweek*, *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 (*Bloomberg Businessweek*, *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.

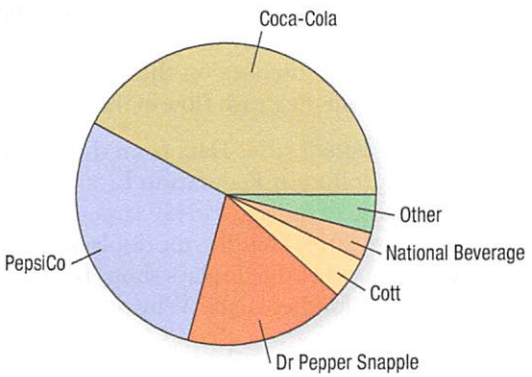
13. Tables in the news. Find a frequency table of categorical data from a business publication (*Bloomberg Businessweek*, *Fortune*, *The Wall Street Journal*, etc.).

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 (*Bloomberg Businessweek*, *Fortune*, *The Wall Street Journal*, etc.).

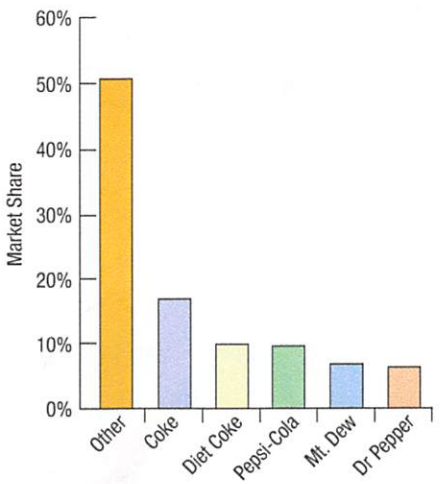
a) Is it clearly labeled?
b) Does it display percentages or counts?
c) Does the accompanying article tell the W's of the variable?
d) Do you think the article correctly interprets the data? Explain.

15. U.S. market share. An article in *The Wall Street Journal* (March 18, 2011) reported the 2010 U.S. market share of leading sellers of carbonated soft drinks, summarized in the following pie chart:



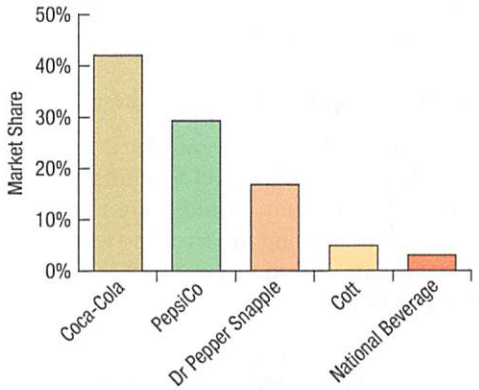
a) Is this an appropriate display for these data? Explain.
b) Which company had the largest share of the market?

16. World market share. *The Wall Street Journal* article described in Exercise 15 also indicated the market share of the leading *brands* of carbonated beverages. The following bar chart displays the values:



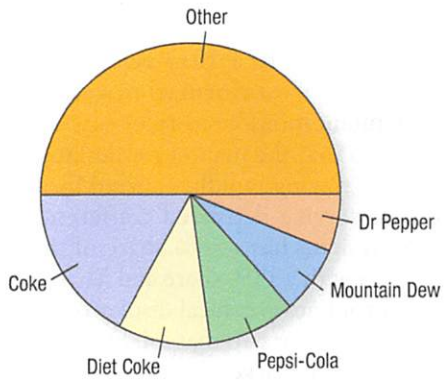
a) Is this an appropriate display for these data? Explain.
b) Which brand had the largest share of the beverage market?
c) Which brand had the larger market share—Mountain Dew or Dr Pepper?

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 somewhat 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 brands? Explain.
b) Does Mountain Dew or Dr Pepper have a bigger market share? Is that comparison easier to make with the pie chart or the bar chart of Exercise 16?

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 and Prevention (www.cdc.gov) 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. College value? In March and April of 2011, the Pew Research Center asked 2142 U.S. adults and 1055 college presidents whether they would "rate the job the higher education system is doing in providing value for the money spent by students and their families" as Excellent, Good, Only Fair, or Poor.

	Poor	Only Fair	Good	Excellent	No Answer/ Don't Know
U.S. Adults	321	900	750	107	64
College Presidents	32	222	622	179	0

a) Compare the distribution of opinions between U.S. adults and college presidents on the value of higher education.
b) Is it reasonable to conclude that 5.00% of *all* U.S. adults think that the higher education system provides an excellent value?

21. SaaS. According to a 2013 report by Synergy Research Group (www.informationweek.com/telecom/unified-communications/cisco-rules-saas-uc-conferencing-market/231601562) the market for desktop conferencing apps has been growing rapidly, spurred by a 33% jump in software-as-a-service-delivered conferencing apps revenue. Cisco Systems has a 58% share of this market, followed by Citrix with a 13% share and Microsoft with 11%. 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 2013 annual report, Mattel Inc. reported that their domestic market sales were broken down as follows: 49.6% Mattel Girls and Boys brand, 36.1% Fisher-Price brand, and the rest of their over $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.

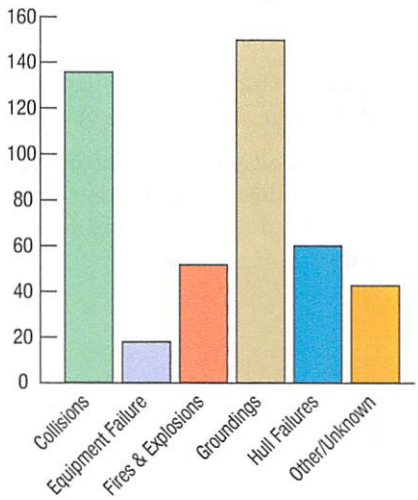
23. Small business financing. The Wells Fargo/Gallup Small Business Index asked 604 small business owners in October 2011 "how difficult or easy do you think it will be for your company to obtain credit when you need it?" 22% said "Very difficult," 21% "Somewhat difficult," 28% "About Average," 11% "Somewhat easy," and 11% "Very Easy."

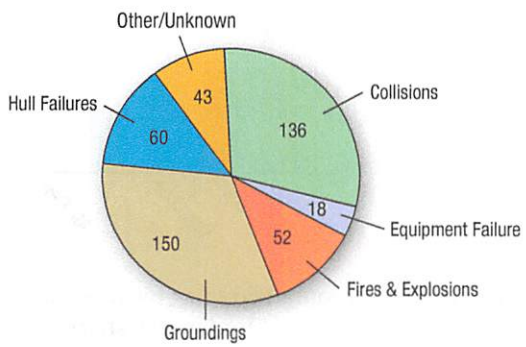
a) What do you notice about the percentages listed? How could this be?
b) Make a bar chart to display the results and label it clearly.
c) Would a pie chart be an effective way of communicating this information? Why or why not?
d) Write a couple of sentences on how small businesses felt about the difficulty of obtaining credit in late 2011.

24. Small business cash flow. The Wells Fargo/Gallup Small Business Index survey from Exercise 23 also asked 604 small businesses about their cash flow over the next 12 months. 13% responded "Very Good," 37% "Somewhat good," 21% "Neither good nor poor," 20% "Somewhat poor," and 7% "Very poor."

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 of small business owners about their cash flow in the next 12 months.

25. Environmental hazard 2014. Data from the International Tanker Owners Pollution Federation Limited (www.itopf.com) give the cause of spillage for 459 large oil tanker accidents from 1970 to 2014. 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?





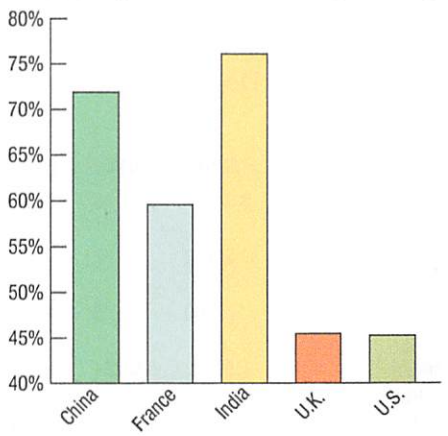
26. Olympic medals. In the history of the modern Olympics, the United States has won more medals than any other country. But the United States has a large population.

Perhaps a better measure of success is the number of medals won *per capita*—that is the number of medals divided by the population. By that measure, the leading countries are Liechtenstein (255.42 medals/cap), Norway (95.271), Finland (86.514), and Sweden (66.455). The following table summarizes the medals/capita counts for the 100 countries with the most medals.

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?

Medals/capita	# Countries	Medals/capita	# Countries
0	72	130	0
10	13	140	0
20	3	150	0
30	3	160	0
40	2	170	0
50	0	180	0
60	1	190	0
70	0	200	0
80	1	210	0
90	1	220	0
100	0	230	0
110	0	240	0
120	0	250	1

27. Importance of wealth. GfK Roper Reports Worldwide surveyed people, 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% U.K., and 45.3% U.S. There were about 1500 respondents per country. A report showed the following bar chart of these percentages.

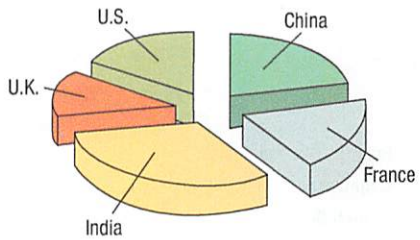


a) How much larger is the proportion of those who said acquiring wealth was important in India than in the United States?
b) Is that the impression given by the display? Explain.
c) How would you improve this display?
d) Make an appropriate display for the percentages.
e) Write a few sentences describing 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%
U.K.	27.8%
U.S.	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 from 2009 to 2013.

	2009	2010	2011	2012	2013
Google Websites	67%	66%	69%	68%	67%
Google Network Members' Websites	30%	30%	28%	27%	24%
Other Revenues	3%	4%	3%	5%	9%

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.

		Price			
		Low	Med Low	Med High	High
Size	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 priced 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 470 of the S&P 500 stocks, on how their one-day change on October 24, 2011 (a day on which the S&P 500 index gained 1.23%) compared with their year to date change.

		Year to Date	
		Positive Change	Negative Change
October 24, 2011	Positive Change	164	233
	Negative Change	48	25

a) What percent of the companies reported a positive change in their stock price over the year to date?
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 a negative change in the other period?
e) Among those companies reporting a positive change in their stock price on October 24 over the prior day what percentage also reported a positive change over the year to date?
f) Among those companies reporting a negative change in their stock price on October 24 over the prior day what percentage reported a positive change over the year to date?
g) What relationship, if any, do you see between the performance of a stock on a single day and its year-to-date 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 *Greenville, South Carolina Real Estate Hub* keeps track of home sales in their area. They reported that sales were down in 2010 by about 3.7% from the previous year. Here are the number of homes sold in Greenville for the last 5 months of 2009 and 2010:

	August	September	October	November	December
2010	475	466	502	423	495
2009	607	597	596	581	447

a) What percent of all homes in these ten months were sold in October of 2009?
b) What percent of all homes in these 10 months were sold in 2010?
c) What percent of all homes in these 10 months were sold in December?
d) How did the percent of homes sold in November change from 2009 to 2010?

34. Google financials, part 2. Google Inc. divides their total costs and expenses into five categories: Costs of Revenues, Research and Development, Sales and Marketing, General and Administrative, and Dept of Justice charges (amounts in $Millions).

	2009	2010	2011	2012	2013
Cost of Revenues	$8844	$10,417	$13,188	$20,634	$21,993
Research and Development	$2843	$3762	$5162	$6793	$7137
Sales and Marketing	$1984	$2799	$4589	$6143	$6554
General and Administrative	$1667	$1962	$2724	$3845	$4432
Dept of Justice	$0	$0	$500	$0	$0
Total Costs and Expenses	**$15,338**	**$18,940**	**$26,163**	**$37,415**	**$40,116**

a) What percent of total costs and expenses were sales and marketing in 2009? In 2013?
b) What percent of total costs and expenses were due to research and development in 2009? In 2013?
c) Have general and administrative costs grown as a percentage of total costs and expenses over this time period?

35. Movie ratings 2014. 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 the ratings of 279 movies that came out in 2014, also classified by their genre.

		Rating				
		G	PG	PG-13	R or NC-17	Total
Genre	Action/Adventure	2	14	26	17	59
	Comedy	1	6	19	53	79
	Drama	0	14	41	62	117
	Thriller/Suspense	0	0	14	44	58
	Others*	0	4	2	2	8
	Total	3	38	102	178	321

*Others include Westerns, musicals etc.

a) Find the conditional distribution (in percentages) of movie ratings for action/adventure films.
b) Find the conditional distribution (in percentages) of Genres for PG-13 rated 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. CyberShopping. It has become more common for shoppers to "comparison shop" using the Internet. Respondents to a Pew survey in 2013 who owned cell phones were asked whether they had, in the past 30 days, looked up the price of a product while they were in a store to see if they could get a better price somewhere else. Here is a table of their responses by income level.

	<$30K	$30K–$49.9K	$50K–$74.9K	>$75K
Yes	207	115	134	204
No	625	406	260	417

(Source: www.pewinternet.org/Reports/2012/In-store-mobile-commerce.aspx)

a) Find the conditional distribution (in percentages) of income distribution for those who do not compare prices on the Internet.
b) Find the conditional distribution (in percentages) of income distribution for shoppers who do compare prices.
c) Create a graph comparing the income distributions of those who compare prices with those who don't.
d) Do you see any differences between the conditional distributions? Write a brief summary of what these data show about Internet 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?

39. Top-producing movies, 2014. The following table shows the Motion Picture Association of America (MPAA; www.mpaa.org) ratings for the top 20 grossing films in the United States for each of the 10 years from 2005 to 2014. (Data are number of films.)

Year	Rating				
	G	PG	PG-13	R/NC-17	Total
2014	0	4	13	3	20
2013	1	4	11	4	20
2012	0	6	12	2	20
2011	1	4	11	4	20
2010	1	9	9	1	20
2009	0	7	12	1	20
2008	2	4	10	4	20
2007	1	5	11	3	20
2006	1	4	13	2	20
2005	1	4	13	2	20
Total	8	51	115	26	200

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 2010?
d) What percent of all top 20 films produced in 2010 or later were PG-13?
e) What percent of all top 20 films produced from 2005 to 2009 were rated PG-13 or R/NC-17?
f) Compare the conditional distributions of the ratings for films produced in 2010 or later to those produced from 2005 to 2009. Write a couple of sentences summarizing what you see.

40. Movie admissions 2013. The following table shows attendance data collected by the Motion Picture Association of America during the period 2009 to 2013. Figures are the number (in millions) of frequent moviegoers in each age group.

	Age						
	2–11	12–17	18–24	25–39	40–49	50–59	60+
2013	4.3	5.5	7.2	8.2	3.2	4.2	3.8
2012	2.8	6.3	8.7	9.9	5.8	3.3	4.6
2011	2.5	5.7	6.6	9.7	3.3	3.1	4.1
2010	3.1	6.1	7.4	7.7	3.5	3.0	4.3
2009	2.8	5.7	6.3	6.3	4.5	2.9	3.4

a) What percent of all frequent moviegoers during this period were people between the ages of 12 and 24?
b) What percent of the frequent moviegoers in 2011 were people between the ages of 12 and 39?
c) What percent of *all* frequent moviegoers during this period were people between the ages of 18 and 24 who went to the movies in 2009?
d) What percent of frequent moviegoers in 2010 were people 60 years old and older?
e) What percent of *all* frequent moviegoers in this period were people 60 years old and older who went to the movies in 2010?
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. Poverty and region 2013. In 2013, the following data were reported by the U.S. Census Bureau. The data show the number of people (in thousands) living above and below the poverty line in each of the four regions of the United States. Based on these data do you think there is an association between region and poverty? Explain.

	Below Poverty Level	Above Poverty Level
Northeast	7046	48,432
Midwest	8590	58,195
South	18,870	98,091
West	10,812	62,930

43. Being successful. In a random sample of U.S. adults surveyed in December 2011, Pew research asked how important it is "to you personally" to be successful in a high-paying career or profession. Here is a table reporting the responses. (Percentages may not add to 100% due to rounding.) (Data from www.pewsocialtrends.org/files/2012/04/Women-in-the-Workplace.pdf)

	Women		Men	
Age	18–34	35–64	18–34	35–64
One of the most important things	18%	7%	11%	9%
Very important, but not the most	48%	35%	47%	34%
Somewhat important	26%	34%	31%	37%
Not important	8%	24%	10%	20%
	100%	100%	100%	100%

a) What percent of young women consider it very important or one of the most important things for them personally to be successful?
b) How does that compare with young men?
c) From this table, can you determine what percent of all women responding felt this way? Explain.
d) Write a few sentences describing the association between the sex of young respondents and their attitudes toward the importance of financial or professional success.

44. Minimum wage workers 2013. 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 in 2013:

	Hourly Workers (in thousands)		At or Below Minimum Wage (in thousands)	
Age	Men	Women	Men	Women
16–24	7558	7552	655	1007
25–34	9281	8326	286	418
35–44	7112	7082	116	239
45–54	7181	7916	83	231
55–64	4915	5798	55	99

a) What percent of all women were in the 16–24 *Age* group?
b) Using side-by-side bar graphs, compare the proportions of men and women who worked at or below minimum wage at each *Age* group. Use the total number of workers at or below minimum wage as the denominator. 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, Caucasian, and Other for the year 2010. Also included are the numbers for the general U.S. population and the number of tickets sold.

	Caucasian	Hispanic	African-American	Other	Total
Population	204.6	49.6	37.2	18.6	310
Moviegoers	88.8	26.8	16.9	8.5	141
Tickets	728	338	143	91	1300
Total	1021.4	414.4	197.1	118.1	1751

a) Compare the conditional distribution of *Ethnicity* for all three groups: the entire population, moviegoers, and ticket holders.
b) Write a brief description of the association between population groups and *Ethnicity*.

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. Success II. Look back at the table in exercise 43 concerning desires for success and a high-paying career. That table presented only the percentages, but Pew Research reported the numbers of respondents in the major categories:

	Women		Men	
Age	18–34	35–64	18–34	35–64
Count	610	571	703	605

With this additional information you should be able to answer these questions. (Note: Percentages were rounded to whole numbers, so estimated cell counts will have fractions. You need not round estimated cell counts to whole numbers for the purpose of answering these questions.)

a) What percentage of 18- to 34-year-olds (both male and female) reported that being successful in a high-paying career or profession was "one of the most important things" to them personally?
b) What percentage of 18- to 34-year-olds who said that such success was "one of the most important things" were women?
c) Write a few sentences describing how the opinions of young women differ from those of older female respondents.

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 (www.allcountries.org/uscensus/424_household_pet_ownership_and_by_selected.html) reports the number of households owning various types of pets. Specifically, they keep track of dogs, cats, birds, and horses.

a) Do you think the income distributions of the households who own these different animals would be roughly the same? Why or why not?

Percent Distribution of Households Owning Pets

Income Range	Pets: Dog	Cat	Bird	Horse
Under $12,500	12.7	13.9	17.3	9.5
$12,500 to $24,999	19.1	19.7	20.9	20.3
$25,000 to $39,999	21.6	21.5	22.0	21.8
$40,000 to $59,999	21.5	21.2	17.5	23.1
$60,000 and over	25.2	23.7	22.3	25.4

b) The table shows the percentages of income levels for each type of animal owned. Are these row percentages, column percentages, or total percentages?
c) Do the data support that the pet food company should not target specific market segments based on household income? Explain.

49. 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.

50. 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?

51. 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?

52. 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.)