Chapter 3 Describing Categorical Data

Mix and Match

In each case, unless noted, bar charts are better to emphasize counts whereas pie charts are better to communicate the relative share of the total amount.

1. Proportion of autos: pie chart is the most common; a bar chart or Pareto chart can also be used.
2. Types of defects: Pareto chart (a bar chart with the categories sorted in order of the most common defect)
3. Coupons: bar chart or Pareto chart (these are counts) or perhaps a table (only three values)
4. Hospital: bar chart or Pareto chart (counts) or pie chart (shares)
5. Destination: bar chart or Pareto chart (counts) or pie chart (shares)
6. Hanging up: Pareto chart
7. Excuses: Pareto chart
8. Brand of phone: bar chart (counts) or pie chart (shares)
9. Software: pie chart (shares) or perhaps a table (only three values)
10. Camera: bar chart (counts), pie chart (shares), or a table (only three values)
11. Ratings: Bar chart or table (only four values). Because the values are ordinal, avoid a pie chart.
12. Loans: Bar chart or table (only three values). Because the data is ordinal, it should not be put into a pie chart – even though the plot shows shares.

True/False

13. True, but only in general. For variables with few categories, a frequency table is often better, particularly when the analysis requires knowing the detailed frequencies.
14. False. The values have to be in one category for the variable to have no variation.
15. False. The frequency is the count of the items.
16. False. A relative frequency is a proportion.
17. True. It would be false if the variable were ordinal; you should not put the shares of an ordinal variable into a pie chart.
18. False. The proportion must match the relative frequency.
19. True.
20. False. It has fewer bars.
21. True.
22. False. The median only applies to ordinal variables and identifies the category of the middle value.

Think About It

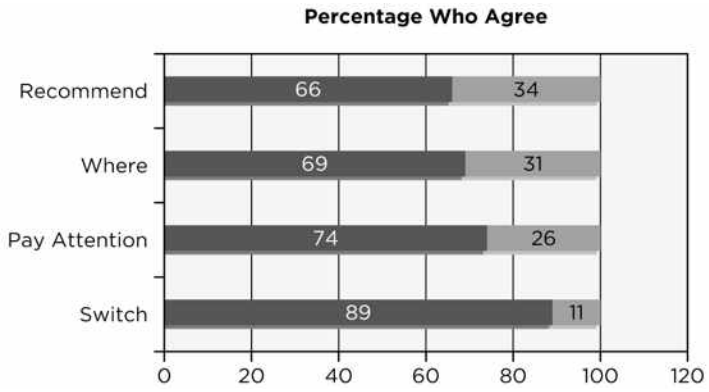
23. The message is that customers tend to stick with manufacturers from the same region. Someone trading in a domestic car tends to get another domestic car whereas someone who trades in an Asian car buys an Asian car. There's not a lot of switching of loyalties. The more subtle message, one that is disturbing to domestic car makers, is that those who own Asian cars are more loyal (78% buy another Asian car compared to 69% who stick with a domestic car). That makes it hard for domestic manufacturers to win back customers, even if they improve the quality of their cars.
24. The answer is yes. Since lighting makes up 37% of the use of electricity, reducing the demand for electricity by using more efficient bulbs can have a substantial impact. Compact fluorescent lamps produce the same amount of light with much less, say one-quarter, of the electricity used by an incandescent lamp. Less energy also

implies less heat and lower cooling costs. That said, the benefit of these savings for utilities is less pronounced because these savings happen mostly at night, not during the times of peak load that occur during the daytime.

25. This is a bar chart if you think about the underlying data as labeling the dollars held in these countries. The intent of the plot is to show the relative sizes of these counts, comparing the shares of U.S. debt held in these countries.

26. No, this is not a bar chart in the sense of this chapter. The chart uses bars to show a very short time series with two data points, the annual revenue in 2004 and 2005. The three other bars are projections of future revenue (the footnote indicates that the article was published in December of 2006). It's really a timeplot.

27. (a) No, these categories are not mutually exclusive. These percentages summarize four dichotomous variables, not one variable.
(b) Divided bars such as these might work well. This style is commonly used in reporting opinion poll results in the news. Sorting the values so that the percentages are in order also makes for a cleaner presentation.



28. (a) No. Each customer could report several of these items, so the categories are not mutually exclusive.
(b) A figure such as the divided bars used in Exercise 27 would be useful to illustrate the varying shares.

29. No. These percentages only list the percent of executives that report each problem and are not the relative frequencies of a categorical variable. The categories are not mutually exclusive; some of the executives listed several issues.

30. The percentages do not add to 100; we need an Other category (which has a 12% share of the market).

31. This variable is continuous and most of the values would be unique, producing a bar chart with close to 200 categories, one for each purchase amount. (Of course, there might be quite a few $0.99 coffee or soda purchases.)

32. This recoding converts the numerical values into an ordinal variable with three categories. The bar chart of the labels, ordered by purchase amount, would then reveal the most common purchase size (on this rounded scale).

33. The bar chart would have one very long bar (height 900) and five shorter bars of height 20 each. The plot would not be very useful, other than to show the predominance of one category.

34. A pie chart would devote 90% of its area to the main category and divide the remaining area into five small slices, each with equal area of 2%.

35. The bar chart would have five bars, each of the same height.

36. The bar chart. It would be hard to tell in the pie chart that the slices were of the same size.

37. The frequency table is simple enough to look at with only two numbers, particularly if it includes proportions.

38. With so many categories (the 51 states, including Washington D.C.), some aggregation by region might be useful. Alternatively, it might be good to highlight the most common states, and combine the rest together into a separate, Other category. A bar chart or pie chart could be used, and a frequency table would be fine if there were only a few states represented.

39. The mode is Public. There's no median for this chart; the data is not ordered.

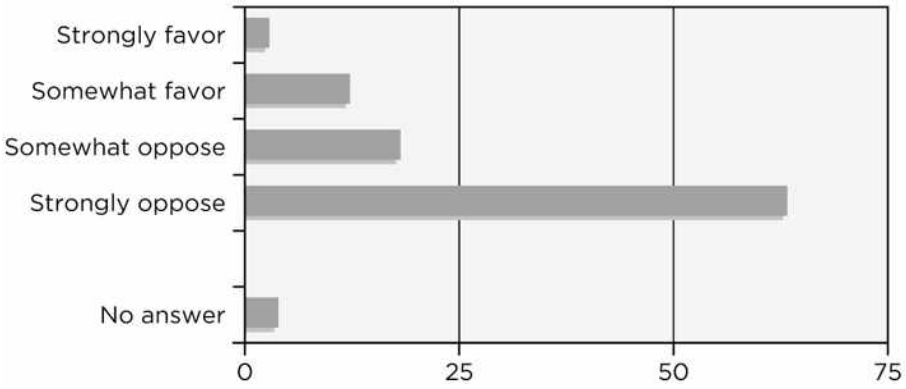
40. The East is the modal location. To find the median size, notice there are 50 sizes given, so the median is the size in position 25 or 26. Both lie in the category 10,000 to 19,999. Enrollment categories are ordinal and should not be shown in a pie chart.

41. The manufacturers want to know the modal preference because it identifies the most common color preference. Color preferences cannot be ordered, the median color preference is not defined.

42. A median rating of Excellent implies that at least one-half rated the service as Excellent. A modal rating of Excellent implies that this is the most common rating, but far fewer than half might have picked this rating.

43. This is ordinal data. Though not evidently a Likert scale (note the missing middle category), the responses are ordered. A pie chart would not be appropriate as it would conceal the ordering.

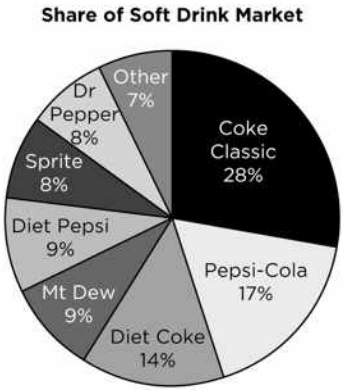
44. The key design element is to keep the categories in order rather than, say, show them alphabetically. Separating those who don't answer also distinguishes this category from the others. Here's an example.



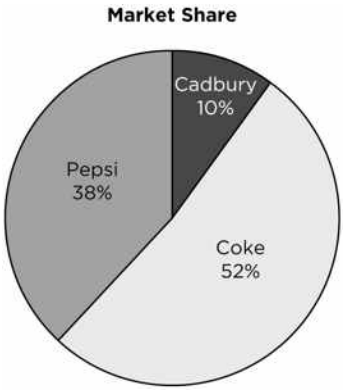
You Do It

45. (a) The underlying data table probably accumulates case sales by brand to some degree, such as the number sold by different retail chains and perhaps in different months. All we have here are the aggregated totals.

(b) To show the shares, a pie chart is most natural because it implies that we are dividing up the total amount. With all the brands shown, however, the small categories make it hard to show the labels. This version accumulates the 3 smallest brands into an "other" category.

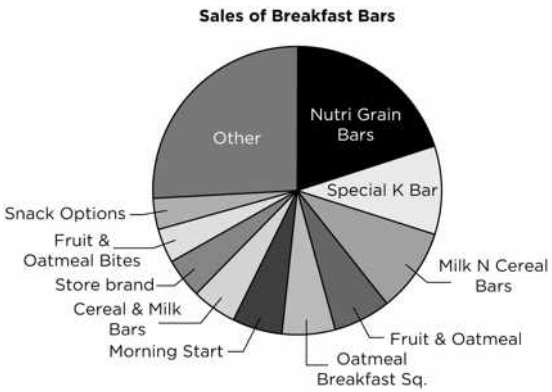


(c) Here is the pie chart, organized by company rather than brand. If we were artists, then a nicer way to prepare this chart would be to show at the same time the division of Pepsi into brands and so forth. The 38% share of Pepsi comes from Pepsi-Cola, Mt Dew, Diet Pepsi, and Sierra Mist. These divisions could then be used to slice up Pepsi's share of the chart.

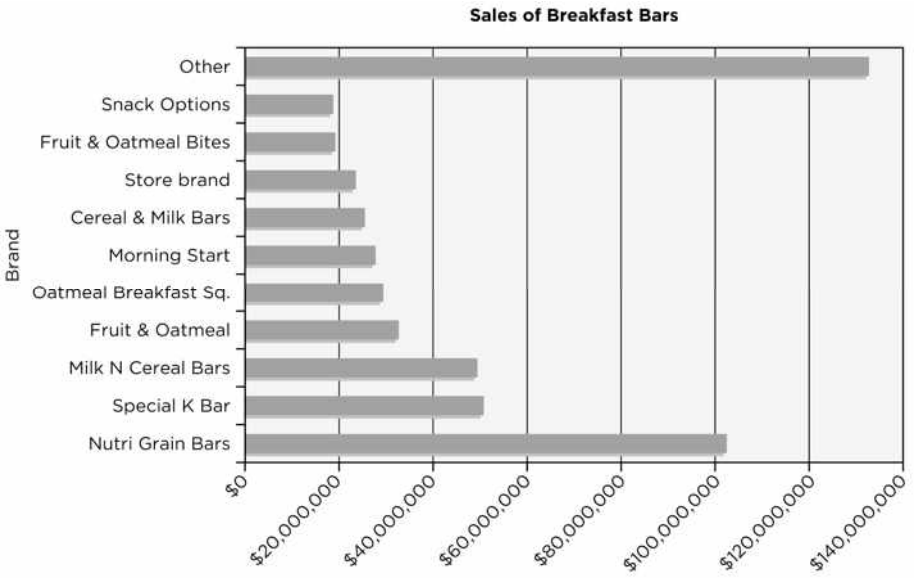


(d) To show the amounts sold, here's a bar chart with the two accumulated types.

46. (a) The underlying data would be organized by the company that sold the product, not by the boxes or bars themselves.

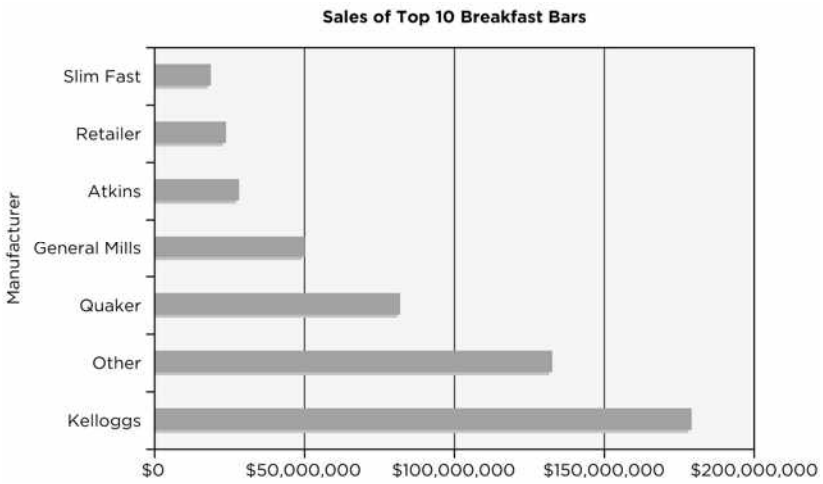


(b) The pie chart, including a category labeled Other for the rest of the market, is a bit cluttered by so many smaller brands. The chart makes the shares of the leading brands rather clear.

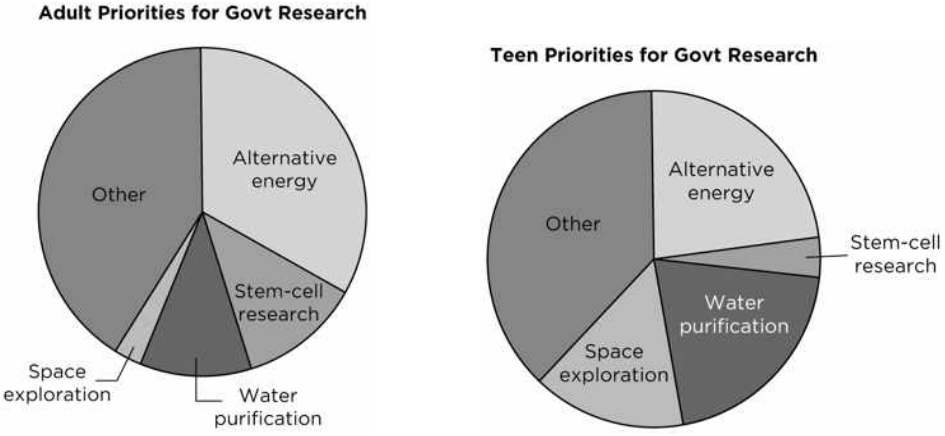


A bar chart might be more useful with so many categories.

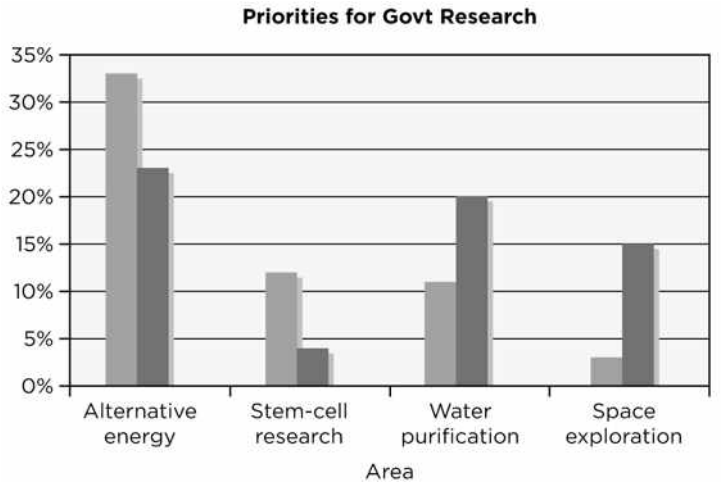
(c) This plot shows the sales of products grouped by the six named manufacturers rather than by brand.



47. (a) The Other category forms an additional row in the tables so that each column adds up to 100%. The addition of this extra row makes up a big part of both pie charts.



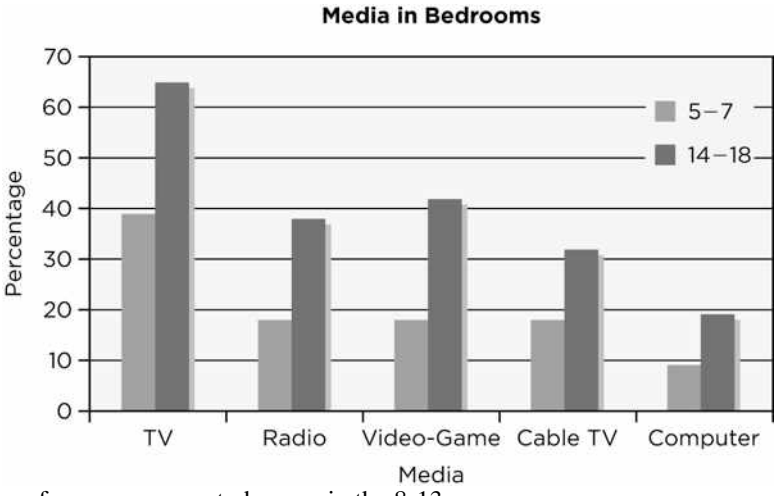
(b) The side-by-side bar chart works well for this. Notice that we no longer need the Other category that dominates the pie charts.



(c) No, because the categories would no longer partition the cases into distinct, non-overlapping subsets. A pie chart should only be used to summarize mutually exclusive groups.

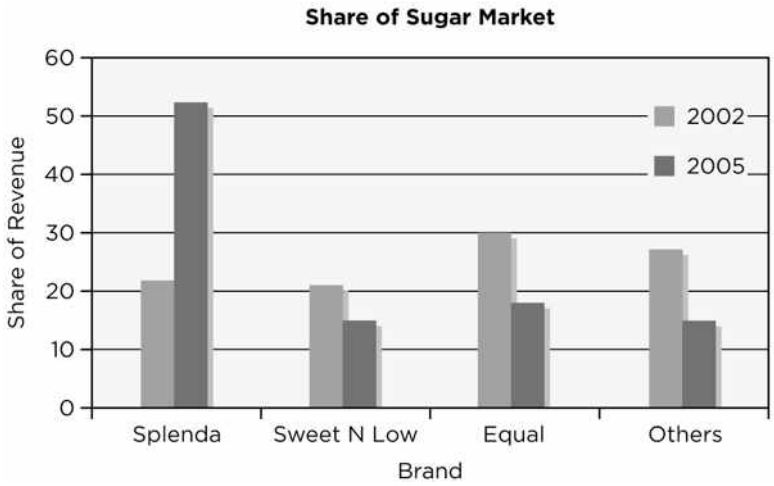
48. (a) The totals within a row do not sum to 100%. The columns give the proportion with different types of media, and these can sum to more than 100% as well. The represented categories do not divide the homes into different groups; a bedroom could have all of these media.

(b) The side-by-side bar chart shows more of every type of media in the rooms of older children.

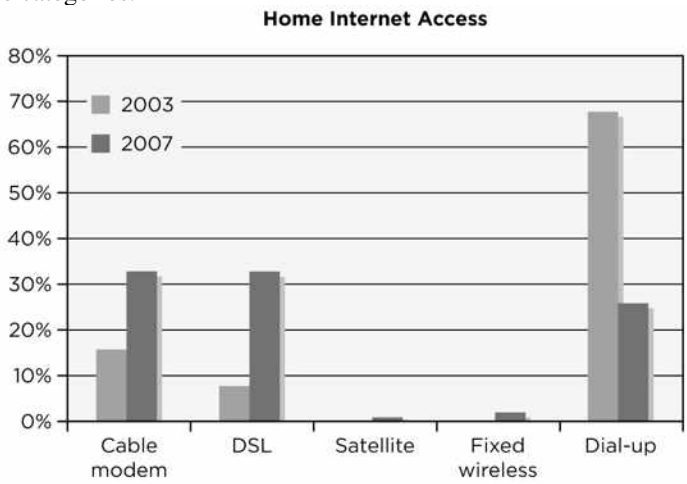


(c) The big adoption of games appears to happen in the 8-13 age range.

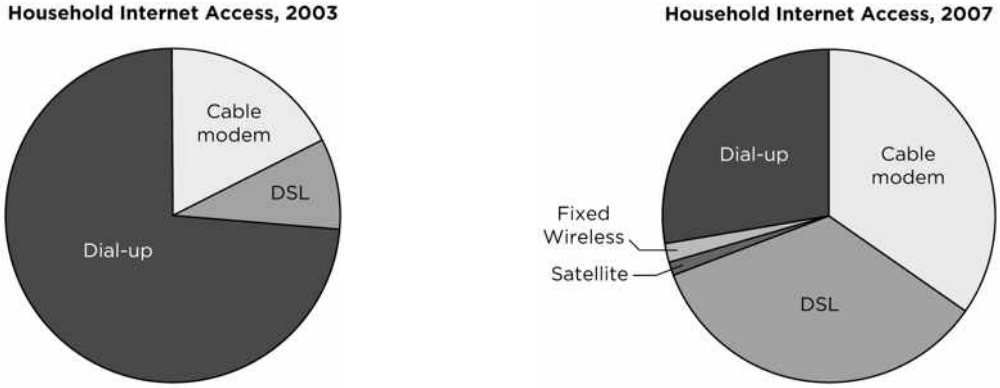
49. It's hard to beat this combined bar chart, though some would rather see two pie charts placed side by side. From the bar chart, you can quickly tell the shares each year by the colors, and the adjacent bars show that Splenda gained shares while the others fell.



50. (a) Side by side bars don't seem as useful in this example because of the very small shares given to satellite and fixed wireless, categories that were not present in 2003. The bars do make it easy to read off the shares, however. It might be better to omit the smaller categories or lump them as Other in order to focus the attention on the larger, more active categories.



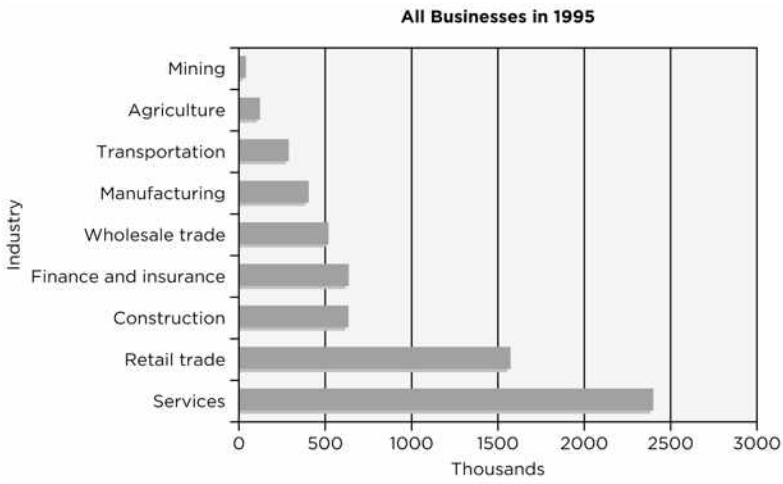
Pie charts put that emphasis more clearly on the larger categories, without leaving so much vacant space in the chart.



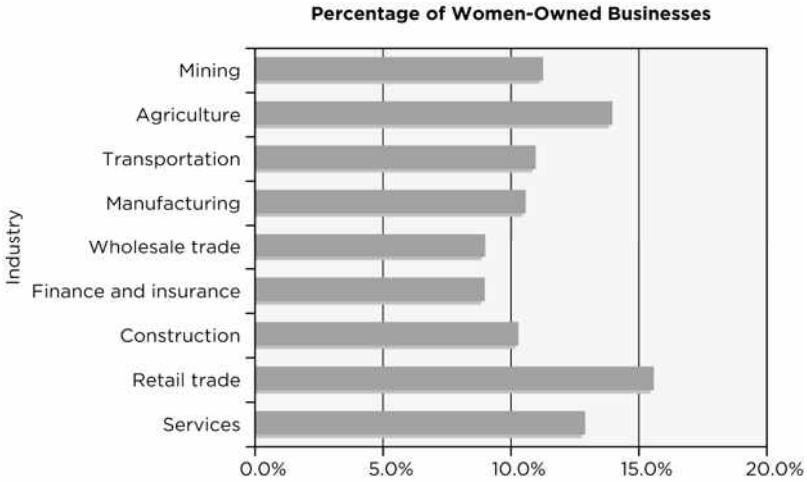
(b) No. These are percentages, not counts. If the number of households with Internet access grew enough during these years, 68% of a small number would be smaller than 26% of a much larger number. That's not the case, but remember that percentages conceal the total counts.

51.

(a) Visually, the distributions in the two bar charts seem similar. You should not use one bar chart because the scale required to show all businesses would hide the counts for the businesses owned by women.



(b) A pie chart shows which industries have the most women-owned businesses, but it does not show the percentage of women-owned businesses *within* each industry. It is necessary to form a column that shows the proportion of women-owned businesses within each industry. This plot shows the percentages within each industry. While some have more and some less, women own about 10-15% across the board, with none particularly standing out.



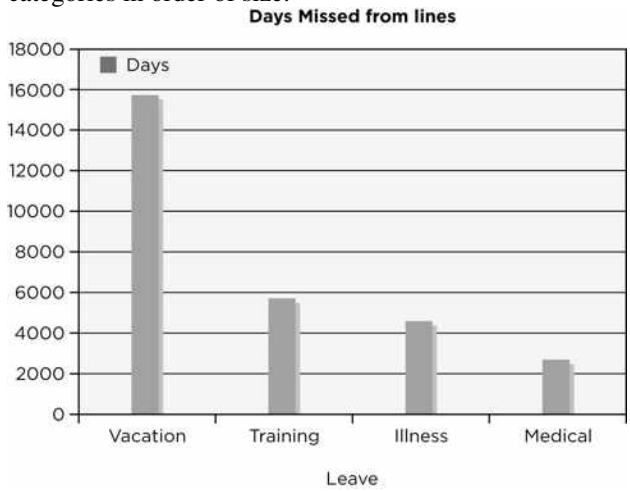
(c) Yes, but slight. The problem is that some industries might have grown or shrunk in number during the three intervening years. The broad nature of these categories suggests that any change was relatively small. If the data tracked more specialized industries, there might have been a larger proportional change over the three years.

52. (a) A pie chart is not appropriate because the categories are not mutually exclusive. A customer could purchase several types of gift cards.
(b) The card would more likely be for a product, but unless we know more about these data (such as the age or nature of the customers) it would be hard to use them to predict the type of card an individual would receive.
(c) With only three numbers, a table is fine. A chart that disguised the particular values would be less informative.

53. (a) Use a table with the two rows and the percentages (or proportions)

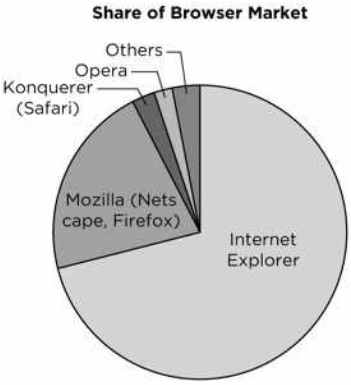
Unexpected illness	4,463	15.8%
Planned leave	23,735	84.2%

(b) A Pareto chart shows the categories in order of size.



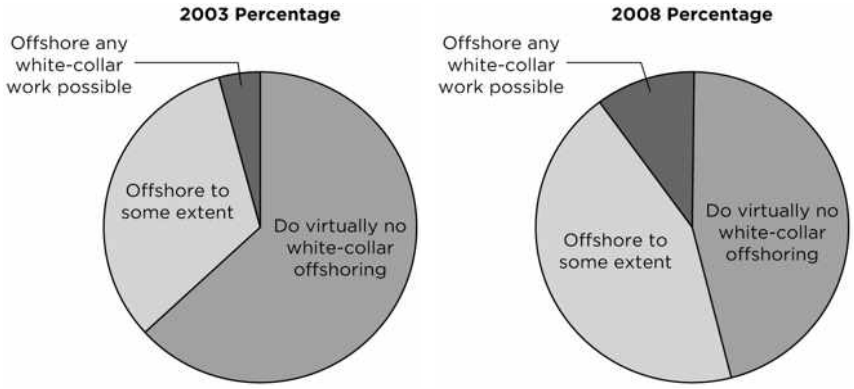
54. (a) Reasonable questions include asking which "popular" websites were monitored and precisely how the type of Web browser was recognized. Not every browser identifies itself, and Web crawlers used to support search engines like Google visit sites masquerading as browsers. Another important question would be to know how many different users were out there rather than just how many times the same user came to these websites. Finally, you might also want to know the total count used to form these percentages.

(b) The pie chart shows clearly that Internet Explorer dominates this market, but Netscape remains its biggest competitor.

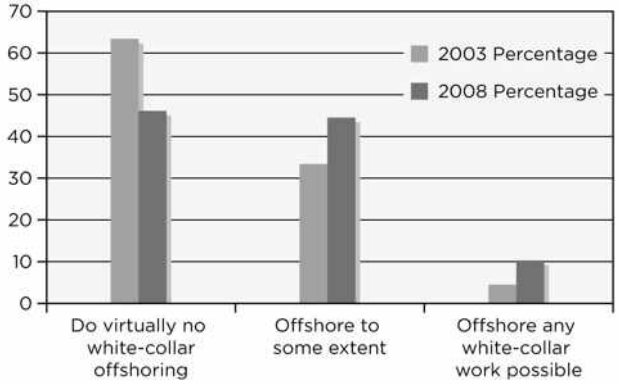


(c) The modal choice used to be Netscape, but has become Internet Explorer.

55. (a) Yes, pie charts are fine because the responses are mutually exclusive and sum to 100%.



(b) Various answers are possible. The following layout is reasonable.

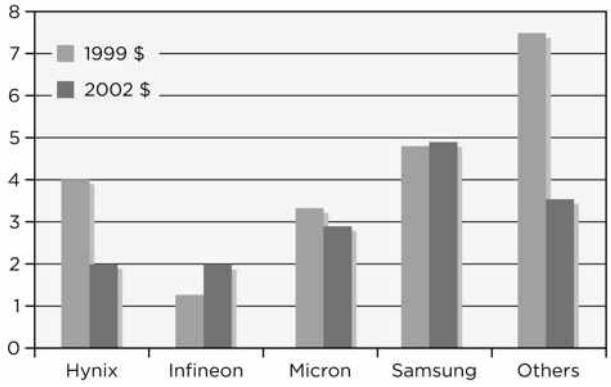


(c) The bar chart facilitates comparison. The pie chart makes the relative shares more apparent. For example, the 2003 pie shows a predominant share for taking no action, the only choice anticipated to fall in 2008.

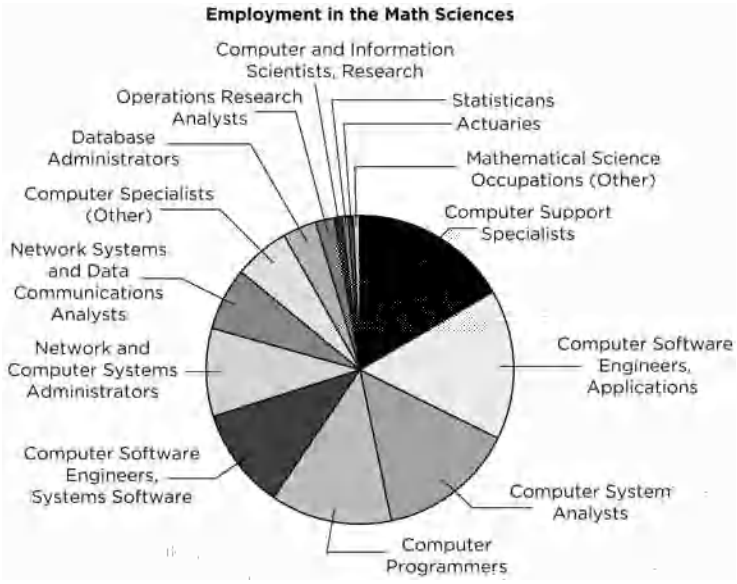
(d) The mode and median agree (virtually none) in 2003, but differ in 2008 as responses shift from the more consistent response to a tendency to do more off shoring.

56. (a) To make the area proportional to the amount of sales, the diameter of the pie for 2002 would have to be smaller by a factor of $\sqrt{15.25/20.8} \approx 0.857$. As shown, the areas of the two are the same (the original article got it right).

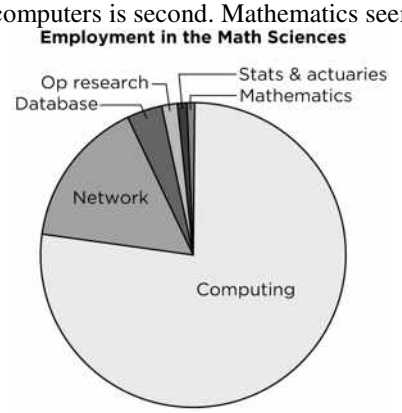
(b) Micron's sales went down, even though its share increased.



57. (a) The breakdown of employment into so many categories hides the dominance of computer-related occupations; you have to look at the labels to see that all of the top categories are related to computing. Statisticians have just about fallen off of the visible portion of the list shown (by Excel) next to the pie chart

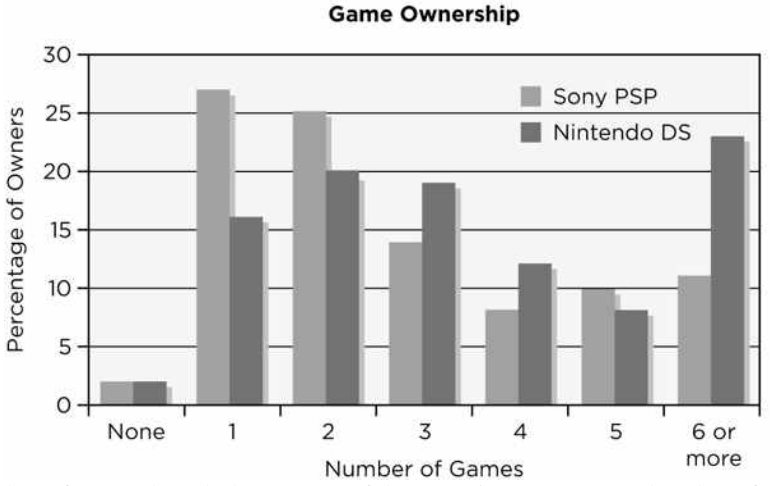


(b) The most common of these categories is computer support specialists.
(c) It makes sense to combine some of the categories. Here's one suggestion that bundles all of the computer titles into one category. Networking computers is second. Mathematics seems smaller than ever.



58. (a) The columns define a collection of mutually exclusive categories that add to 100% so a pie chart of either would be appropriate.

(b) Multiply the percentages times the number of owners of each system, and divide by 100 to get a count of the number who own the various numbers of games.
(c) Because there are more who own the Nintendo than the Sony, use percentages.



(d) The median number of games bought by owners of the Sony is two compared to three for Nintendo. You can tell that the median for the Sony is two because 26% of Sony owners have one and 25% have two. These two categories contain more than half of the people, so the middle person in the sorted list would have two games.
(e) Comparing the heights of the bars for each number of games, the bar for Sony owners is higher at the left (few games), whereas the Nintendo bar is higher at the right (more games). Not only has Nintendo sold about twice as many units (among these respondents), but each respondent has tended to buy more games.

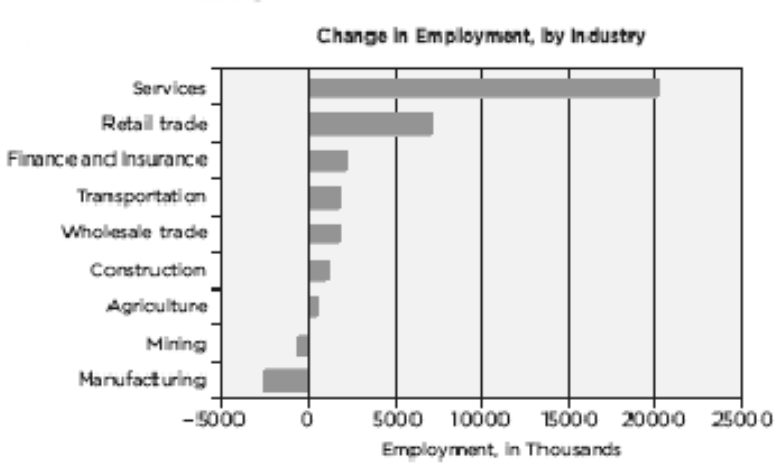
4M Growth Industries

(a) It could use trends suggested by the table to indicate how to shift its sales force from declining industries to those that appear stronger and growing.

(b) A bar chart would show the counts and make it simpler to compare the counts within a year. A pie chart would emphasize the shares among industries in the two years.

(c) By looking at the changes from 1980 to 1997, you can see which industries are growing and which are shrinking. You could also use a side-by-side chart, but a chart of the differences does the subtraction for us.

(d) This bar chart shows the change in employment. This chart shows industries in order, from those with the greatest increase down to those with the largest decline.



(e) The negative bars identify industries that employ fewer workers. One can also see the growth in side-by-side bar charts, but the comparison is less immediate.

(f) Show all nine. There are not enough categories to have an advantage in collapsing the number of categories.

(g) The growth in service-type industries (called service, but retail is a service in a way as well), the decline in basic manufacturing and mining.

(h) This chart hides the size of the changes relative to the size of the industry. The fall off in manufacturing looks really large compared to that in mining, but manufacturing employs far more workers than mining. The drop-off in mining, proportionally, is much larger than the drop-off in manufacturing (12% drop in manufacturing versus 41% drop in mining).

Other charts should be used to contrast shares. One could use a bar chart (though these make it harder to recognize the shares) or two pie charts. The layout below shows that manufacturing once employed more than one-quarter of the workforce but has become much smaller. The service industry grew to fill the void.