Displaying and Describing Categorical Data

CHAPTER 4

CONNECTIONS: CHAPTER

In Chapter 2 we learned that data can be categorical or quantitative. In this chapter we concentrate on categorical data. We learn how to display them with graphs and describe them with numerical summaries, first examining one variable at a time. Then we learn how to display and assess the relationship between two categorical variables.

LEARNING OBJECTIVES

1. Choose an appropriate display of categorical data and determine its effectiveness
2. Analyze a contingency table of counts or percentages
3. Create and analyze relative frequency distributions from tabulated data
4. Compute and interpret marginal and conditional distributions from contingency tables
5. Identify misleading results that are due to data aggregation (Simpson's paradox)

MEC: Mountain Equipment Co-op

Mountain Equipment Co-op, or MEC, is a leader in active outdoor lifestyle equipment: gear, clothing, and services. A Canadian success story founded over 40 years ago by a group of climbers at the University of British Columbia, it has grown to more than $250 million in annual sales with 3.5 million members, while retaining its co-operative member-owned structure.

It has 16 retail locations across Canada, with a global supply chain hub in Surrey, British Columbia. MEC has embraced internet sales and marketing. In 2001 the MEC website became transactional so that members could buy clothing and gear online. It is now also bilingual English/French. In MEC's words, their aim is to provide quality gear and excellent value, and to minimize environmental impact by building products that last.

MEC is a leader in ethical sourcing, sustainability initiatives, and charitable contributions to the environmental sector. The Community Contributions page on their website details an extensive program of grants and product donations, national and regional partnerships, and outreach and advocacy programs. It is a member of 1% For the Planet, investing one percent of annual revenue to environmental causes.

MEC is a fast-growing company. Approximately 10% of adult Canadians are MEC members, and the number increased by an average of over 10 000 new members every month, as at the end of 2013. MEC employs about 1500 people and was recognized as one of Canada's Top 100 Employers in 2011.

Based on information from www.mec.ca

WHO	Visits to the MEC.ca website
WHAT	Originating province of search on MEC's website
WHEN	Jan. 1 – Dec. 31, 2012
WHERE	Canada-wide
HOW	Data compiled via Google Analytics from MEC website
WHY	To understand regional differences in where customers come from

There is a well-known saying that the three most important principles of real estate are: location, location, location. And in French cooking, the three most important principles are: use butter, use butter, use butter. A simple three-fold rule also applies to data analysis.

MEC, like most companies, collects data on visits to its website. Actual data are proprietary, and companies either need to invest in their own resources to handle the large data files, or rely on third party resources such as Google Analytics to summarize the data. Without formal access to a company's data, a researcher can turn to online resources such as Google Trends (www.google.com/trends) to analyze search volume for "Mountain Equipment Co-op," which is a useful proxy for total traffic. As well, Google AdWords (https://adwords.google.com) gives actual measures of the number of times a particular item was searched for, or can identify the most common keywords that brought a visitor to the site. In this illustration we have actual data courtesy of MEC.

Raw data are rarely informative. And rarely can we see what is going on, but 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 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 (as in Figure 4.1) will reveal things you are not likely to see in a table of numbers and will help you to *plan* your approach to

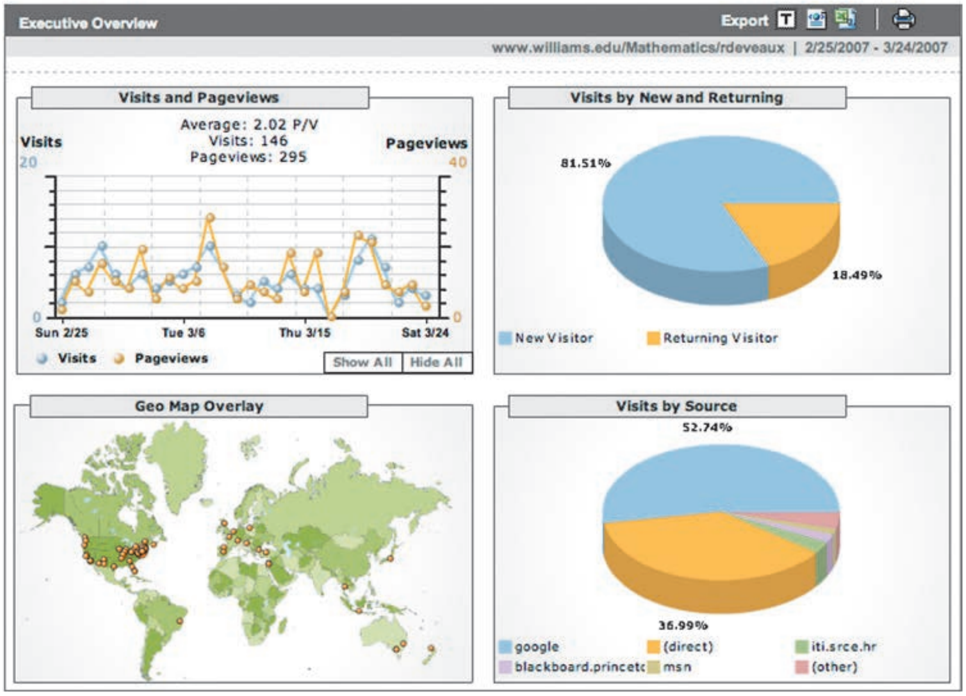


Figure 4.1 Part of the output from Google Analytics (www.google.com) for the period Feb. 25 to March 24, 2007 displaying website traffic.

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 and effectively in this chapter.

4.2 Frequency Tables

To make an informed business decision, it is often important to know how a variable distributes it values. For example, as part of planning new retail locations and spending advertising dollars, MEC managers might want to know how much activity their website attracts from different provinces in the country. Since *Province* is a categorical variable, the possible values of the variables are just the categories, namely the provinces, and we can start by counting the number of cases in each category.

Table 4.1 has a summary of information of the originating province of search traffic to MEC.ca, created using Google Analytics. Only provinces that have bricks-and-mortar MEC retail stores are included.

Province	Organic Search Visits
British Columbia	1609 160
Alberta	1031 830
Manitoba	208 185
Ontario	2108 643
Quebec	1441 269
Nova Scotia	138 393
Total	**6537 470**

Table 4.1 Frequency table of organic search traffic to MEC.ca, Jan. 1 – Dec. 31, 2012, by province. An organic search visit originates from a search engine, not from an advertisement.

Source: MEC and Google Analytics, Feb. 2013

Organizing the counts in this way (i.e., as in Table 4.1) is called a **frequency table**; it shows the number of visits (cases) for each category and records totals and category names. The names of the categories label each row in the frequency table. For *Province* these are "British Columbia," "Alberta," 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. 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 a dozen categories. In the MEC case, we could include another category that includes searches originating in all the other provinces.

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 tables 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.

Province	Organic Search Visits (%)
British Columbia	24.61%
Alberta	15.78%
Manitoba	3.18%
Ontario	32.25%
Quebec	22.05%
Nova Scotia	2.12%
Total	**100.00%**

Table 4.2 A relative frequency table for the same data.

Most often a frequency table will contain both the counts and the percentages, to give views of the data in both absolute and relative ways. But be careful about using percentages when the total count is small. Suppose you read that 67% of students in a particular class got grades of A. You would be very impressed (and surprised) if it happened in a large class of 100 students. But if the class had only three students, 67% would just mean that two of three students got an A. You wouldn't be so impressed.

99.9%?

If you are careful to add the percentages in Table 4.2, you will notice the total is 99.99%. 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.

4.3 Charts

The Area Principle

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

The figure does not accurately represent the information in the table. What's gone wrong? The height of the images in the figure do match the percentages 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 image, and it's that aspect of the image that we notice.

Since there were nearly three times as many visits from Ontario and BC as from Quebec, the images depicting the number from Ontario and from BC is almost three times higher than the image for Quebec, but it occupies almost nine times the area, since both the height and the width were increased threefold to keep the image looking proportional. As you can see from the frequency table, that 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 shoes, 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 about three times as many visits came from BC and Ontario as from Quebec—not the impression that the images in Figure 4.2. conveyed. We can also see that Manitoba and Nova

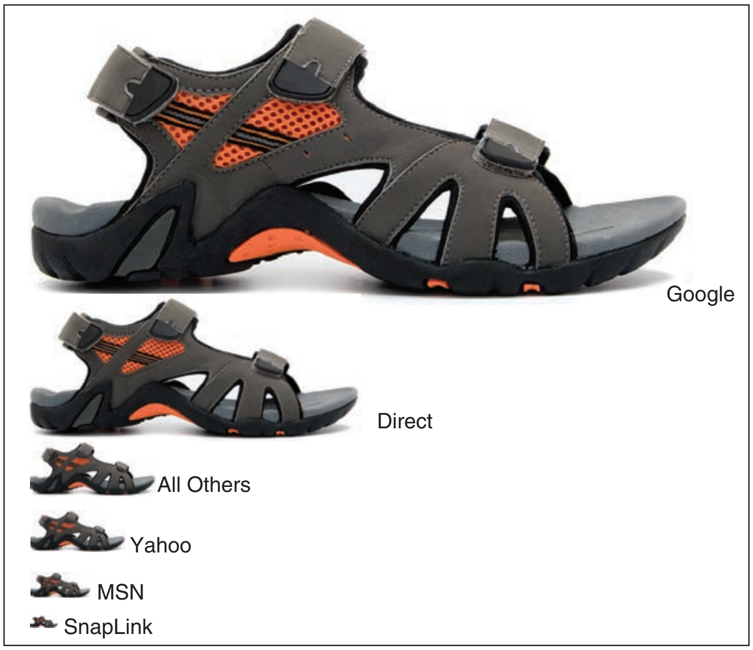


Figure 4.2 Although the length of each shoe corresponds to the correct number, the impression we get is all wrong because we perceive the entire area of the shoe. In fact, only a little more than 50% of all website visits originated in BC or Ontario.

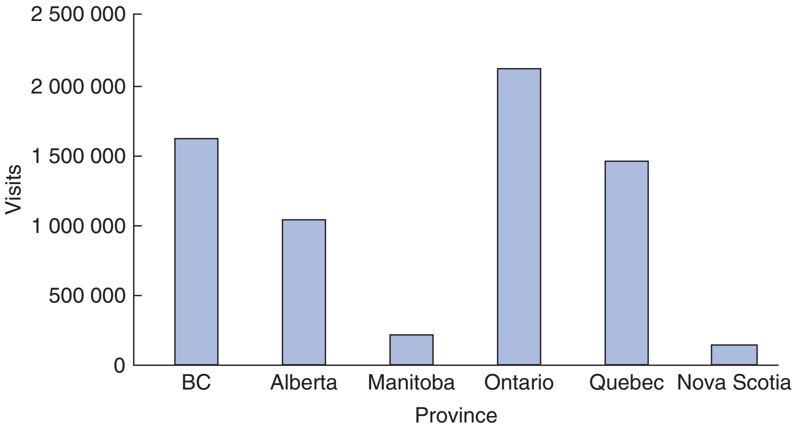
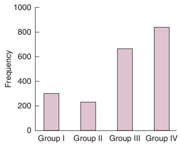


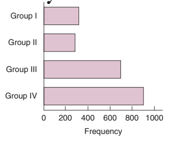
Figure 4.3 Visits to MEC website by *Province*. With the area principle satisfied, the true distribution is clear.

Scotia had about half as many as Quebec. 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.

Bar charts are usually drawn vertically in columns, 

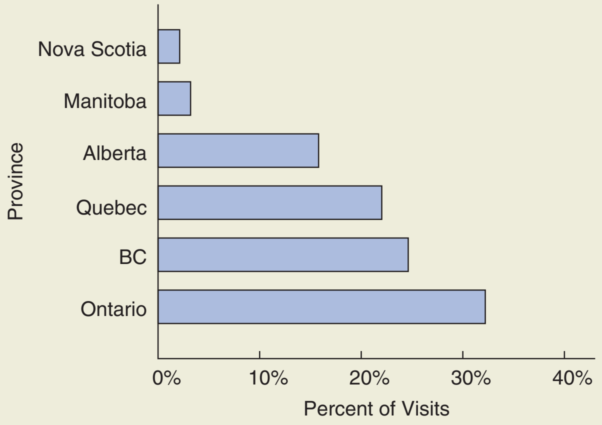
but sometimes they are drawn with horizontal bars, like this.[1]



Horizontal bars are very useful when the category labels are quite long, as they are in Figure 4.3. If the spaces between bars in Figure 4.3 were reduced, the labels would either be in much smaller print or printed on an angle, making them harder to read.

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

It is often also very useful to rearrange the order of the categories, and the bars, so that they go from highest to lowest (or lowest to highest). This version is referred to as a Pareto chart, but it is just a variation of the bar chart we have discussed. Here is another look at Figure 4.3, with reordered, and horizontal, bars.



Pie Charts

Unfortunately, another display of categorical data is still in wide use. **Pie charts** were designed to show how a whole group breaks into several categories. The whole group of cases is represented as a circle, and the circle in

[1] Excel refers to this display as a bar graph.

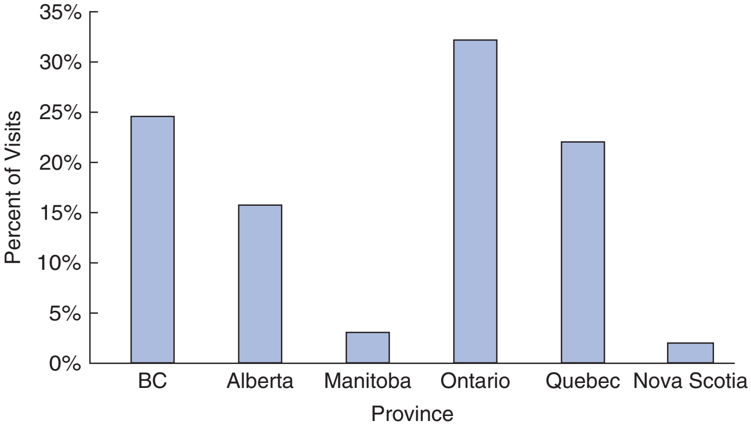


Figure 4.4 The relative frequency bar chart looks the same as the bar chart (Figure 4.3) but shows the proportion of visits in each category rather than the counts.

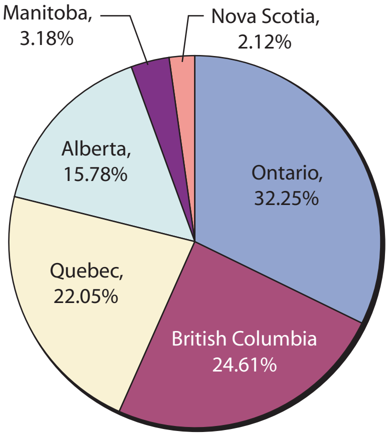


Figure 4.5 Relative frequency of visits by *Province*.

sliced into pieces whose size is proportional to the fraction of the whole in each category.

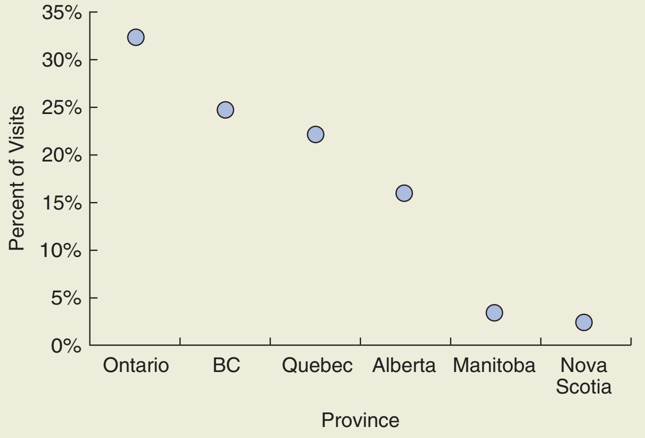
Becker and Cleveland (1996) wrote, "Pie charts have severe perceptual problems." Tufte (1983) prefers using a frequency table to a ". . . dumb pie chart; the only worse design than a pie chart is several of them, for then the viewer is asked to compare quantities located in spatial disarray both within and between pies. . . .Given their low data-density and failure to order numbers along a visual dimension, pie charts should never be used." That's a very strong indictment of pie charts.

Pie charts may give a quick impression of how a whole group is partitioned into smaller groups, but mostly for seeing relative frequencies near 1/2, 1/4, or 1/8. That's because we're used to cutting up pies into 2, 4, or 8 pieces. However, in Figure 4.5 it is very difficult to compare Alberta to Ontario or British Columbia, or Quebec to Manitoba or Nova Scotia. What's worse is that since the areas are hard to interpret, the relative frequencies are included. And if you have the relative frequencies, why is the pie needed? Compare the pie chart to the bar chart in Figure 4.4 and to the dot plot above.

The worst examples of pie charts occur when the variable is binary so that the pie has only two pieces. For example, a pie chart of the male/female split in a population is a graph that displays one piece of information, namely the percentage of males. You might think there are two pieces, but if you know the percentage of males, you also know the percentage of females. They have to sum to 100%! And when multiple two-slice pies are shown, for example, to show the male/female split across a number of countries, it is clear that the graphic designer did not think about communicating information clearly.

Tufte makes the challenge that it is *always* possible to find a better way to display data than by pie charts. We challenge you to take up his challenge!

There are a number of alternatives to bar charts. A simple and effective one is a dot plot, where dots replace the bars. After all, it is only the height (or length) of the bars that matters, not the bars themselves. Here are the data in Table 4.2 displayed as a *dot plot*. Note that the dots should not be joined up with line segments since the data are categorical.



Statisticians and psychologists have studied our ability to decode quantitative information. We are best at finding positions on a common scale (e.g., dot plot or bar chart). Next best are our abilities if the scales are identical but not aligned (e.g., comparing two side-by-side but separate dot plots or bar charts). We perceive length more accurately than angles or area, which is why pie charts are hard to interpret. We are worst at perceiving volume and colour. So three-dimensional charts of single-variable data should never be used.

Beware of chartjunk, the term coined by Tufte to describe decorations in graphics that generate a lot of ink but do not tell the viewer anything new. It is possible to be both artistically interesting and accurate in communicating information, but it takes work.

For additional advice on good and bad graphs, see What Can Go Wrong near the end of this chapter.

◆ **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, always check the **Categorical Data Condition:** that the data are counts or percentages of individuals in categories.

If you want to make a relative frequency bar chart or insist on making a pie 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 one of these charts, since the percentages won't add up to 100%. For the MEC search 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. Our Categorical Data Condition is just the first of many such checks.

4.4 Contingency Tables

GfK Roper Consulting gathers 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 people in different countries differ in their attitudes toward the food they eat. One question in the Roper survey asked respondents how they felt about 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.

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

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?

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

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 five 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."

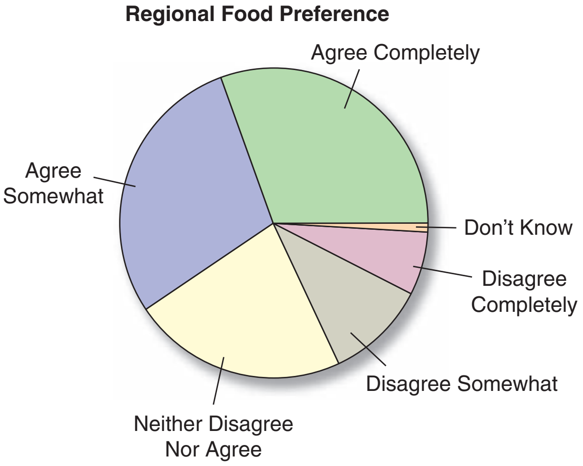
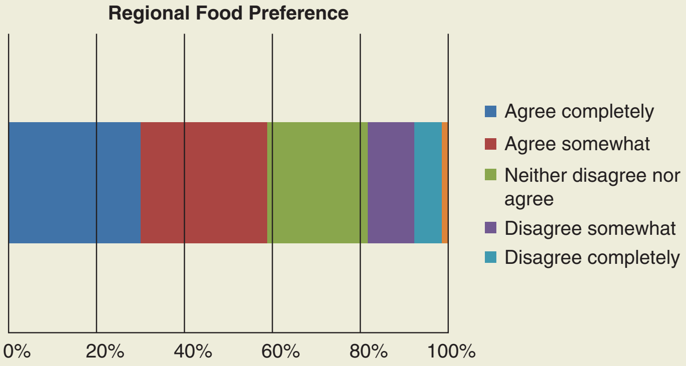


Figure 4.6 It's clear from the pie chart that the majority of respondents identify with their local foods.

A segmented or stacked 100% bar chart is another alternative to a pie chart. Instead of displaying each bar separately, the bars are "stacked" into one bar representing the 100% total. In other words, a 100% circle has become a 100% rectangle. It is even clearer from this chart that the majority of respondents identify with their local foods. In fact, this chart shows that almost 60% agree; that figure is much harder to ascertain from the pie chart.



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**.

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**.

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. If you look across the row in Table 4.4 for the

		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.

United Kingdom, you can see that 504 people 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 people from India or China? 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 Table	*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 Table	*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 Table	*12.48*	*3.67*	*1.68*	*0.845*	*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 Table	*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 Table	*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 Table	***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 first pull out the values corresponding to the percentages of the total.

Regional Preference—Percentage of Total

		Agree Completely	Agree Somewhat	Neither Disagree Nor Agree	Disagree Somewhat	Disagree Completely	Don't Know	Total
Country	**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?

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 5 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 5 0.26%.

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.

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*.

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.

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

In effect, we can temporarily restrict the *Who* first to Indians and look at how their response 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. 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 4.7 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 shows the same comparison of India and the United Kingdom, but this time using pie charts. We'll leave it to you decide which graphical comparison is easier to interpret.

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.

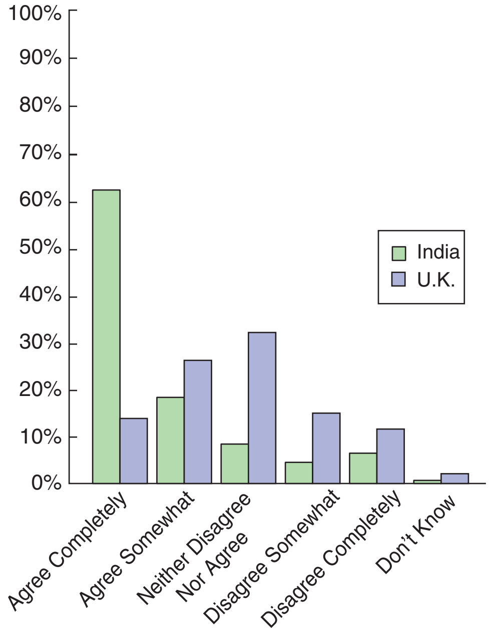


Figure 4.7 Side-by-side bar charts of the conditional distributions of *Regional Food Preference* importance for India and the United Kingdom. The percentage of people who agree is much higher in India than in the United Kingdom.

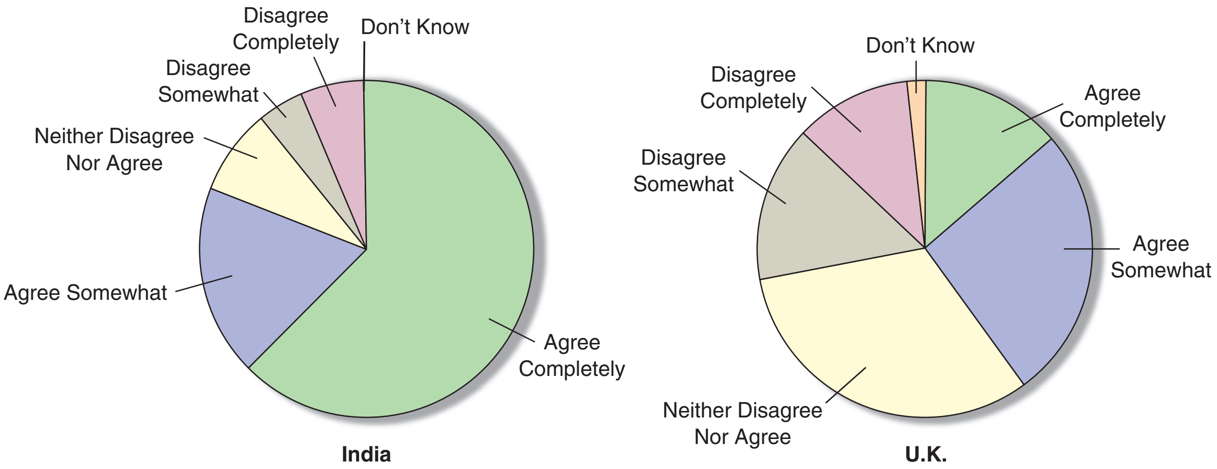


Figure 4.8 Pie charts of the conditional distributions of Regional Food Preference importance for India and the United Kingdom. It's much harder to compare percentages within each country using pie charts than with side-by-side bar charts.

From Figure 4.7, 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, we say that the 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.

JUST CHECKING

So that they can balance their inventory, an optometric practice collects the following data about its patients.

Sex		Eye Condition			
		Near Sighted	Far Sighted	Need Bifocals	Total
	Males	6	20	6	**32**
	Females	4	16	12	**32**
	Total	**10**	**36**	**18**	**64**

1 What percent of females are far-sighted?
2 What percent of near-sighted customers are female?
3 What percent of all customers are far-sighted 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 near-sighted customers to the percent of all customers who are female.
7 Does it seem that *Eye Condition* and *Sex* might be dependent? Explain.

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

Segmented Bar Charts

We could display the Roper survey information on India and the United Kingdom comparatively, with two segmented bar charts. Instead of dividing up circles as we did when making pie charts, we divide up bars. The resulting **segmented bar chart** (Figure 4.9) 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*.

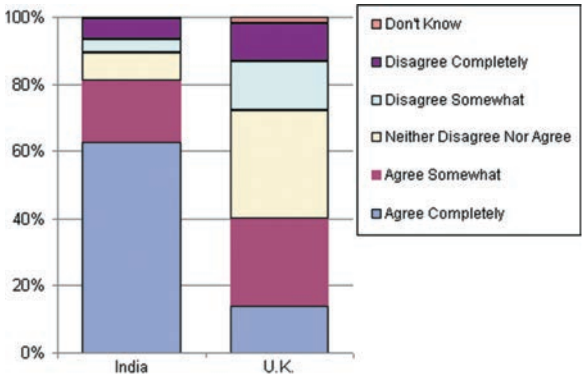


Figure 4.9 Although the totals for India and the United Kingdom are different, the bars are the same height because we have converted the numbers to percentages. Compare this display with the side-by-side bar charts in Figure 4.7 and the side-by-side pie charts of the same data in Figure 4.8.

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 how they felt about 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

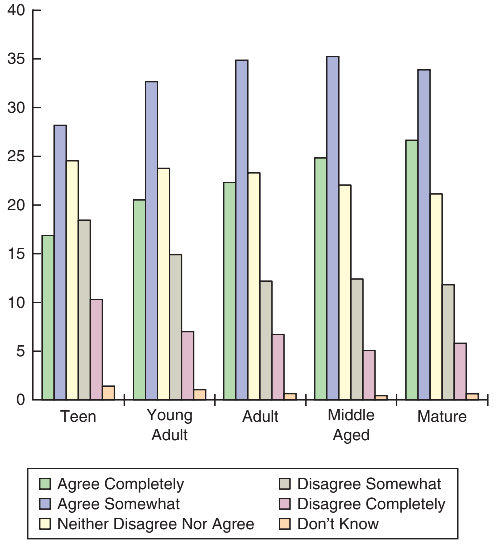
Mechanics For a large data set like this, we rely on technology to make table and displays. Because we want to compare the response distribution by age, we will examine the row percentages for each age group.

Food Safety

Age	Agree Completely	Agree Somewhat	Neither Disagree Nor Agree	Disagree Somewhat	Disagree Completely	Don't Know	Total
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

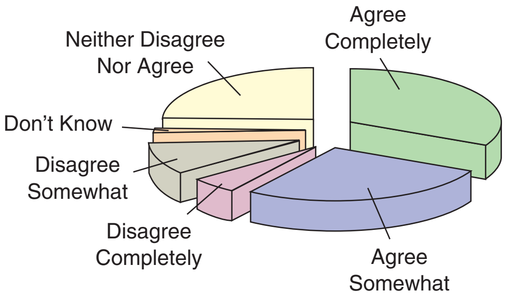
Our analysis of the GfK Roper Reports™ Worldwide survey data for 2006 shows a pattern of concern about food safety that generally increases from youngest to oldest.

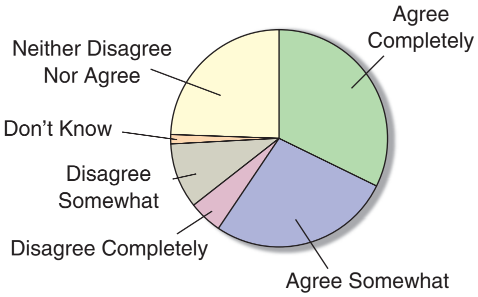
Our analysis thus far has not considered whether this trend is consistent across countries. If it were of interest to your group, we could perform a similar analysis for each of the countries.

The enclosed tables and plots provide support for these conclusions.

WHAT CAN GO WRONG?

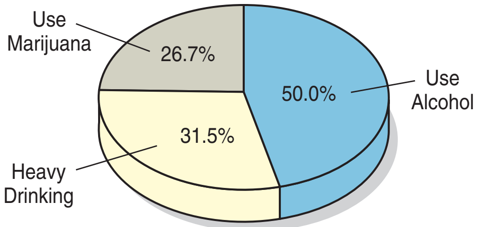
- **Don't violate the area principle.** This is probably the most common mistake in a graphical display. Violations of the area principle are often made for the sake of artistic presentation. Earlier we quoted Tufte, who said that "the only worse design than a pie chart is several of them..." Here is another take. The only worse design than a 2-D pie chart is a 3-D pie chart. If you insist on using a pie chart, don't compound the visual perception difficulty by adding irrelevant and misleading perspective. While we're on the subject, don't use 3-D bar charts either (see below). Here, for example, are two versions of the same pie chart for the *Regional Preference* data.





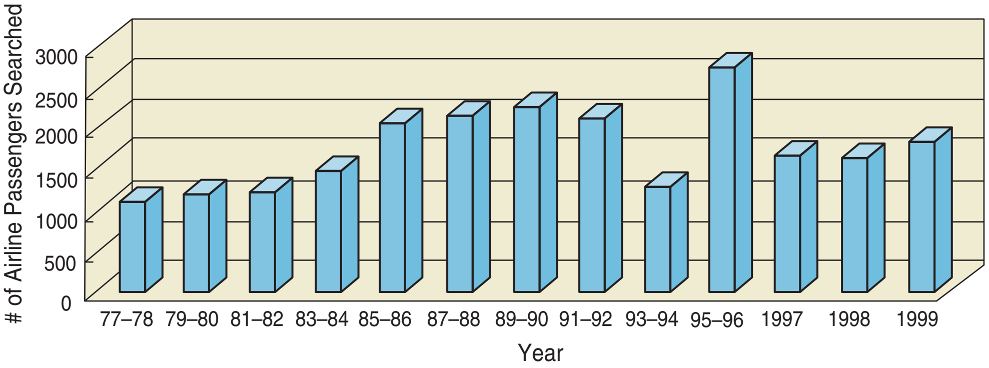
The one on the left looks interesting, doesn't it? But showing the pie three dimensionally on a slant violates the area principle and makes it much more difficult to compare fractions of the whole made up of each category of the response—the principal feature that a pie chart ought to show. Those sections of the pie in the forefront are emphasized. If you want to mislead the viewer, put the "bad news" category in the back and the "good news" category in the front. The viewer sees not only the top of the pie but the side as well.

- **Keep it honest.** Here's a pie chart that displays data on the percentage of high school students who engage in specified dangerous behaviours. What's wrong with this plot?



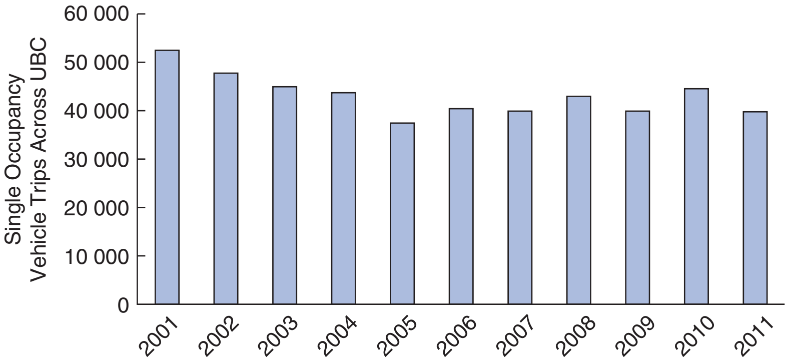
Try adding up the percentages. Or look at the 50% slice. Does it look right? Then think: What are these percentages of? Is there a "whole" that has been sliced up? In a pie chart, the proportions shown by each slice of the pie must add up to 100%, and each individual must fall into only one category.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

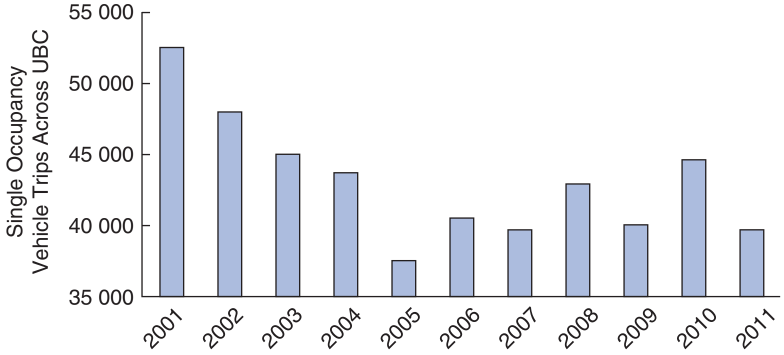


Looks like things didn't change much in the final years of the twentieth century—until you read the bar labels and see that the last three bars represent single years, while all the others are for *pairs* of years. The false depth makes it even harder to see the problem.

Here's yet another example. If the vertical axis is truncated, a slight change across bars can become a pronounced one. The following two bar charts show the change in single occupancy vehicle (SOV) trips at the University of British Columbia over the years. The first chart shows a modest decline, the second shows a strong decline, and it makes 2005 look like the Year of the Carpool!



(continued)



- **Don't confuse percentages.** Many percentages based on a conditional and joint distributions sound similar, but are different:
    - The percentage of French who answered "Agree Completely": This is 347/1539 or 22.5%.
    - The percentage of those who answered "Don't Know" who are French: This is 15/80 or 18.75%.
    - The percentage of those who were French *and* answered "Agree Completely": This is 347/7690 or 4.5%.

Regional Preference

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

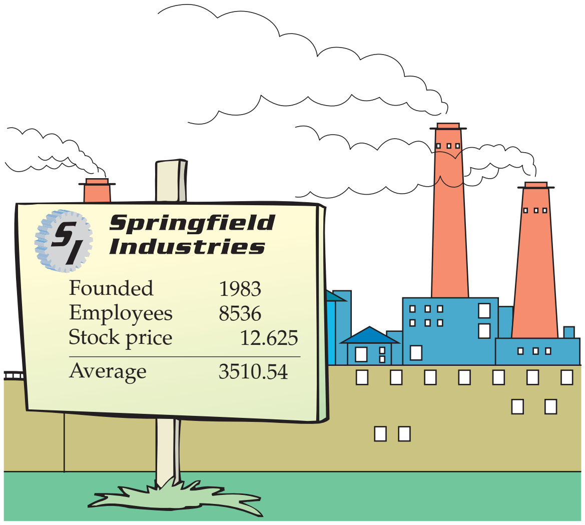
In each instance, pay attention to the wording that makes a restriction to a smaller group (those who are French, those who answered "Don't Know," and all respondents, respectively) before a percentage is found. This restricts the *Who* of the problem and the associated denominator for the percentage. 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	10 out of 20	100 out of 120
	90%	50%	83%
Katrina	19 out of 20	75 out of 100	94 out of 120
	95%	75%	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

(*continued*)

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 the sciences. 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.

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.

Simpson's Paradox shows the perils of aggregation, which can happen in many different contexts. A political party could win a majority of seats, but lose the popular vote if in the ridings it wins, it wins by a small number of votes, and in the ridings it loses, it loses by a lot. In the 1960 World Series, the New York Yankees lost to the Pittsburgh Pirates even though the Yankees outscored the Pirates 55 to 27 in the seven-game series. New York's victories were by scores of: 16–3, 10–0, and 12–0, while Pittsburgh's victories were by scores of: 6–4, 3–2, 5–2, and 10–9. Best in the aggregate didn't mean best in wins and losses.

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 favourable 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?

We've learned that we can summarize categorical data by counting the number of cases in each category, sometimes expressing the resulting distribution as percents. We can display the distribution in a bar chart or, if you insist, a pie chart. 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.

- 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.

- We can display these conditional and marginal distributions using bar charts or related graphs.
- If the conditional distributions of one variable are (roughly) the same for every category of the other, the variables are independent.

Terms

Area principle A principle that helps to interpret statistical information without distortion by insisting that in a statistical display, each data value be 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.

Categorical data condition Data are counts or percentages of individuals in categories.

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

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

Contingency table A contingency table displays counts and, sometimes, percentages of individuals falling into named categories on two or more variables. The table categorizes the individuals on all variables at once, to reveal possible patterns in one variable that may be contingent on the category of the other.

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 A table that lists the categories in a categorical variable and gives the number (i.e., count) 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.

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. We recommend not using them. Use other alternatives wherever possible.

Relative frequency table A frequency table showing proportions (i.e., relative frequencies) or percentages instead of numbers or counts. But be careful when using percentages. Always consider the size of the base; that is, the denominator being used to compute the percentages.

Row percent The proportion of each row contained in the cell of a contingency table.

Segmented bar chart A bar representing the "whole" divided proportionally into segments corresponding to the percentage in each group.

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 contingency table.

Skills

PLAN

- Recognize when a variable is categorical and choose an appropriate display for it.
- Understand how to examine the association between categorical variables by comparing conditional and marginal percentages.

DO

- Summarize the distribution of a categorical variable with a frequency table.
- Display the distribution of a categorical variable with a bar chart.
- Construct and examine a contingency table.
- Construct and examine displays of the conditional distributions of one variable for two or more groups.

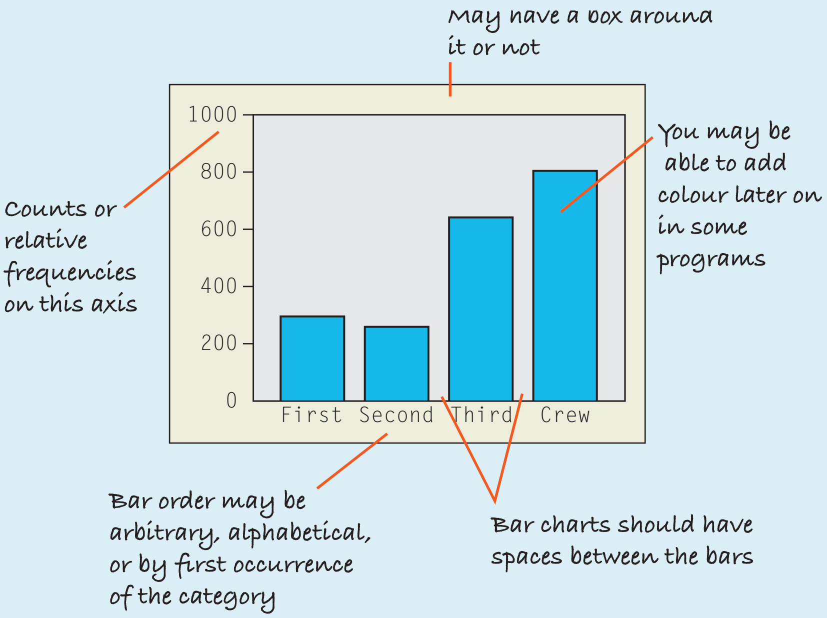
REPORT

- Describe the distribution of a categorical variable in terms of its possible values and relative frequencies.
- Describe any anomalies or extraordinary features revealed by the display of a variable.
- Describe and discuss patterns found in a contingency table and associated displays of conditional distributions.

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

First make a pivot table (Excel's name for a frequency table). From the **Data** menu, choose **Pivot Table** and **Pivot Chart Report.**

When you reach the Layout window, drag your variable to the row area and drag your variable again to the data area. This tells Excel to count the occurrences of each category.

Once you have an Excel pivot table, you can construct bar charts and pie charts.

Click inside the Pivot Table.

Click the Pivot Table Chart Wizard button. Excel creates a bar chart.

A longer path leads to a pie chart; see your Excel documentation.

Comments

Excel uses the pivot table to specify the category names and find counts within each category. If you already have that information, you can proceed directly to the Chart Wizard.

EXCEL 2007

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.

MINI CASE STUDY PROJECT

Eddie's Hang-Up Display

Chances are very high that when you walk into a retail store you notice the merchandise. But do you notice the store fixtures: hangers and size dividers, clothing racks, display cases, signs, tagging and price labels, mannequins, and the myriad of supplies that retailers need to run a business? Eddie's Hang-Up Display Ltd. (www.eddies.com) is one of Canada's leading importers and distributors of store fixtures and retail supplies. They have suppliers in Taiwan, China, Korea, Thailand, Italy, Turkey, France, and the United States. Eddie's has stores in Vancouver and Edmonton and offers over 3000 different display and supply items in addition to custom manufacturing.

Like MEC in the chapter's opening illustration, Eddie's relies on Google Analytics to analyze web traffic and a variety of other data. The Excel spreadsheet **ch04_MCPS_Eddies** has data on *Visits*, *Pages*, and *New Visits* for each of ten regions for March and October 2012. These are two peak months in their business as retailers prepare for spring and Christmas sales periods. The spreadsheet also has 2012 monthly data on these variables for British Columbia and Alberta, where Eddie's has retail stores.

Using Excel or your statistics package, create frequency tables and bar charts of the three variables by region, for each of the two months separately. Next, create a bar chart that compares the two months on the same graph. Then create frequency tables and bar charts to compare data from British Columbia and Alberta across the year. Write a case report summarizing your analysis and results.

EXERCISES

1. Graphs in the news. Find a bar graph of categorical data from a business publication (e.g., *Financial Post*, *The Economist*, *The Wall Street Journal*, etc.). LO❶

a) Is the graph clearly labelled?
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.

2. Graphs in the news, part 2. Find a pie chart of categorical data from a business publication (e.g., *Financial Post*, *The Economist*, *The Wall Street Journal*, etc.). LO❶

a) Is the graph clearly labelled?
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.

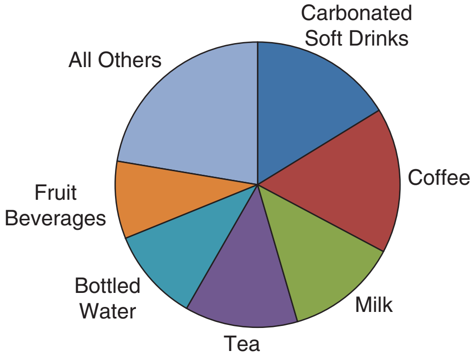
3. Tables in the news. Find a frequency table of categorical data from a business publication (e.g., *Financial Post*, *The Economist*, *The Wall Street Journal*, etc.). LO❶

a) Is it clearly labelled?
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.

4. Tables in the news, part 2. Find a contingency table of categorical data from a business publication (e.g., *Financial Post*, *The Economist*, *The Wall Street Journal*, etc.). LO❶

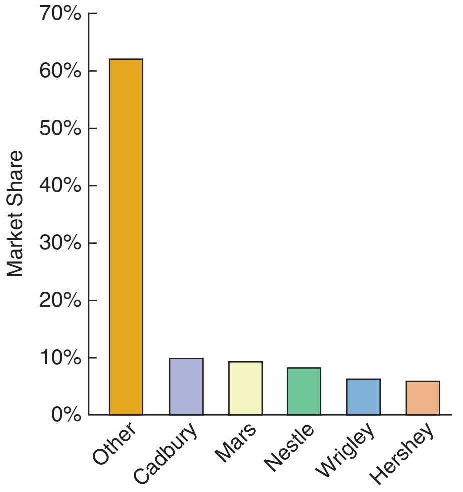
a) Is it clearly labelled?
b) Does it display percentages or counts?
c) Does the accompanying article tell the W's of the variable?
d) Do you think the article correctly interprets the data? Explain.

5. Canadian market share. A report on the Canadian Soft Drink Industry, prepared by Agriculture and Agri-Food Canada (AAFC), summarized Canada's non-alcoholic beverage market in 2009. Here is a pie chart with the results. LO❶



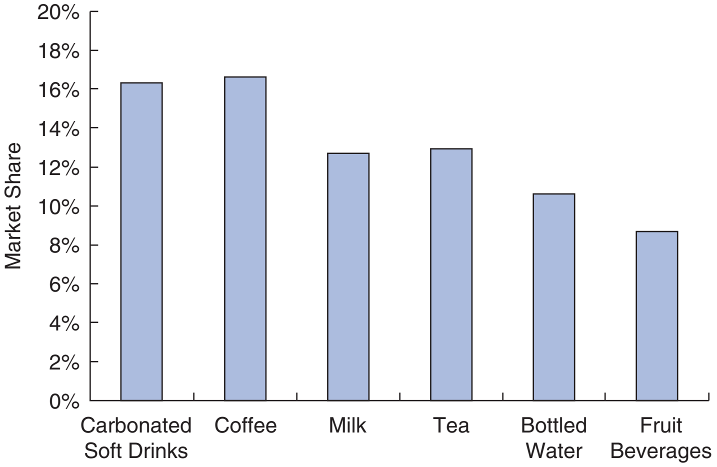
a) Is this an appropriate display for these data? Explain.
b) Compare the relative market share of carbonated soft drinks with that of coffee, tea, milk, and bottled water.
c) Approximately what percentage is in the "All Others" category?

6. World market share. An article that appeared in 2005 *The Wall Street Journal* indicated the world market share for leading distributors of total confectionery products. The following bar chart displays the values: LO❶



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

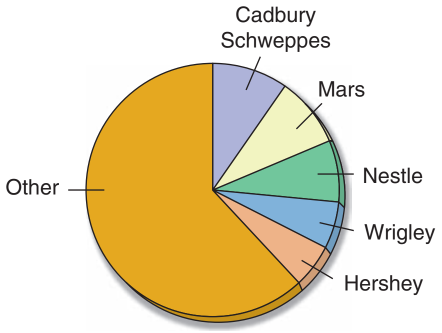
7. Canadian market share again. Here's a bar chart of the data in Exercise 5. LO❶



a) Compared to the pie chart in Exercise 5, which is better for displaying the relative portions of market share? Explain.

b) What is missing from this display that might make it misleading?

8. World market share again. Here's a pie chart of the data in Exercise 6. LO❶



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?

9. 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. Statistics Canada reported the leading causes of death in Canada in 2009 as follows. LO❶

Cause of Death	Percent
Cancer	29.8
Heart disease	20.7
Circulatory diseases and stroke	5.9
Respiratory diseases	4.6
Accidents	4.3

a) Is it reasonable to conclude that heart or respiratory diseases were the cause of approximately 25% of Canadian deaths in 2009?

b) What percent of deaths were from causes not listed here?

c) Create an appropriate display for these data.

10. 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. LO❶

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.

11. 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 venders in the area of Web conferencing in 2006 were 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. LO❶

12. Mattel. In their 2011 annual report, Mattel Inc. reported that their worldwide market gross sales were broken down as follows: 60.7% Mattel Girls and Boys brand, 31.6% Fisher-Price brand and the rest of the over $6.8 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. LO❶

13. 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. LO❶

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.

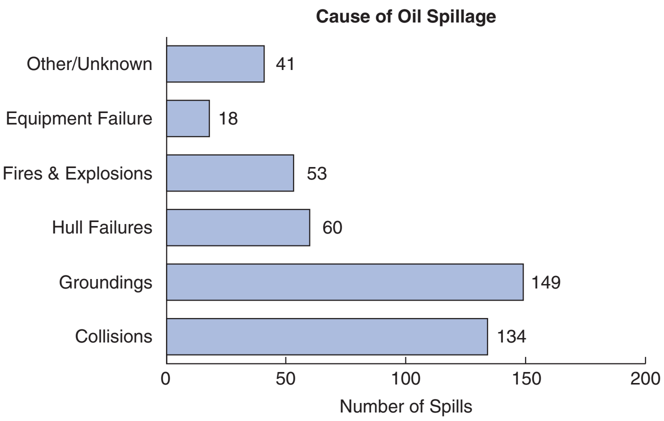
c) Would a pie chart be an effective way of communicating this information? Why or why not?

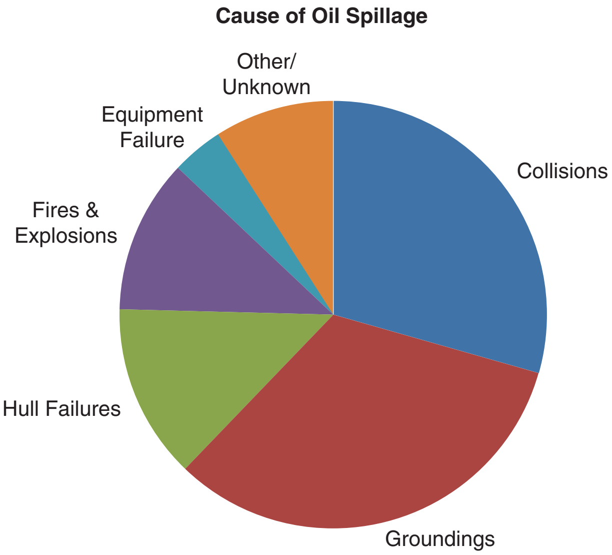
d) Write a couple of sentences on the steps taken by small businesses to increase productivity.

14. 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 labour. (See Exercise 13.) 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%. LO❶

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.

15. Environmental hazard. Data from the International Tanker Owners Pollution Federation Limited (www.itopf.com) give the cause of spillage for 455 large (>700 tonnes) oil tanker accidents from 1970–2012. 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? LO❶


Cause of Oil Spillage


Cause of Oil Spillage

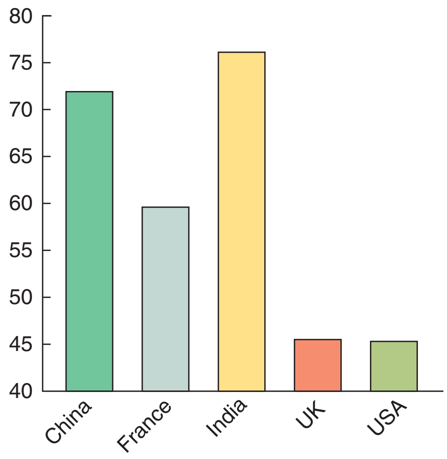
16. Winter Olympics 2010. Twenty-six countries won medals in the 2010 Winter Olympics in Vancouver-Whistler. The following table lists them, along with the total number of medals each won. Note that by virtue of winning more gold medals than any other country (including men’s and women’s hockey–hurray!), the Olympic Committee officially ranks Canada number one. LO❶

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

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?

17. 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. LO❶

a) How much larger is the proportion of those who said acquiring wealth was important in India than in the United States?

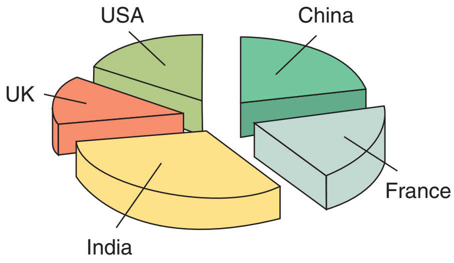


b) Is that the impression given by the display? Explain.
c) How would you improve this display?
d) Make an appropriate display for the percentages.
e) Write a few sentences describing what you have learned about attitudes toward acquiring wealth.

18. Importance of power. In the same survey as that discussed in Exercise 17, 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: **LO❶**

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.

19. 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 2005 to 2012. **LO❶**

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

		Year							
		2005	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**
Revenue Source	**Google websites**	55%	60%	64%	66%	67%	66%	69%	68%
	Google network websites	44%	39%	35%	31%	30%	30%	27%	27%
	Licensing & other revenue	1%	1%	1%	3%	3%	4%	4%	5%

20. Real estate pricing. A study of a sample of 1057 houses reports the following percentages of houses falling into different *Price* and *Size* categories. **LO❸**

		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 prices houses, what percent were small or medium small?
e) Write a few sentences describing the association between *Price* and *Size*.

21. Stock performance. The following table displays information for 40 widely held stocks that are popular among Canadian investors, on how their one-day change on March 15, 2007, compared with their previous 52-week change. **LO❸**

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

a) What percent of the companies reported a positive change in their share price over the prior 52 weeks?
b) What percent of the companies reported a positive change in their share price over both time periods?
c) What percent of the companies reported a negative change in their share price over both time periods?
d) What percent of the companies reported a positive change in their share price over one period and then a negative change in the other period?
e) Among those companies reporting a positive change in their share 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 share 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?

22. 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. LO④

		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?

23. Real estate. The Edmonton Real Estate Board (Realtors Association of Edmonton) website (www.ereb.com) provides data on sales activity in the Edmonton CMA (Census Metropolitan Area). The following table compares the number of sales in the January 2012 to those in January in 2013, the year over year change. LO③

		Type of Sale				
		Single Family	**Condos**	**Multi-family**	**Rural**	Total
Year	**2012**	543	219	51	52	**865**
	2013	496	265	59	57	**877**
	Total	**1039**	**484**	**110**	**109**	**1742**

a) What percent of all sales in January 2012 were condominiums (condos)? In January 2013?
b) What percent of all sales in January 2012 were multi-family? In January 2013?
c) Overall, what was the percentage change in January real estate sales in Edmonton from 2012 to 2013?

24. 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 2011? In 2012? LO③
b) What percent of all costs and expenses were due to research and development in 2011? In 2012?
c) Have general administrative costs grown as a percentage of all costs and expenses over this time period?

Cost & Expenses (millions of $)	2007	2008	2009	2010	2011	2012
Cost of revenues	$6649	$8622	$8844	$10 417	$13 188	$20 634
Research and development	$2120	$2793	$2843	$3762	$5162	$6793
Sales and marketing	$1461	$1946	$1984	$2799	$4589	$6143
General administrative	$1279	$1803	$1668	$1962	$2724	$3845
Miscellaneous	$0	$0	$0	$0	$500	$0
Total Costs and Expenses	**$11 510**	**$15 164**	**$15 339**	**$18 940**	**$26 163**	**$37 415**

Note: 2012 cost of revenues includes Motorola Mobile.

25. 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 the ratings of a random sample of 120 movies that were released last year, also classified by their genre. LO④

		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.

26. Smartphone use. A 2012 survey by Angus Reid/Vision Critical for Rogers asked smartphone users a variety of questions about their attitudes and behaviours with the device. The following table, adapted from the report, is a breakdown by age group of the question, "How close do you keep your cellphone/smartphone from you when you sleep at night?" LO❹

	Age		
Phone location at night	**18–34**	**35–54**	**55+**
In the bed with me	24	11	4
Nightstand beside bed	162	138	92
In the same room	35	46	33
In the next room	22	74	129
Downstairs/on another floor	22	64	129
Other	**5**	**21**	**29**

a) Complete the table by calculating the marginal distributions for the rows and columns.

b) Find the conditional distribution (in percentages) for each age group.

c) Create a graph that compares location by age group (in percentages).

d) Write a brief summary of what these data show about "phone location at night" and its relationship to age.

27. MBAs. Records of entering MBA students at the University of British Columbia from 2011 to 2013 include country of birth of the students. The following table compares the full-time program (FT) and part-time program (PT) by region of birth. LO❹

	Type		
Region of Birth	Full-time MBA	Part-time MBA	Total
North America	154	81	**235**
Asia/Pacific Rim	139	42	**181**
Europe	27	16	**43**
Middle East	6	13	**19**
Other	13	3	**16**
Total	**339**	**155**	**494**

a) What percent of all MBA students were from North America?

b) What percent of the full-time MBAs were from North America?

c) What percent of the part-time MBAs were from North America?

d) What is the marginal distribution of region of birth?

e) Obtain the column percentages and show the conditional distributions of region of birth by MBA program.

f) Do you think that region of birth of the MBA student is independent of the MBA program? Explain.

28. MBAs, part 2. The same university as in Exercise 27 reported the following data on the gender of their students in their two MBA programs. LO❹

	Full-time MBA	Part-time MBA	Total
Men	230	106	**336**
Women	109	49	**158**
Total	**339**	**155**	**494**

a) What percent of all MBA students are women?

b) What percent of full-time MBAs are women?

c) What percent of part-time 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?

Ⓣ **29. Top producing movies.** 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 2003 to 2012. (Data are number of films.) LO❹

	Rating				
Year	**G**	**PG**	**PG-13**	**R**	Total
2012	0	6	12	2	**20**
2011	3	3	12	2	**20**
2010	1	9	8	2	**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**
2004	1	6	10	3	**20**
2003	1	3	11	5	**20**
Total	**11**	**51**	**112**	**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 2006?

d) What percent of all top 20 films produced in 2007 or later were PG-13?

e) What percent of all top 20 films produced from 2003 to 2006 were PG-13?

f) Compare the conditional distributions of the ratings for films produced in 2007 or later to those produced in 2003 to 2006. Write a couple of sentences summarizing what you see.

30. 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. LO❹

	Patron Age						
Year	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 over 60 years old?

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.

31. Tattoos. A study by a medical centre 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. LO❷

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

32. 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. LO❷

	Year	
Response	**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

33. Revenue growth, last one. The study completed in 2005 and described in Exercise 10 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.) LO❷

	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*.

34. Low wage workers. Statistics Canada's Labour Force Survey 2004 data were analyzed to examine the incidence of low pay wages (defined as the percentage of employees earning less than $10.00 per hour). From the analysis, here is a table that presents the percentages, split by age and sex (www.rhdcc-hrsdc.gc.ca/eng/labour/employment_standards/fls/research/research02/page05.shtml) LO❷

Age	Male	Female
17–24	60.2%	69.2%
25–34	14.5%	22.8%
35–44	8.8%	19.6%
45–54	7.1%	19.4%
55–64	12.1%	24.9%

a) Is this a contingency table? Why or why not? Are segmented bar charts appropriate here?
b) Prepare a graphical display to compare the incidence of low pay for men to the incidence for women. Write a couple of sentences summarizing what you see.

35. 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. LO❹

		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 five years with a segmented bar graph.
d) Write a brief description of the association between *Year* and *Ethnicity* among these respondents.

36. Department store. A department store is planning its next advertising campaign. Because 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. LO❹

		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.

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

	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 centres by location.

38. Advertising. A company that distributes a variety of pet foods is planning their next advertising campaign. Because 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
Owning Pets (Percent)

		Pet			
		Dog	**Cat**	**Bird**	**Horse**
Income	**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.

39. 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 North America in 2003, accumulated by the International Council of Toy Industries (www.toy-icti.org). LO❷

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 North America are sold in Europe? Why or why not?
c) Use a graphical display to compare the distribution of channels between Europe and North America.
d) Summarize the distribution of toy sales by channel in a few sentences. What are the biggest differences between these two continents?

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

40. Internet users. Internet World Stats tracks internet usage and population for over 233 individual countries and world regions. The website (www.internetworldstats.com) reports that, as of June 30, 2012, there were 2.4 billion internet users worldwide. The site also reports users by World Region, as follows: LO❸

	Population (millions)	Internet Users (millions)
Africa	1073	167
Asia	3922	1077
Europe	821	519
Middle East	224	90
North America	348	274
Latin America/Caribbean	594	255
Oceania/Australia	36	24
World Total	7018	2406

a) What percent of North Americans use the internet?
b) What percent of internet users are from North America?
c) Draw a graph to compare the percentage of the population who are internet users across regions.

41. Health care. A provincial ministry of health is concerned that patients who undergo surgery at large hospitals have their discharges delayed for various reasons—which results in increased medical costs. The recent data for area hospitals and two types of surgery (major and minor) are shown in the following table. LO❸, LO❺

		Discharge Delayed	
		Large Hospital	**Small Hospital**
Procedure	**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 size of hospital?
d) What were the delay rates at each size of hospital for each kind of surgery?
e) The ministry of health is considering advising patients 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.

42. 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. LO❸, LO❺

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?

43. Graduate admissions. This case is old but it's a "classic". 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. LO❸, LO❺

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?

44. 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. LO❸, LO❺

JUST CHECKING ANSWERS

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