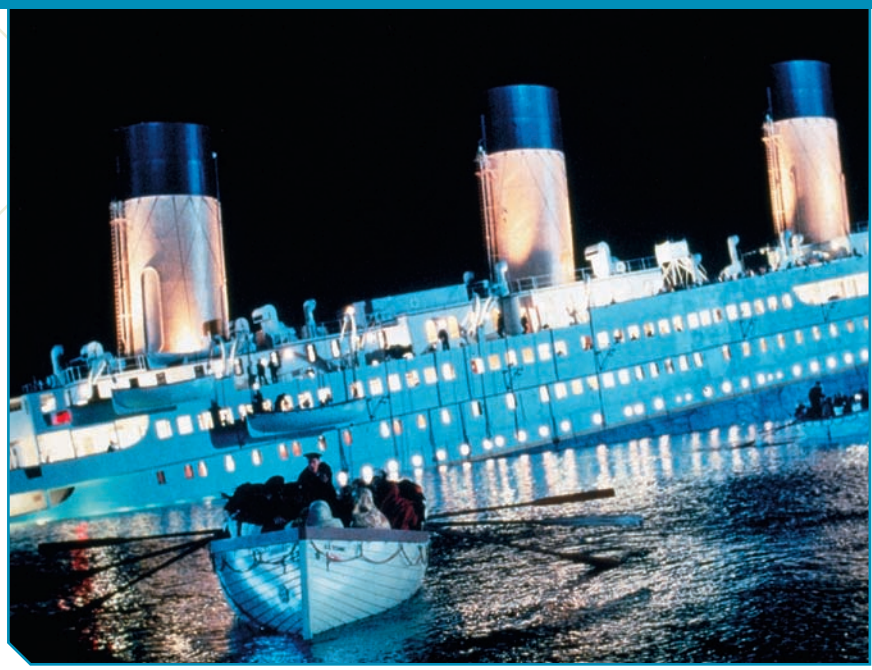
CHAPTER 3

Displaying and Describing Categorical Data



WHO	People on the *Titanic*
WHAT	Survival status, age, sex, ticket class
WHEN	April 14, 1912
WHERE	North Atlantic
HOW	A variety of sources and Internet sites
WHY	Historical interest

What happened on the *Titanic* at 11:40 on the night of April 14, 1912, is well known. Frederick Fleet's cry of "Iceberg, right ahead" and the three accompanying pulls of the crow's nest bell signaled the beginning of a nightmare that has become legend. By 2:15 a.m., the *Titanic*, thought by many to be unsinkable, had sunk, leaving more than 1500 passengers and crew members on board to meet their icy fate.

Here are some data about the passengers and crew aboard the *Titanic*. Each case (row) of the data table represents a person on board the ship. The variables are the person's *Survival* status (Dead or Alive), *Age* (Adult or Child), *Sex* (Male or Female), and ticket *Class* (First, Second, Third, or Crew).

The problem with a data table like this—and in fact with all data tables—is that you can't *see* what's going on. And seeing is just what we want to do. We need ways to show the data so that we can see patterns, relationships, trends, and exceptions.

AS ***Video:* The Incident** tells the story of the *Titanic*, and includes rare film footage.

Survival	Age	Sex	Class
Dead	Adult	Male	Third
Dead	Adult	Male	Crew
Dead	Adult	Male	Third
Dead	Adult	Male	Crew
Dead	Adult	Male	Crew
Dead	Adult	Male	Crew
Alive	Adult	Female	First
Dead	Adult	Male	Third
Dead	Adult	Male	Crew

Table 3.1
Part of a data table showing four variables for nine people aboard the *Titanic*.

The Three Rules of Data Analysis

So, what should we do with data like these? There are three things you should always do first with data:

1. **Make a picture.** A display of your data will reveal things you are not likely to see in a table of numbers and will help you to *Think* clearly about the patterns and relationships that may be hiding in your data.
2. **Make a picture.** A well-designed display will *Show* the important features and patterns in your data. A picture will also show you the things you did not expect to see: the extraordinary (possibly wrong) data values or unexpected patterns.
3. **Make a picture.** The best way to *Tell* others about your data is with a well-chosen picture.

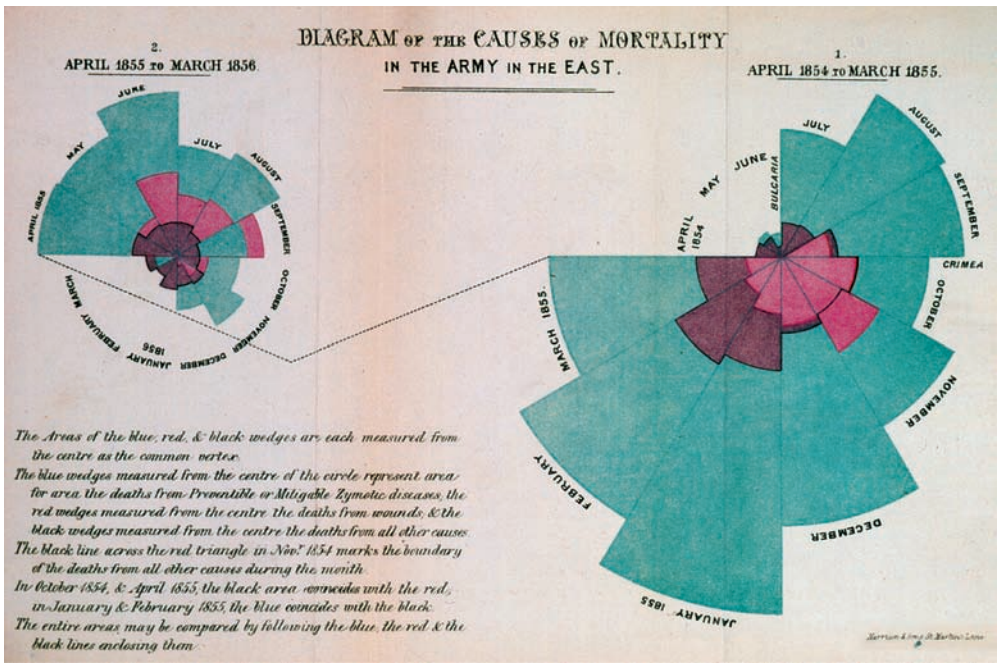
These are the three rules of data analysis. There are pictures of data throughout the book, and new kinds keep showing up. These days, technology makes drawing pictures of data easy, so there is no reason not to follow the three rules.

FIGURE 3.1 ***A Picture to Tell a Story***

Florence Nightingale (1820–1910), a founder of modern nursing, was also a pioneer in health management, statistics, and epidemiology. She was the first female member of the British Statistical Society and was granted honorary membership in the newly formed American Statistical Association.

To argue forcefully for better hospital conditions for soldiers, she and her colleague, Dr. William Farr, invented this display, which showed that in the Crimean War, far more soldiers died of illness and infection than of battle wounds. Her campaign succeeded in improving hospital conditions and nursing for soldiers.

Florence Nightingale went on to apply statistical methods to a variety of important health issues and published more than 200 books, reports, and pamphlets during her long and illustrious career.



Frequency Tables: Making Piles

A S ***Activity:*** **Make and examine a table of counts.** Even data on something as simple as hair color can reveal surprises when you organize it in a data table.

To make a picture of data, the first thing we have to do is to make piles. Making piles is the beginning of understanding about data. We pile together things that seem to go together, so we can see how the cases distribute across different categories. For categorical data, piling is easy. We just count the number of cases corresponding to each category and pile them up.

One way to put all 2201 people on the *Titanic* into piles is by ticket *Class*, counting up how many had each kind of ticket. We can organize these counts into a **frequency table,** which records the totals and the category names.

Even when we have thousands of cases, a variable like ticket *Class*, with only a few categories, has a frequency table that's easy to read. A frequency table with dozens or hundreds of categories would be much harder to read. We use the names of the categories to label each row in the frequency table. For ticket *Class*, these are "First," "Second," "Third," and "Crew."

Class	Count
First	325
Second	285
Third	706
Crew	885

Table 3.2

A frequency table of the *Titanic* passengers.

Class	%
First	14.77
Second	12.95
Third	32.08
Crew	40.21

Table 3.3
A relative frequency table for the same data.

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** 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.

The Area Principle

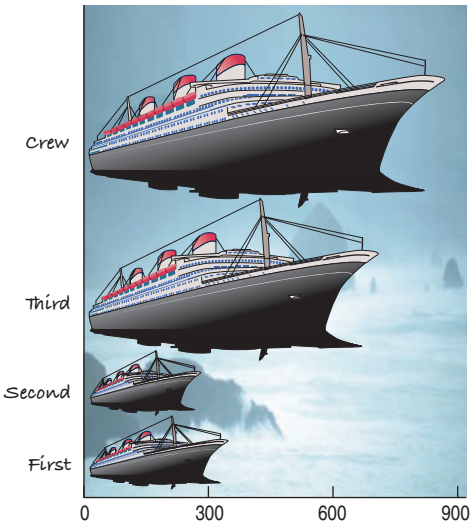


FIGURE 3.2
How many people were in each class on the Titanic? From this display, it looks as though the service must have been great, since most aboard were crew members. Although the length of each ship here corresponds to the correct number, the impression is all wrong. In fact, only about 40% were crew.

Now that we have the frequency table, we're ready to follow the three rules of data analysis and make a picture of the data. But a bad picture can distort our understanding rather than help it. Here's a graph of the *Titanic* data. What impression do you get about who was aboard the ship?

It sure looks like most of the people on the *Titanic* were crew members, with a few passengers along for the ride. That doesn't seem right. What's wrong? The lengths of the ships *do* match the totals in the table. (You can check the scale at the bottom.) However, experience and psychological tests show that our eyes tend to be more impressed by the *area* than by other aspects of each ship image. So, even though the *length* of each ship matches up with one of the totals, it's the associated *area* in the image that we notice. Since there were about 3 times as many crew as second-class passengers, the ship depicting the number of crew is about 3 times longer than the ship depicting second-class passengers, but it occupies about 9 times the area. As you can see from the frequency table (Table 3.2), that just isn't a correct impression.

The best data displays observe a fundamental principle of graphing data called the **area principle**. The area principle says that the area occupied by a part of the graph should correspond to the magnitude of the value it represents. Violations of the area principle are a common way to lie (or, since most mistakes are unintentional, we should say err) with Statistics.

Bar Charts

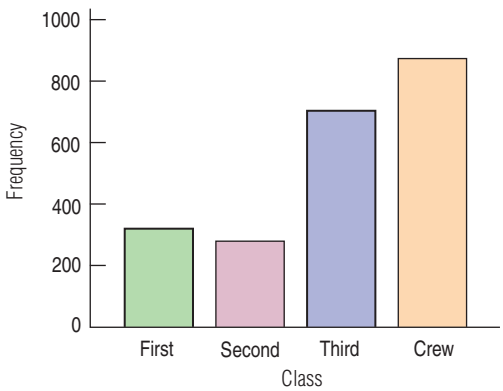
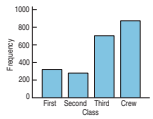
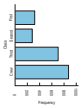


FIGURE 3.3 ***People on the* Titanic *by Ticket Class***
With the area principle satisfied, we can see the true distribution more clearly.

Here's a chart that obeys the area principle. It's not as visually entertaining as the ships, but it does give an *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 the majority of people on board were *not* crew, as the ships picture led us to believe. We can also see that there were about 3 times as many crew as second-class passengers. And there were more than twice as many third-class passengers as either first- or second-class passengers, something you may have missed in the frequency table. 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.

Usually they stick up like this  but sometimes they run sideways like this 

If we really want to draw attention to the relative *proportion* of passengers falling into each of these classes, we could replace the counts with percentages and use a **relative frequency bar chart.**

A S ***Activity:*** **Bar Charts.** Watch bar charts grow from data; then use your statistics package to create some bar charts for yourself.

For some reason, some computer programs give the name "bar chart" to any graph that uses bars. And others use different names according to whether the bars are horizontal or vertical. Don't be misled. "Bar chart" is the term for a *display of counts of a categorical variable* with bars.

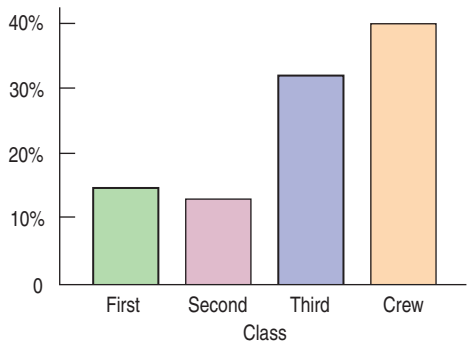


FIGURE 3.4
The relative frequency bar chart looks the same as the bar chart (Figure 3.3) but shows the proportion of people in each category rather than the counts.

Pie Charts

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

Pie charts give a quick impression of how a whole group is partitioned into smaller groups. Because we're used to cutting up pies into 2, 4, or 8 pieces, pie charts are good for seeing relative frequencies near 1/2, 1/4, or 1/8. For example, you may be able to tell that the pink slice, representing the second-class passengers, is very close to 1/8 of the total. It's harder to see that there were about twice as many third-class as first-class passengers. Which category had the most passengers? Were there more crew or more third-class passengers? Comparisons such as these are easier in a bar chart.

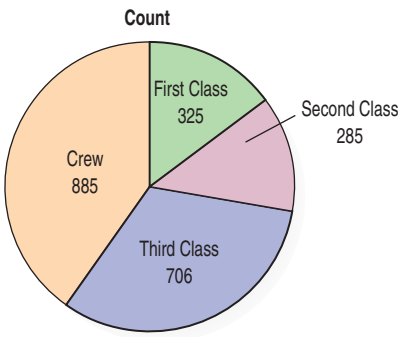


FIGURE 3.5 ***Number of*** **Titanic** ***passengers in each class***

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 each type of graph is a critical first step in data analysis. That decision depends in part on what type of data we have.

It's important to check that the data are appropriate for whatever method of analysis you choose. Before you make a bar chart or a pie chart, always check the

Categorical Data Condition: The data are counts or percentages of individuals in categories.

If you want to make a relative frequency bar chart or a pie chart, you'll need to also make sure that the categories don't overlap so that no individual is counted twice. If the categories do overlap, you can still make a bar chart, but the percentages won't add up to 100%. For the *Titanic* data, either kind of display is appropriate because the categories don't overlap.

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, then, is to check that the type of analysis you plan is appropriate. The Categorical Data Condition is just the first of many such checks.

Contingency Tables: Children and First-Class Ticket Holders First?

A S ***Activity:* Children at Risk.** This activity looks at the fates of children aboard the *Titanic*; the subsequent activity shows how to make such tables on a computer.

We know how many tickets of each class were sold on the *Titanic,* and we know that only about 32% of all those aboard the *Titanic* survived. After looking at the distribution of each variable by itself, it's natural and more interesting to ask how they relate. Was there a relationship between the kind of ticket a passenger held and the passenger's chances of making it into the lifeboat? To answer this question, we need to look at the two categorical variables *Class* and *Survival* together.

To look at two categorical variables together, we often arrange the counts in a two-way table. Here is a two-way table of those aboard the *Titanic,* classified according to the class of ticket and whether the ticket holder survived or didn't. Because the table shows how the individuals are distributed along each variable, contingent on the value of the other variable, such a table is called a **contingency table.**

Contingency table of ticket *Class* and *Survival.* The bottom line of "Totals" is the same as the previous frequency table.
Table 3.4

		Class				
		First	Second	Third	Crew	Total
Survival	Alive	203	118	178	212	**711**
	Dead	122	167	528	673	**1490**
	Total	**325**	**285**	**706**	**885**	**2201**

A bell-shaped artifact from the Titanic.

The margins of the table, both on the right and at the bottom, give totals. The bottom line of the table is just the frequency distribution of ticket *Class.* The right column of the table is the frequency distribution of the variable *Survival.* When presented like this, in the margins of a contingency table, the frequency distribution of one of the variables is called its **marginal distribution.**

Each **cell** of the table gives the count for a combination of values of the two variables. If you look down the column for second-class passengers to the first cell, you can see that 118 second-class passengers survived. Looking at the third-class passengers, you can see that more third-class passengers (178) survived. Were second-class passengers more likely to survive? Questions like this are easier to address by using percentages. The 118 survivors in second class were 41.4% of the total 285 second-class passengers, while the 178 surviving third-class passengers were only 25.2% of that class's total.

We know that 118 second-class passengers survived. We could display this number as a percentage—but as a percentage of what? The total number of passengers? (118 is 5.4% of the total: 2201.) The number of second-class passengers?

(118 is 41.4% of the 285 second-class passengers.) The number of survivors? (118 is 16.6% of the 711 survivors.) 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 holds lots of information, but it can be hard to understand:

Another contingency table of ticket Class. This time we see not only the counts for each combination of *Class* and *Survival* (in bold) but 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.

Table 3.5

			Class				
			First	Second	Third	Crew	Total
Survival	Alive	Count	203	118	178	212	711
		% of Row	28.6%	16.6%	25.0%	29.8%	100%
		% of Column	62.5%	41.4%	25.2%	24.0%	32.3%
		% of Table	9.2%	5.4%	8.1%	9.6%	32.3%
	Dead	Count	122	167	528	673	1490
		% of Row	8.2%	11.2%	35.4%	45.2%	100%
		% of Column	37.5%	58.6%	74.8%	76.0%	67.7%
		% of Table	5.6%	7.6%	24.0%	30.6%	67.7%
	Total	Count	325	285	706	885	2201
		%of Row	14.8%	12.9%	32.1%	40.2%	100%
		% of Column	100%	100%	100%	100%	100%
		% of Table	14.8%	12.9%	32.1%	40.2%	100%

To simplify the table, let's first pull out the percent of table values:

A contingency table of *Class* by *Survival* with only the table percentages

Table 3.6

		Class				
		First	Second	Third	Crew	Total
Survival	Alive	9.2%	5.4%	8.1%	9.6%	32.3%
	Dead	5.6%	7.6%	24.0%	30.6%	67.7%
	Total	14.8%	12.9%	32.1%	40.2%	100%

These percentages tell us what percent of *all* passengers belong to each combination of column and row category. For example, we see that although 8.1% of the people aboard the *Titanic* were surviving third-class ticket holders, only 5.4% were surviving second-class ticket holders. Is this fact useful? Comparing these percentages, you might think that the chances of surviving were better in third class than in second. But be careful. There were many more third-class than second-class passengers on the *Titanic,* so there were more third-class survivors. That group is a larger percentage of the passengers, but is that really what we want to know?

Percent of what? The English language can be tricky when we talk about percentages. If you're asked "What percent *of the survivors* were in second class?" it's pretty clear that we're interested only in survivors. It's as if we're restricting the *Who* in the question to the survivors, so we should look at the number of second-class passengers among all the survivors—in other words, the row percent.

But if you're asked "What percent were second-class passengers who survived?" you have a different question. Be careful; here, the *Who* is everyone on board, so 2201 should be the denominator, and the answer is the table percent.

And if you're asked "What percent of the second-class passengers survived?" you have a third question. Now the *Who* is the second-class passengers, so the denominator is the 285 second-class passengers, and the answer is the column percent.

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

FOR EXAMPLE Finding marginal distributions

In January 2007, a Gallup poll asked 1008 Americans age 18 and over whether they planned to watch the upcoming Super Bowl. The pollster also asked those who planned to watch whether they were looking forward more to seeing the football game or the commercials. The results are summarized in the table:

Response	Sex: Male	Female	Total
Game	279	200	479
Commercials	81	156	237
Won't watch	132	160	292
Total	492	516	1008

Question: What's the marginal distribution of the responses?

To determine the percentages for the three responses, divide the count for each response by the total number of people polled:

$$\frac{479}{1008} = 47.5\% \quad \frac{237}{1008} = 23.5\% \quad \frac{292}{1008} = 29.0\%$$

According to the poll, 47.5% of American adults were looking forward to watching the Super Bowl game, 23.5% were looking forward to watching the commercials, and 29% didn't plan to watch at all.

Conditional Distributions

The more interesting questions are *contingent*. We'd like to know, for example, what percentage of *second-class passengers* survived and how that compares with the survival rate for third-class passengers.

It's more interesting to ask whether the chance of surviving the *Titanic* sinking *depended* on ticket class. We can look at this question in two ways. First, we could ask how the distribution of ticket *Class* changes between survivors and nonsurvivors. To do that, we look at the *row percentages:*

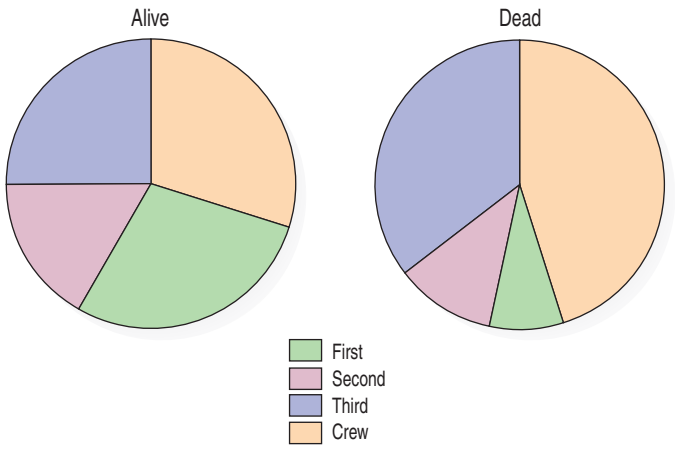
Table 3.7 **The conditional distribution of ticket *Class*** conditioned on each value of *Survival: Alive* and *Dead*.

Survival	Class: First	Second	Third	Crew	Total
Alive	203 28.6%	118 16.6%	178 25.0%	212 29.8%	711 100%
Dead	122 8.2%	167 11.2%	528 35.4%	673 45.2%	1490 100%

By focusing on each row separately, we see the distribution of class under the *condition* of surviving or not. The sum of the percentages in each row is 100%, and we divide that up by ticket class. In effect, we temporarily restrict the *Who* first to survivors and make a pie chart for them. Then we refocus the *Who* on the nonsurvivors and make their pie chart. These pie charts show the distribution of ticket classes *for each row* of the table: survivors and nonsurvivors. The distributions we create this way are called **conditional distributions,** because they show the distribution of one variable for just those cases that satisfy a condition on another variable.

FIGURE 3.6

Pie charts of the conditional distributions of ticket Class *for the survivors and nonsurvivors, separately. Do the distributions appear to be the same? We're primarily concerned with percentages here, so pie charts are a reasonable choice.*



FOR EXAMPLE Finding conditional distributions

Recap: The table shows results of a poll asking adults whether they were looking forward to the Super Bowl game, looking forward to the commercials, or didn't plan to watch.

Question: How do the conditional distributions of interest in the commercials differ for men and women?

Response	Sex: Male	Female	Total
Game	279	200	479
Commercials	81	156	237
Won't watch	132	160	292
Total	492	516	1008

Look at the group of people who responded "Commercials" and determine what percent of them were male and female:

$$\frac{81}{237} = 34.2\% \quad \frac{156}{237} = 65.8\%$$

Women make up a sizable majority of the adult Americans who look forward to seeing Super Bowl commercials more than the game itself. Nearly 66% of people who voiced a preference for the commercials were women, and only 34% were men.

But we can also turn the question around. We can look at the distribution of *Survival* for each category of ticket *Class*. To do this, we look at the *column percentages*. Those show us whether the chance of surviving was roughly the same *for each of the four classes*. Now the percentages in each column add to 100%, because we've restricted the *Who*, in turn, to each of the four ticket classes:

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

Table 3.8

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

Looking at how the percentages change across each row, it sure looks like ticket class mattered in whether a passenger survived. To make it more vivid, we could show the distribution of *Survival* for each ticket class in a display. Here's a side-by-side bar chart showing percentages of surviving and not for each category:

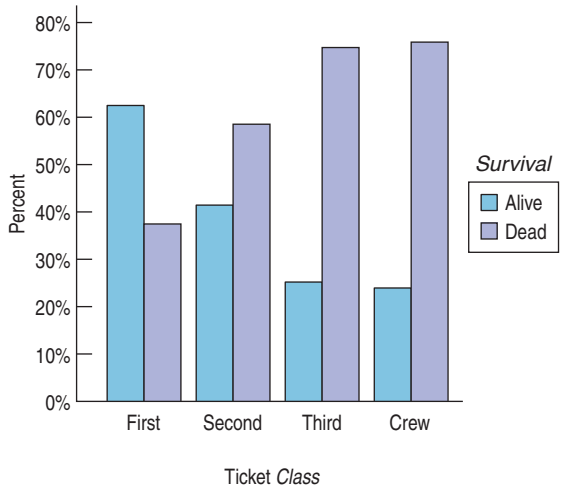


FIGURE 3.7
Side-by-side bar chart *showing the conditional distribution of Survival for each category of ticket Class. The corresponding pie charts would have only two categories in each of four pies, so bar charts seem the better alternative.*

These bar charts are simple because, for the variable *Survival*, we have only two alternatives: Alive and Dead. When we have only two categories, we really need to know only the percentage of one of them. Knowing the percentage that survived tells us the percentage that died. We can use this fact to simplify the display even more by dropping one category. Here are the percentages of dying *across the classes* displayed in one chart:

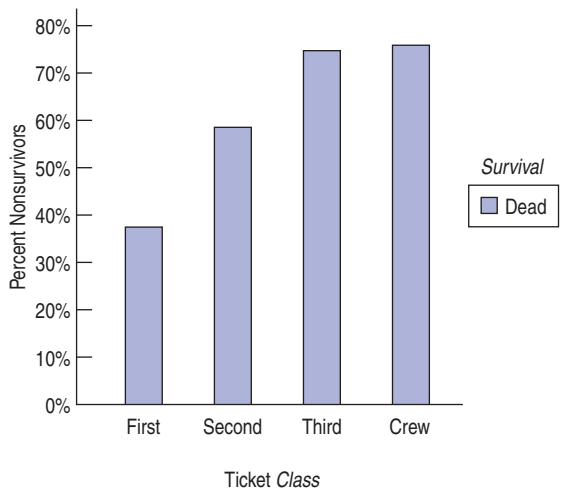


FIGURE 3.8
Bar chart showing just nonsurvivor percentages for each value of ticket Class. Because we have only two values, the second bar doesn't add any information. Compare this chart to the side-by-side bar chart shown earlier.

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

If the risk had been about the same across the ticket classes, we would have said that survival was *independent* of class. But it's not. The differences we see among these conditional distributions suggest that survival may have depended on ticket class. You may find it useful to consider conditioning on each variable in a contingency table in order to explore the dependence between them.

TI-*nspire*

Conditional distributions and association. Explore the *Titanic* data to see which passengers were most likely to survive.

It is interesting to know that *Class* and *Survival* are associated. That's an important part of the *Titanic* story. And we know how important this is because the margins show us the actual numbers of people involved.

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*.[1] 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.

FOR EXAMPLE Looking for associations between variables

Recap: The table shows results of a poll asking adults whether they were looking forward to the Super Bowl game, looking forward to the commercials, or didn't plan to watch.

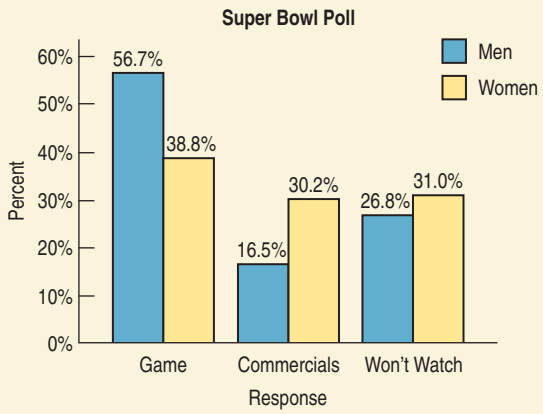
Question: Does it seem that there's an association between interest in Super Bowl TV coverage and a person's sex?

Response	Sex: Male	Sex: Female	Total
Game	279	200	479
Commercials	81	156	237
Won't watch	132	160	292
Total	492	516	1008

First find the distribution of the three responses for the men (the column percentages):

$$\frac{279}{492} = 56.7\% \qquad \frac{81}{492} = 16.5\% \qquad \frac{132}{492} = 26.8\%$$

Then do the same for the women who were polled, and display the two distributions with a side-by-side bar chart:



Based on this poll it appears that women were only slightly less interested than men in watching the Super Bowl telecast: 31% of the women said they didn't plan to watch, compared to just under 27% of men. Among those who planned to watch, however, there appears to be an association between the viewer's sex and what the viewer is most looking forward to. While more women are interested in the game (39%) than the commercials (30%), the margin among men is much wider: 57% of men said they were looking forward to seeing the game, compared to only 16.5% who cited the commercials.

[1] This kind of "backwards" reasoning shows up surprisingly often in science—and in Statistics. We'll see it again.

JUST CHECKING

A Statistics class reports the following data on Sex and Eye Color for students in the class:

		Eye Color			
		Blue	Brown	Green/Hazel/Other	Total
Sex	Males	6	20	6	32
	Females	4	16	12	32
	Total	10	36	18	64

1. What percent of females are brown-eyed?
2. What percent of brown-eyed students are female?
3. What percent of students are brown-eyed females?
4. What's the distribution of Eye Color?
5. What's the conditional distribution of Eye Color for the males?
6. Compare the percent who are female among the blue-eyed students to the percent of all students who are female.
7. Does it seem that Eye Color and Sex are independent? Explain.

Segmented Bar Charts

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

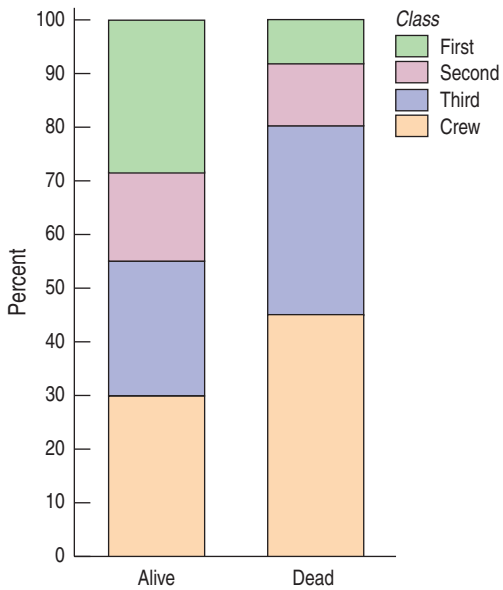


FIGURE 3.9 ***A segmented bar chart for Class by Survival***

Notice that although the totals for survivors and nonsurvivors are quite different, the bars are the same height because we have converted the numbers to percentages. *Compare this display with the side-by-side pie charts of the same data in Figure 3.6.*

STEP-BY-STEP EXAMPLE — Examining Contingency Tables

Medical researchers followed 6272 Swedish men for 30 years to see if there was any association between the amount of fish in their diet and prostate cancer ("Fatty Fish Consumption and Risk of Prostate Cancer," *Lancet*, June 2001). Their results are summarized in this table:

We asked for a picture of a man eating fish. This is what we got.

Fish Consumption	Prostate Cancer: No	Prostate Cancer: Yes
Never/seldom	110	14
Small part of diet	2420	201
Moderate part	2769	209
Large part	507	42

Table 3.9

Question: Is there an association between fish consumption and prostate cancer?

Plan Be sure to state what the problem is about.

I want to know if there is an association between fish consumption and prostate cancer.

Variables Identify the variables and report the W's.

The individuals are 6272 Swedish men followed by medical researchers for 30 years. The variables record their fish consumption and whether or not they were diagnosed with prostate cancer.

Be sure to check the appropriate condition.

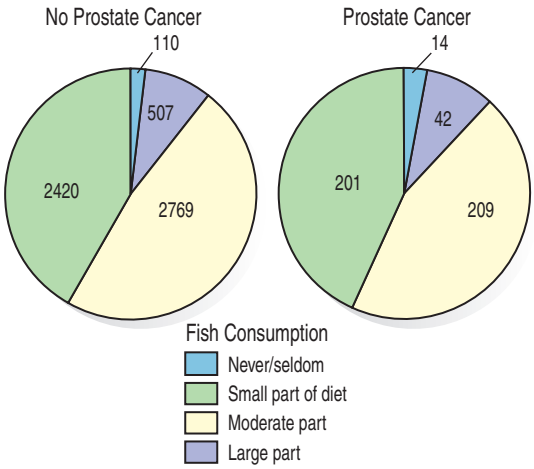
✓ ***Categorical Data Condition:*** I have counts for both fish consumption and cancer diagnosis. The categories of diet do not overlap, and the diagnoses do not overlap. It's okay to draw pie charts or bar charts.

Mechanics It's a good idea to check the marginal distributions first before looking at the two variables together.

Fish Consumption	Prostate Cancer: No	Prostate Cancer: Yes	Total
Never/seldom	110	14	**124 (2.0%)**
Small part of diet	2420	201	**2621 (41.8%)**
Moderate part	2769	209	**2978 (47.5%)**
Large part	507	42	**549 (8.8%)**
Total	**5806 (92.6%)**	**466 (7.4%)**	**6272 (100%)**

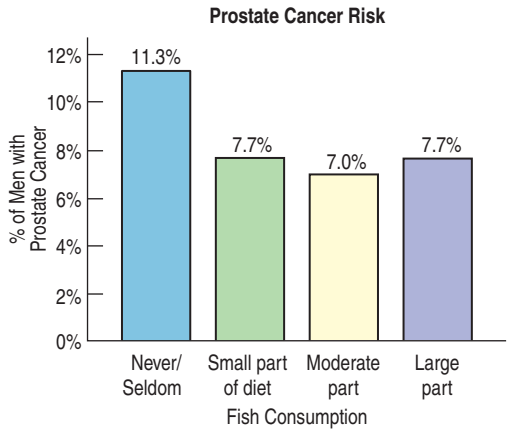
Two categories of the diet are quite small, with only 2.0% Never/Seldom eating fish and 8.8% in the "Large part" category. Overall, 7.4% of the men in this study had prostate cancer.

Then, make appropriate displays to see whether there is a difference in the relative proportions. These pie charts compare fish consumption for men who have prostate cancer to fish consumption for men who don't.



It's hard to see much difference in the pie charts. So, I made a display of the row percentages. Because there are only two alternatives, I chose to display the risk of prostate cancer for each group:

Both pie charts and bar charts can be used to compare conditional distributions. Here we compare prostate cancer rates based on differences in fish consumption.



Conclusion Interpret the patterns in the table and displays in context. If you can, discuss possible real-world consequences. Be careful not to overstate what you see. The results may not generalize to other situations.

Overall, there is a 7.4% rate of prostate cancer among men in this study. Most of the men (89.3%) ate fish either as a moderate or small part of their diet. From the pie charts, it's hard to see a difference in cancer rates among the groups. But in the bar chart, it looks like the cancer rate for those who never/seldom ate fish may be somewhat higher.

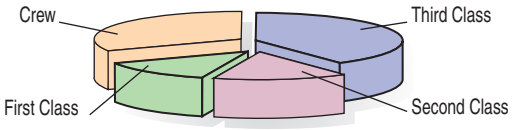
However, only 124 of the 6272 men in the study fell into this category, and only 14 of them developed prostate cancer. More study would probably be needed before we would recommend that men change their diets.[2]

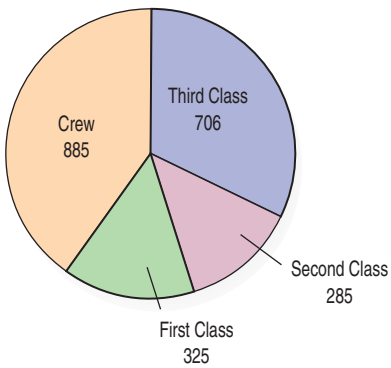
[2] The original study actually used pairs of twins, which enabled the researchers to discern that the risk of cancer for those who never ate fish actually *was* substantially greater. Using pairs is a special way of gathering data. We'll discuss such study design issues and how to analyze the data in the later chapters.

This study is an example of looking at a sample of data to learn something about a larger population. We care about more than these particular 6272 Swedish men. We hope that learning about their experiences will tell us something about the value of eating fish in general. That raises the interesting question of what population we think this sample might represent. Do we hope to learn about all Swedish men? About all men? About the value of eating fish for all adult humans? [3] Often, it can be hard to decide just which population our findings may tell us about, but that also is how researchers decide what to look into in future studies.

WHAT CAN GO WRONG?

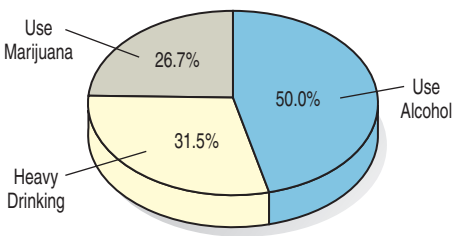
- **Don't violate the area principle.** This is probably the most common mistake in a graphical display. It is often made in the cause of artistic presentation. Here, for example, are two displays of the pie chart of the *Titanic* passengers by class:





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

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

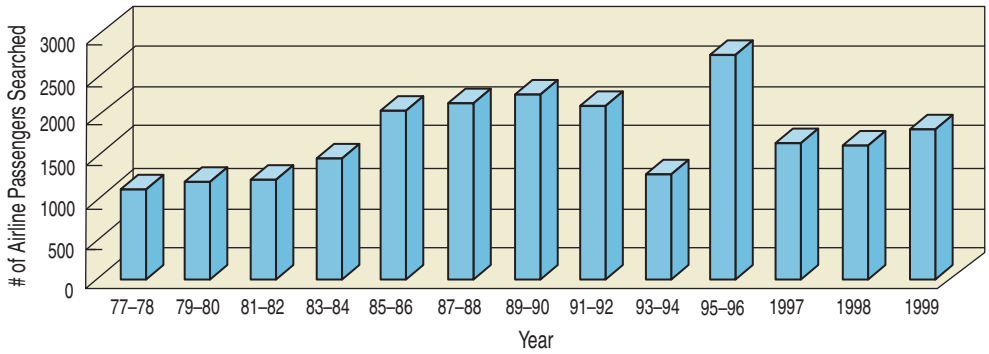


Try adding up the percentages. Or look at the 50% slice. Does it look right? Then think: What are these percentages of? Is there a "whole" that has been sliced up? In a pie chart, the proportions shown by each slice of the pie must add up to 100% and each individual must fall into only one category.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

(*continued*)

[3] Probably not, since we're looking only at prostate cancer risk.

Here's another. This bar chart shows the number of airline passengers searched in security screening, by year:



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

- **Don't confuse similar-sounding percentages.** These percentages sound similar but are different:
  - The percentage of the passengers who were both in first class and survived: This would be 203/2201, or 9.4%.
  - The percentage of the first-class passengers who survived: This is 203/325, or 62.5%.
  - The percentage of the survivors who were in first class: This is 203/711, or 28.6%.

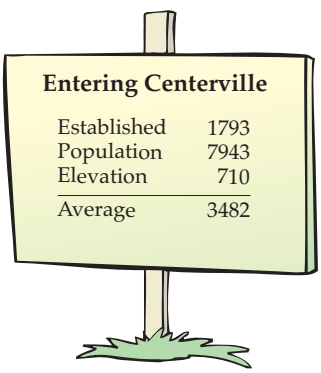
		Class				
		First	Second	Third	Crew	Total
Survival	Alive	203	118	178	212	711
	Dead	122	167	528	673	1490
	Total	325	285	706	885	2201

In each instance, pay attention to the *Who* implicitly defined by the phrase. Often there is a restriction to a smaller group (all aboard the *Titanic*, those in first class, and those who survived, respectively) before a percentage is found. 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 you 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. Take care not to make a report such as this one:

  We found that 66.67% of the rats improved their performance with training. The other rat died.



- **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.

Simpson's Paradox

- **Don't use unfair or silly averages.** Sometimes averages can be misleading. Sometimes they just don't make sense at all. Be careful when averaging different variables that the quantities you're averaging are comparable. The Centerville sign says it all.

  When using averages of proportions across several different groups, it's important to make sure that the groups really are comparable.

It's easy to make up an example showing that averaging across very different values or groups can give absurd results. Here's how that might work: Suppose there are two pilots, Moe and Jill. Moe argues that he's the better pilot of the two, since he managed to land 83% of his last 120 flights on time compared with Jill's 78%. But let's look at the data a little more closely. Here are the results for each of their last 120 flights, broken down by the time of day they flew:

Table 3.10
On-time flights by *Time of Day* and *Pilot*. Look at the percentages within each *Time of Day* category. Who has a better on-time record during the day? At night? Who is better overall?

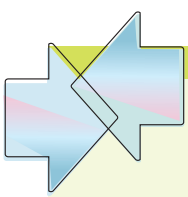
		Time of Day		
		Day	**Night**	**Overall**
Pilot	**Moe**	90 out of 100 90%	10 out of 20 50%	100 out of 120 83%
	Jill	19 out of 20 95%	75 out of 100 75%	94 out of 120 78%

Look at the daytime and nighttime flights separately. For day flights, Jill had a 95% on-time rate and Moe only a 90% rate. At night, Jill was on time 75% of the time and Moe only 50%. So Moe is better "overall," but Jill is better both during the day and at night. How can this be?

What's going on here is a problem known as **Simpson's paradox,** named for the statistician who discovered it in the 1960s. It comes up rarely in real life, but there have been several well-publicized cases. As we can see from the pilot example, the problem is *unfair averaging* over different groups. Jill has mostly night flights, which are more difficult, so her *overall average* is heavily influenced by her nighttime average. Moe, on the other hand, benefits from flying mostly during the day, with its higher on-time percentage. With their very different patterns of flying conditions, taking an overall average is misleading. It's not a fair comparison.

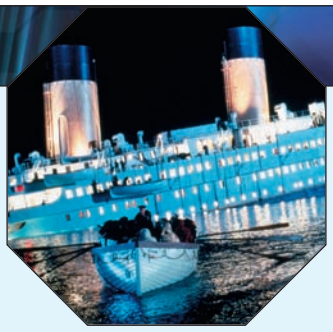
The moral of Simpson's paradox is to be careful when you average across different levels of a second variable. It's always better to compare percentages or other averages *within* each level of the other variable. The overall average may be misleading.

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



CONNECTIONS

All of the methods of this chapter work with *categorical variables.* You must know the *Who* of the data to know who is counted in each category and the *What* of the variable to know where the categories come from.



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 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 by using bar charts or pie charts.
- If the conditional distributions of one variable are (roughly) the same for every category of the other, the variables are independent.

Terms

Frequency table (Relative frequency table)
21. A frequency table lists the categories in a categorical variable and gives the count (or percentage of observations for each category.

Distribution
22. The distribution of a variable gives
- the possible values of the variable and
- the relative frequency of each value.

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

Bar chart (Relative frequency bar chart)
22. Bar charts show a bar whose area represents the count (or percentage) of observations for each category of a categorical variable.

Pie chart
23. 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.

Categorical data condition
24. The methods in this chapter are appropriate for displaying and describing categorical data. Be careful not to use them with quantitative data.

Contingency table
24. 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.

Marginal distribution
24. In a contingency table, the distribution of either variable alone is called the marginal distribution. The counts or percentages are the totals found in the margins (last row or column) of the table.

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

Independence
29. Variables are said to be independent if the conditional distribution of one variable is the same for each category of the other. We'll show how to check for independence in a later chapter.

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

Simpson's paradox
34. When averages are taken across different groups, they can appear to contradict the overall averages. This is known as "Simpson's paradox."

Skills

THINK
- Be able to 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.

SHOW
- Be able to summarize the distribution of a categorical variable with a frequency table.
- Be able to display the distribution of a categorical variable with a bar chart or pie chart.
- Know how to make and examine a contingency table.

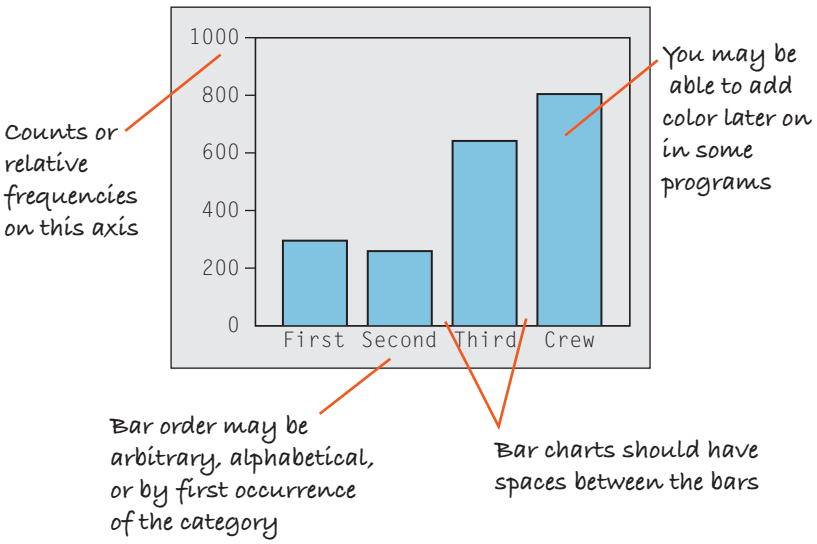
- Know how to make and examine displays of the conditional distributions of one variable for two or more groups.

TELL

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

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.

EXERCISES

1. **Graphs in the news.** Find a bar graph of categorical data from a newspaper, a magazine, or the Internet.
   a) Is the graph clearly labeled?
   b) Does it violate the area principle?
   c) Does the accompanying article tell the W's of the variable?
   d) Do you think the article correctly interprets the data? Explain.

2. **Graphs in the news II.** Find a pie chart of categorical data from a newspaper, a magazine, or the Internet.
   a) Is the graph clearly labeled?
   b) Does it violate the area principle?
   c) Does the accompanying article tell the W's of the variable?
   d) Do you think the article correctly interprets the data? Explain.