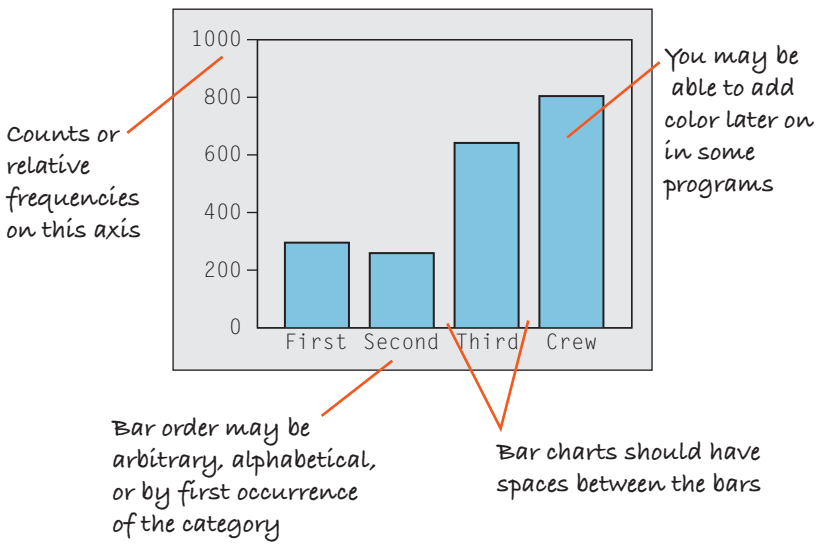
- Know how to make and examine displays of the conditional distributions of one variable for two or more groups.

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

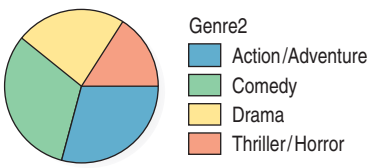
1. **Graphs in the news.** Find a bar graph of categorical data from a newspaper, a magazine, or the Internet.
   a) Is the graph clearly labeled?
   b) Does it violate the area principle?
   c) Does the accompanying article tell the W's of the variable?
   d) Do you think the article correctly interprets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

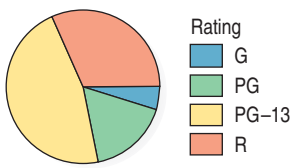
3. **Tables in the news.** Find a frequency table of categorical data from a newspaper, a magazine, or the Internet.
   a) Is it clearly labeled?
   b) Does it display percentages or counts?
   c) Does the accompanying article tell the W's of the variable?
   d) Do you think the article correctly interprets the data? Explain.

4. **Tables in the news II.** Find a contingency table of categorical data from a newspaper, a magazine, or the Internet.
   a) Is it clearly labeled?
   b) Does it display percentages or counts?
   c) Does the accompanying article tell the W's of the variables?
   d) Do you think the article correctly interprets the data? Explain.

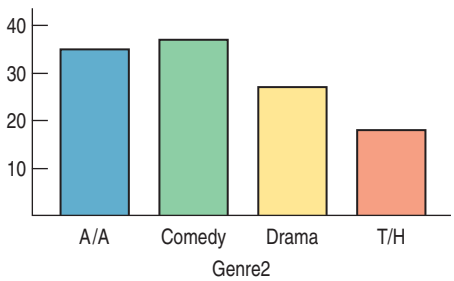
T 5. **Movie genres.** The pie chart summarizes the genres of 120 first-run movies released in 2005.
   a) Is this an appropriate display for the genres? Why/why not?
   b) Which genre was least common?



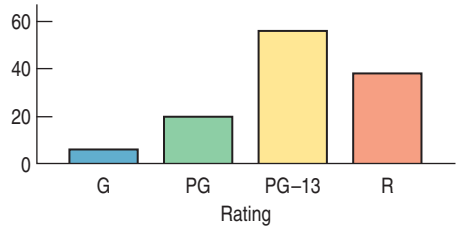
T 6. **Movie ratings.** The pie chart shows the ratings assigned to 120 first-run movies released in 2005.
   a) Is this an appropriate display for these data? Explain.
   b) Which was the most common rating?



T 7. **Genres again.** Here is a bar chart summarizing the 2005 movie genres, as seen in the pie chart in Exercise 5.
   a) Which genre was most common?
   b) Is it easier to see that in the pie chart or the bar chart? Explain.



T 8. **Ratings again.** Here is a bar chart summarizing the 2005 movie ratings, as seen in the pie chart in Exercise 6.
   a) Which was the least common rating?
   b) An editorial claimed that there's been a growth in PG-13 rated films that, according to the writer, "have too much sex and violence," at the expense of G-rated films that offer "good, clean fun." The writer offered the bar chart below as evidence to support his claim. Does the bar chart support his claim? Explain.



9. **Magnet schools.** An article in the Winter 2003 issue of *Chance* magazine reported on the Houston Independent School District's magnet schools programs. Of the 1755 qualified applicants, 931 were accepted, 298 were waitlisted, and 526 were turned away for lack of space. Find the relative frequency distribution of the decisions made, and write a sentence describing it.

10. **Magnet schools again.** The *Chance* article about the Houston magnet schools program described in Exercise 9 also indicated that 517 applicants were black or Hispanic, 292 Asian, and 946 white. Summarize the relative frequency distribution of ethnicity with a sentence or two (in the proper context, of course).

11. **Causes of death 2004.** The Centers for Disease Control and Prevention (www.cdc.gov) lists causes of death in the United States during 2004:

Cause of Death	Percent
Heart disease	27.2
Cancer	23.1
Circulatory diseases and stroke	6.3
Respiratory diseases	5.1
Accidents	4.7

   a) Is it reasonable to conclude that heart or respiratory diseases were the cause of approximately 33% of U.S. deaths in 2003?
   b) What percent of deaths were from causes not listed here?
   c) Create an appropriate display for these data.

12. **Plane crashes.** An investigation compiled information about recent nonmilitary plane crashes (www.planecrashinfo.com). The causes, to the extent that they could be determined, are summarized in the table.

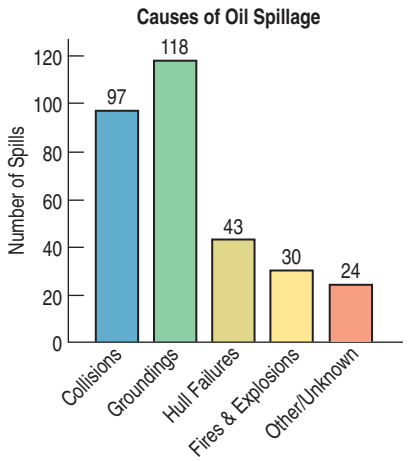
Cause	Percent
Pilot error	40
Other human error	5
Weather	6
Mechanical failure	14
Sabotage	6

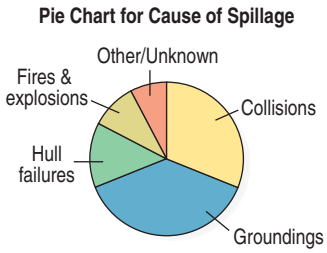
   a) Is it reasonable to conclude that the weather or mechanical failures caused only about 20% of recent plane crashes?
   b) In what percent of crashes were the causes not determined?
   c) Create an appropriate display for these data.

13. Oil spills 2006. Data from the International Tanker Owners Pollution Federation Limited (www.itopf.com) give the cause of spillage for 312 large oil tanker accidents from 1974–2006. Here are displays.

a) Write a brief report interpreting what the displays show.

b) Is a pie chart an appropriate display for these data? Why or why not?





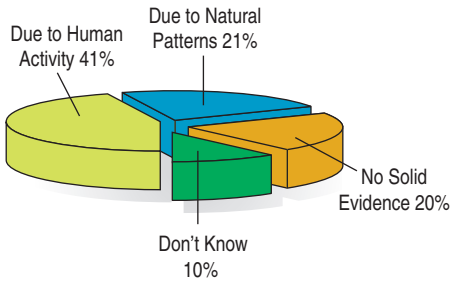
14. Winter Olympics 2006. Twenty-six countries won medals in the 2006 Winter Olympics. The table lists them, along with the total number of medals each won:

Country	Medals	Country	Medals
Germany	29	Finland	9
United States	25	Czech Republic	4
Canada	24	Estonia	3
Austria	23	Croatia	3
Russia	22	Australia	2
Norway	19	Poland	2
Sweden	14	Ukraine	2
Switzerland	14	Japan	1
South Korea	11	Belarus	1
Italy	11	Bulgaria	1
China	11	Great Britain	1
France	9	Slovakia	1
Netherlands	9	Latvia	1

a) Try to make a display of these data. What problems do you encounter?

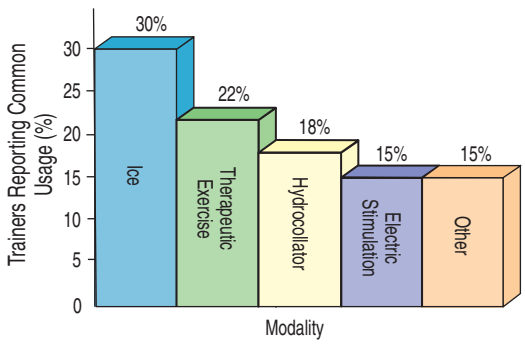
b) Can you find a way to organize the data so that the graph is more successful?

15. Global Warming. The Pew Research Center for the People and the Press (http://people-press.org) has asked a representative sample of U.S. adults about global warming, repeating the question over time. In January 2007, the responses reflected an increased belief that global warming is real and due to human activity. Here's a display of the percentages of respondents choosing each of the major alternatives offered:



List the errors in this display.

16. Modalities. A survey of athletic trainers (Scott F. Nadler, Michael Prybicien, Gerard A. Malanga, and Dan Sicher. "Complications from Therapeutic Modalities: Results of a National Survey of Athletic Trainers." *Archives of Physical Medical Rehabilitation* 84 [June 2003]) asked what modalities (treatment methods such as ice, whirlpool, ultrasound, or exercise) they commonly use to treat injuries. Respondents were each asked to list three modalities. The article included the following figure reporting the modalities used:



a) What problems do you see with the graph?

b) Consider the percentages for the named modalities. Do you see anything odd about them?

17. Teen smokers. The organization Monitoring the Future (www.monitoringthefuture.org) asked 2048 eighth graders who said they smoked cigarettes what brands they preferred. The table below shows brand preferences for two regions of the country. Write a few sentences describing the similarities and differences in brand preferences among eighth graders in the two regions listed.

Brand preference	South	West
Marlboro	58.4%	58.0%
Newport	22.5%	10.1%
Camel	3.3%	9.5%
Other (over 20 brands)	9.1%	9.5%
No usual brand	6.7%	12.9%

18. Handguns. In an effort to reduce the number of gun-related homicides, some cities have run buyback programs in which the police offer cash (often $50) to anyone who turns in an operating handgun. *Chance* magazine looked at results from a four-year period in Milwaukee. The table on the next page shows what types of guns were turned in and what types were used in homicides during a four-year period. Write a few sentences comparing the two distributions.

Caliber of gun	Buyback	Homicide
Small (.22, .25, .32)	76.4%	20.3%
Medium (.357, .38, 9 mm)	19.3%	54.7%
Large (.40, .44, .45)	2.1%	10.8%
Other	2.2%	14.2%

19. Movies by Genre and Rating. Here's a table that classifies movies released in 2005 by genre and MPAA rating:

	G	PG	PG-13	R	Total
Action/Adventure	66.7	25	30.4	23.7	**29.2**
Comedy	33.3	60.0	35.7	10.5	**31.7**
Drama	0	15.0	14.3	44.7	**23.3**
Thriller/Horror	0	0	19.6	21.1	**15.8**
Total	**100%**	**100%**	**100%**	**100%**	**100%**

a) The table gives column percents. How could you tell that from the table itself?
b) What percentage of these movies were comedies?
c) What percentage of the PG-rated movies were comedies?
d) Which of the following can you learn from this table? Give the answer if you can find it from the table.
   i) The percentage of PG-13 movies that were comedies
   ii) The percentage of dramas that were R-rated
   iii) The percentage of dramas that were G-rated
   iv) The percentage of 2005 movies that were PG-rated comedies

20. The Last Picture Show. Here's another table showing information about 120 movies released in 2005. This table gives percentages of the table total:

	G	PG	PG-13	R	Total
Action/Adventure	3.33%	4.17	14.2	7.50	**29.2**
Comedy	1.67	10	16.7	3.33	**31.7**
Drama	0	2.50	6.67	14.2	**23.3**
Thriller/Horror	0	0	9.17	6.67	**15.8**
Total	**5**	**16.7**	**46.7**	**31.7**	**100%**

a) How can you tell that this table holds table percentages (rather than row or column percentages)?
b) What was the most common genre/rating combination in 2005 movies?
c) How many of these movies were PG-rated comedies?
d) How many were G-rated?
e) An editorial about the movies noted, "More than three-quarters of the movies made today can be seen only by patrons 13 years old or older." Does this table support that assertion? Explain.

21. Seniors. Prior to graduation, a high school class was surveyed about its plans. The following table displays the results for white and minority students (the "Minority" group included African-American, Asian, Hispanic, and Native American students):

Seniors

Plans	White	Minority
4-year college	198	44
2-year college	36	6
Military	4	1
Employment	14	3
Other	16	3

a) What percent of the seniors are white?
b) What percent of the seniors are planning to attend a 2-year college?
c) What percent of the seniors are white and planning to attend a 2-year college?
d) What percent of the white seniors are planning to attend a 2-year college?
e) What percent of the seniors planning to attend a 2-year college are white?

22. Politics. Students in an Intro Stats course were asked to describe their politics as "Liberal," "Moderate," or "Conservative." Here are the results:

Politics

Sex	L	M	C	Total
Female	35	36	6	77
Male	50	44	21	115
Total	85	80	27	192

a) What percent of the class is male?
b) What percent of the class considers themselves to be "Conservative"?
c) What percent of the males in the class consider themselves to be "Conservative"?
d) What percent of all students in the class are males who consider themselves to be "Conservative"?

23. More about seniors. Look again at the table of post-graduation plans for the senior class in Exercise 21.
a) Find the conditional distributions (percentages) of plans for the white students.
b) Find the conditional distributions (percentages) of plans for the minority students.
c) Create a graph comparing the plans of white and minority students.
d) Do you see any important differences in the post-graduation plans of white and minority students? Write a brief summary of what these data show, including comparisons of conditional distributions.

24. Politics revisited. Look again at the table of political views for the Intro Stats students in Exercise 22.
a) Find the conditional distributions (percentages) of political views for the females.
b) Find the conditional distributions (percentages) of political views for the males.
c) Make a graphical display that compares the two distributions.
d) Do the variables *Politics* and *Sex* appear to be independent? Explain.

25. Magnet schools revisited. The *Chance* magazine article described in Exercise 9 further examined the impact of an applicant's ethnicity on the likelihood of admission to the Houston Independent School District's magnet schools programs. Those data are summarized in the table below:

Ethnicity	Admission Decision: Accepted	Wait-listed	Turned away	Total
Black/Hispanic	485	0	32	**517**
Asian	110	49	133	**292**
White	336	251	359	**946**
Total	**931**	**300**	**524**	**1755**

a) What percent of all applicants were Asian?
b) What percent of the students accepted were Asian?
c) What percent of Asians were accepted?
d) What percent of all students were accepted?

26. More politics. Look once more at the table summarizing the political views of Intro Stats students in Exercise 22.
a) Produce a graphical display comparing the conditional distributions of males and females among the three categories of politics.
b) Comment briefly on what you see from the display in a.

27. Back to school. Examine the table about ethnicity and acceptance for the Houston Independent School District's magnet schools program, shown in Exercise 25. Does it appear that the admissions decisions are made independent of the applicant's ethnicity? Explain.

28. Cars. A survey of autos parked in student and staff lots at a large university classified the brands by country of origin, as seen in the table.

Origin	Driver: Student	Staff
American	107	105
European	33	12
Asian	55	47

a) What percent of all the cars surveyed were foreign?
b) What percent of the American cars were owned by students?
c) What percent of the students owned American cars?
d) What is the marginal distribution of origin?
e) What are the conditional distributions of origin by driver classification?
f) Do you think that the origin of the car is independent of the type of driver? Explain.

29. Weather forecasts. Just how accurate are the weather forecasts we hear every day? The following table compares the daily forecast with a city's actual weather for a year:

Forecast	Actual Weather: Rain	No rain
Rain	27	63
No rain	7	268

a) On what percent of days did it actually rain?
b) On what percent of days was rain predicted?
c) What percent of the time was the forecast correct?
d) Do you see evidence of an association between the type of weather and the ability of forecasters to make an accurate prediction? Write a brief explanation, including an appropriate graph.

30. Twins. In 2000, the *Journal of the American Medical Association (JAMA)* published a study that examined pregnancies that resulted in the birth of twins. Births were classified as preterm with intervention (induced labor or cesarean), preterm without procedures, or term/post-term. Researchers also classified the pregnancies by the level of prenatal medical care the mother received (inadequate, adequate, or intensive). The data, from the years 1995–1997, are summarized in the table below. Figures are in thousands of births. (*JAMA* 284 [2000]:335–341)

Twin Births 1995–1997 (In Thousands)

Level of Prenatal Care	Preterm (induced or cesarean)	Preterm (without procedures)	Term or post-term	Total
Intensive	18	15	28	**61**
Adequate	46	43	65	**154**
Inadequate	12	13	38	**63**
Total	**76**	**71**	**131**	**278**

a) What percent of these mothers received inadequate medical care during their pregnancies?
b) What percent of all twin births were preterm?
c) Among the mothers who received inadequate medical care, what percent of the twin births were preterm?
d) Create an appropriate graph comparing the outcomes of these pregnancies by the level of medical care the mother received.
e) Write a few sentences describing the association between these two variables.

31. Blood pressure. A company held a blood pressure screening clinic for its employees. The results are summarized in the table below by age group and blood pressure level:

Blood Pressure	Age: Under 30	30–49	Over 50
Low	27	37	31
Normal	48	91	93
High	23	51	73

a) Find the marginal distribution of blood pressure level.
b) Find the conditional distribution of blood pressure level within each age group.
c) Compare these distributions with a segmented bar graph.
d) Write a brief description of the association between age and blood pressure among these employees.
e) Does this prove that people's blood pressure increases as they age? Explain.

32. Obesity and exercise. The Centers for Disease Control and Prevention (CDC) has estimated that 19.8% of Americans over 15 years old are obese. The CDC conducts a survey on obesity and various behaviors. Here is a table on self-reported exercise classified by body mass index (BMI):

		Body Mass Index		
		Normal (%)	Overweight (%)	Obese (%)
Physical Activity	Inactive	23.8	26.0	35.6
	Irregularly active	27.8	28.7	28.1
	Regular, not intense	31.6	31.1	27.2
	Regular, intense	16.8	14.2	9.1

a) Are these percentages column percentages, row percentages, or table percentages?
b) Use graphical displays to show different percentages of physical activities for the three BMI groups.
c) Do these data prove that lack of exercise causes obesity? Explain.

33. Anorexia. Hearing anecdotal reports that some patients undergoing treatment for the eating disorder anorexia seemed to be responding positively to the antidepressant Prozac, medical researchers conducted an experiment to investigate. They found 93 women being treated for anorexia who volunteered to participate. For one year, 49 randomly selected patients were treated with Prozac and the other 44 were given an inert substance called a placebo. At the end of the year, patients were diagnosed as healthy or relapsed, as summarized in the table:

	Prozac	Placebo	Total
Healthy	35	32	67
Relapse	14	12	26
Total	**49**	**44**	**93**

Do these results provide evidence that Prozac might be helpful in treating anorexia? Explain.

34. Antidepressants and bone fractures. For a period of five years, physicians at McGill University Health Center followed more than 5000 adults over the age of 50. The researchers were investigating whether people taking a certain class of antidepressants (SSRIs) might be at greater risk of bone fractures. Their observations are summarized in the table:

	Taking SSRI	No SSRI	Total
Experienced fractures	14	244	258
No fractures	123	4627	4750
Total	**137**	**4871**	**5008**

Do these results suggest there's an association between taking SSRI antidepressants and experiencing bone fractures? Explain.

35. Drivers' licenses 2005. The following table shows the number of licensed U.S. drivers by age and by sex (www.dot.gov):

Age	Male Drivers (number)	Female Drivers (number)	Total
19 and under	4,777,694	4,553,946	**9,331,640**
20–24	8,611,161	8,398,879	**17,010,040**
25–29	8,879,476	8,666,701	**17,546,177**
30–34	9,262,713	8,997,662	**18,260,375**
35–39	9,848,050	9,576,301	**19,424,351**
40–44	10,617,456	10,484,149	**21,101,605**
45–49	10,492,876	10,482,479	**20,975,355**
50–54	9,420,619	9,475,882	**18,896,501**
55–59	8,218,264	8,265,775	**16,484,039**
60–64	6,103,732	6,147,569	**12,251,361**
65–69	4,571,157	4,643,913	**9,215,070**
70–74	3,617,908	3,761,039	**7,378,947**
75–79	2,890,155	3,192,408	**6,082,563**
80–84	1,907,743	2,222,412	**4,130,155**
85 and over	1,170,817	1,406,271	**2,577,088**
Total	**100,389,881**	**100,275,386**	**200,665,267**

a) What percent of total drivers are under 20?
b) What percent of total drivers are male?
c) Write a few sentences comparing the number of male and female licensed drivers in each age group.
d) Do a driver's age and sex appear to be independent? Explain?

36. Tattoos. A study by the University of Texas Southwestern Medical Center examined 626 people to see if an increased risk of contracting hepatitis C was 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.

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

37. Hospitals. Most patients who undergo surgery make routine recoveries and are discharged as planned. Others suffer excessive bleeding, infection, or other postsurgical complications and have their discharges from the hospital delayed. Suppose your city has a large hospital and a small hospital, each performing major and minor surgeries. You collect data to see how many surgical patients have their discharges delayed by postsurgical complications, and you find the results shown in the following table.

	Discharge Delayed	
	Large hospital	**Small hospital**
Major surgery	120 of 800	10 of 50
Minor surgery	10 of 200	20 of 250

a) Overall, for what percent of patients was discharge delayed?
b) Were the percentages different for major and minor surgery?
c) Overall, what were the discharge delay rates at each hospital?
d) What were the delay rates at each hospital for each kind of surgery?
e) The small hospital advertises that it has a lower rate of postsurgical complications. Do you agree?
f) Explain, in your own words, why this confusion occurs.

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

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

a) Compare the two services' overall percentage of late deliveries.
b) On the basis of the results in part a, the company has decided to hire Pack Rats. Do you agree that Pack Rats delivers on time more often? Explain.
c) The results here are an instance of what phenomenon?

39. Graduate admissions. A 1975 article in the magazine *Science* examined the graduate admissions process at Berkeley for evidence of sex discrimination. The table below shows the number of applicants accepted to each of four graduate programs:

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

a) What percent of total applicants were admitted?
b) Overall, was 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?

40. Be a Simpson! Can you design a Simpson's paradox? Two companies are vying for a city's "Best Local Employer" award, to be given to the company most committed to hiring local residents. Although both employers hired 300 new people in the past year, Company A brags that it deserves the award because 70% of its new jobs went to local residents, compared to only 60% for Company B. Company B concedes that those percentages are correct, but points out that most of its new jobs were full-time, while most of Company A's were part-time. Not only that, says Company B, but a higher percentage of its full-time jobs went to local residents than did Company A's, and the same was true for part-time jobs. Thus, Company B argues, it's a better local employer than Company A.

Show how it's possible for Company B to fill a higher percentage of both full-time and part-time jobs with local residents, even though Company A hired more local residents overall.

JUST CHECKING Answers

1. 50.0%
2. 44.4%
3. 25.0%
4. 15.6% Blue, 56.3% Brown, 28.1% Green/Hazel/Other
5. 18.8% Blue, 62.5% Brown, 18.8% Green/Hazel/Other
6. 40% of the blue-eyed students are female, while 50% of all students are female.
7. Since blue-eyed students appear less likely to be female, it seems that *Sex* and *Eye Color* may not be independent. (But the numbers are small.)