Chapter 2

Organizing and Summarizing Data

2.1 Organizing Qualitative Data

1. Raw data are the data as originally collected, before they have been organized or coded.

2. number (or count); proportion (or percent)

3. It is a good idea to add up the frequencies in a frequency distribution as a check to see if you missed data or possibly counted the same data more than once. If the total of the frequencies does not equal the total number of data values, the distribution should be done again.

4. The relative frequencies should add to 1 (though deviations may occur due to rounding error).

5. A Pareto chart is a bar chart with bars drawn in order of decreasing frequency or relative frequency.

6. Relative frequencies should be used when sample or population sizes are different because comparisons of frequencies in such cases can be difficult or misleading. When comparing samples with different sample sizes, the sample sizes should also be reported because the sample size affects the precision of the results.

7. Arrange the ordinal data from least to greatest on either the horizontal or vertical axis. For example, if the possible data values are "poor", "fair", "good", or "excellent", then arrange the data in that order. The order of the data in a pie chart cannot be made apparent because pie charts don't have a natural starting point.

8. The data as reported could not be made into a pie chart because the percentages do not add to 100%. They add to more than 100% which implies that some people selected more than one category as a reason for not losing weight.

9. (a) The largest segment in the pie chart is for "Washing your hands" so the most commonly used approach to beat the flu bug is washing your hands. 61% of respondents selected this as their primary method for beating the flu.

(b) The smallest segment in the pie chart is for "Drinking Orange Juice" so the least used method is drinking orange juice. 2% of respondents selected this as their primary method for beating the flu.

(c) 25% of respondents felt that flu shots were the best way to beat the flu.

10. (a) $\frac{146,000}{1,900,000} = \frac{73}{950} \approx 0.077$; approximately 7.7% of cosmetic surgeries in 2006 were for tummy tucks.

(b) $\frac{307,000}{1,900,000} \approx 0.162$; approximately 16.2% of cosmetic surgeries in 2006 were for nose reshaping.

(c) The graph accounts for 329,000 + 307,000 + 303,000 + 233,000 + 146,000 = 1,318,000 surgeries. Thus, 1,900,000 − 1,318,000 = 582,000 surgeries are not accounted for in the graph.

11. (a) The largest bar corresponds to the United States, so the U.S. had the most internet users in 2007.

(b) The bar for Germany appears to reach the line for 50. Thus, we estimate that there were 50 million internet users in Germany in 2007.

(c) The bar for China appears to reach 160 on the vertical axis. Since, $160 - 50 = 110$, we estimate that there were about 110 million more internet users in China than in Germany during 2007.

12. (a) 22,657,000 whites were living in poverty.

(b) $\frac{9293}{22657 + 8969 + 9293 + 6138}$
$= \frac{9293}{47057} \approx 0.197$ or 19.7%
In 2006, about 19.7% of the impoverished were Hispanic.

(c) The graph does not account for the different population size of each ethnic group. Without knowing the population sizes, we cannot determine whether a group is disproportionally impoverished. For example, there were an estimated 44.3 million, or 14.8%, Hispanics in the U.S. in 2006.

13. (a) 10.5% of game console owners were planning to buy a replacement device within the next 12 months.

(b) $250(0.149) = 37.25$

37.25 million cell phone owners planned to buy a replacement phone within the next 12 months.

(c) No; the bars are not arranged in decreasing order.

(d) No; in a relative frequency bar chart, the percents refer to the whole and sum to 1 (or 100%). In this chart, it is possible for someone to own several (if not all) of the listed devices so the percents refer to each category individually.

14. (a) Approximately 7% of identify fraud victims were victimized by someone who obtained their information over the Internet.

(b) From the graph, roughly 20% of victims were victimized by service employees.
$8.9(0.2) = 1.78$

Thus, 1.78 million identify fraud victims in 2006 were victimized by service employees.

(c) No, the percents add up to less than 100%. Add another category such as "Someone Else" to cover the remaining cases.

15. (a) In 2000, 55% of parents felt the Internet was a good thing. In 2006, 59% of parents felt this way.

(b) Good Thing: $59 - 67 = -8$
No Effect: $30 - 25 = 5$
Bad Thing: $7 - 5 = 2$
The opinion "No Effect Either Way" saw the greatest increase from 2004 to 2006.

(c) No; the decrease of 8% for "Good Thing" corresponded to a 5% increase for "Bad Thing". The difference is accounted for by an increase in the "No Effect" category.

(d) The percentages within each year may not add to 100% because some participants may not have answered the question (nonresponse). In addition, rounding error could lead to small discrepancies.

16. (a) $7191 - 5079 = 2112$
There were 2112 more engineering doctorates awarded in 2006 than in 2002.

(b) The total number of doctorates awarded was 4682 + 7538 + 10443 + 7191 + 6124 + 9618 = 45,596

$\frac{4682}{45,596} \approx 0.103$; $\frac{6124}{45,596} \approx 0.134$

In 2006, approximately 10.3% of doctorates awarded were in physical sciences and 13.4% in education.

(c) The field of education saw a consistent decrease in number of doctorates awarded while the number of doctorates in the professional/other category saw a consistent increase.

(d) Physical Science: $4682 - 4565 = 117$
Social Science: $7538 - 7728 = -190$
Other Science: $10443 - 9061 = 1382$
Engineering: $7191 - 5921 = 1270$
Education: $6124 - 6569 = -445$
Professional/Other: $9618 - 8793 = 825$
"Other Sciences" saw the greatest increase in the number of doctorates awarded over the 3 years of the survey.

(e) Physical Science: $\frac{117}{4565} \approx 0.026$

Social Science: $\frac{-190}{7728} \approx -0.025$

Other Science: $\frac{1382}{9061} \approx 0.153$

Engineering: $\frac{1270}{5921} \approx 0.214$

Education: $\frac{-445}{6569} \approx -0.068$

Professional/Other: $\frac{825}{8793} \approx 0.094$

"Engineering" saw the largest percent increase in doctorates awarded between 1998 and 2006.

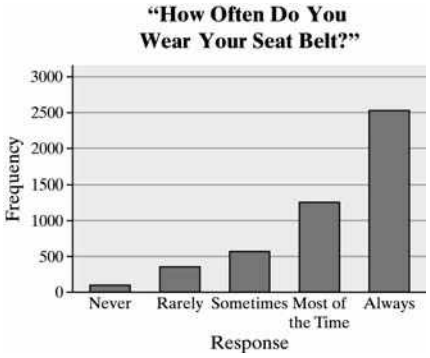
17. **(a)** Total students surveyed = 125 + 324 + 552 + 1257 + 2518 = 4776
Relative frequency of "Never"
$= 125/4776 \approx 0.0262$ and so on.

Response	Relative Frequency
Never	0.0262
Rarely	0.0678
Sometimes	0.1156
Most of the time	0.2632
Always	0.5272

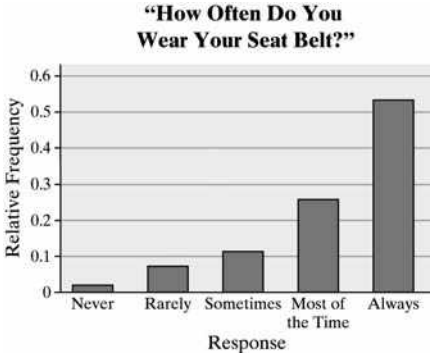
(b) 52.72%

(c) $0.0262 + 0.0678 = 0.0940$ or 9.40%

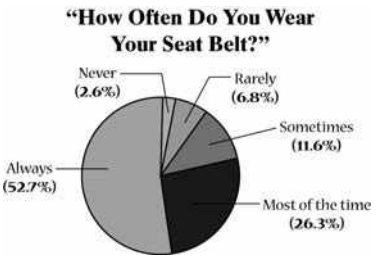
(d)



(e)



(f)



(g) The statement is inferential since it is inferring something about the entire population based on the results of a sample survey.

18. **(a)** Total students surveyed = 249 + 118 + 249 + 345 + 716 + 3093 = 4770
Relative frequency of " I do not drive"
$= \frac{249}{4770} \approx 0.0522$ and so on.

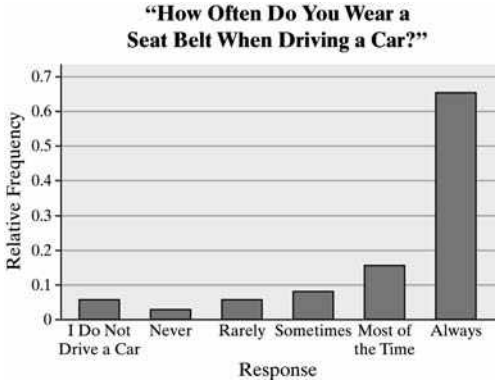
Response	Relative Frequency
I do not drive	0.0522
Never	0.0247
Rarely	0.0522
Sometimes	0.0723
Most of the time	0.1501
Always	0.6484

(b) 64.84%

(c) $0.0247 + 0.0522 = 0.0769$ or 7.69%

(d)

(e)



(f)

(g) Total students = 118 + 249 + 345 + 716 + 3093 = 44521
Relative frequency of "Never"
$= \frac{118}{4521} \approx 0.0261$ and so on.

Response	Relative Frequency
Never	0.0261
Rarely	0.0551
Sometimes	0.0763
Most of the time	0.1584
Always	0.6841

The relative frequencies of all categories are very similar except that students are more likely to wear their seatbelt 'Always' when driving than when riding in a car driven by another.

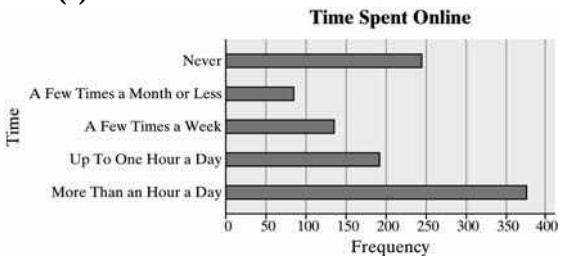
(h) The statement is descriptive because it is describing the particular sample.

19. (a) Relative frequency of "More than 1 hour a day" = $377/1025 \approx 0.3678$ and so on.

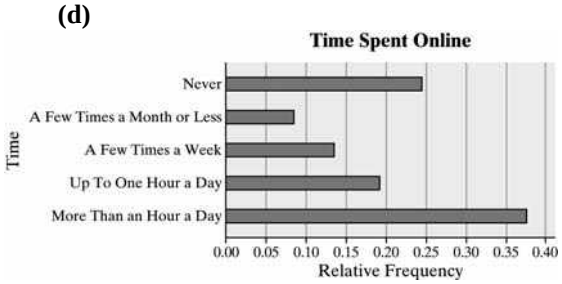
Response	Relative Frequency
More than 1 hr a day	0.3678
Up to 1 hr a day	0.1873
A few times a week	0.1288
A few times a month or less	0.0790
Never	0.2371

(b) 0.2371 (about 24%)

(c)



(d)



(e)

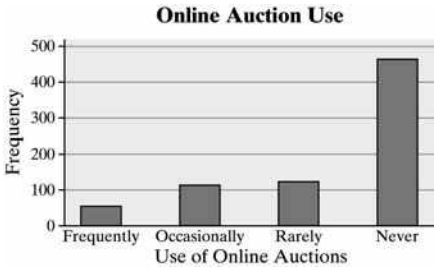
(f) The statement provides an estimate, but no level of confidence is given.

20. (a) Relative frequency of "Frequently" $= 54/770 \approx 0.0701$ and so on.

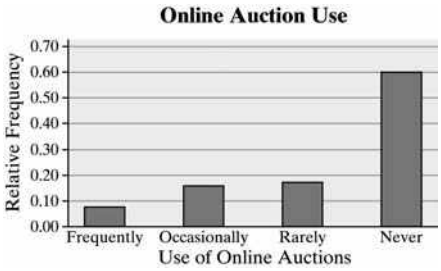
Response	Relative Frequency
Frequently	0.0701
Occasionally	0.1597
Rarely	0.1701
Never	0.6000

(b) The proportion surveyed who never use online auctions is 0.6 (or 60%).

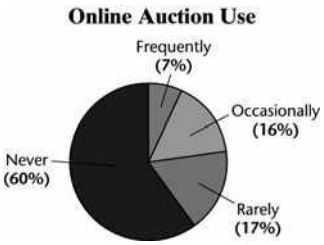
(c)



(d)



(e)



(f) The population being studied is adult Internet users.

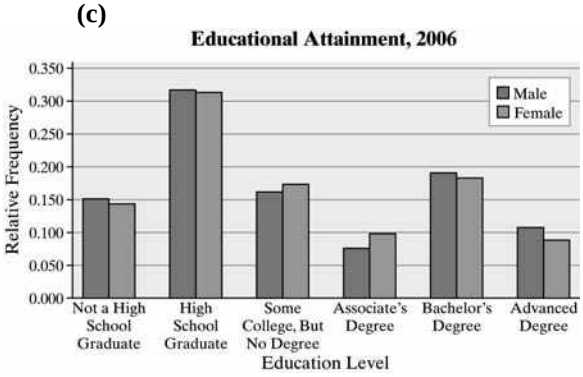
21. (a) Total males = 92.2 (million)
Relative frequency for "Not HS graduate" $= \frac{13.8}{92.2} \approx 0.1497$ and so on.

Educational Attainment	Males Rel. Freq.
Not a high school graduate	0.1497
High school graduate	0.3189
Some college, but no degree	0.1627
Associate's degree	0.0770
Bachelor's degree	0.1855
Advanced degree	0.1063

(b) Total females = 99.6 (million)
Relative frequency for "Not HS graduate" $= \frac{14.1}{99.6} \approx 0.1416$ and so on.

Educational Attainment	Females Rel. Freq.
Not a high school graduate	0.1416
High school graduate	0.3163
Some college, but no degree	0.1767
Associate's degree	0.0964
Bachelor's degree	0.1817
Advanced degree	0.0873

(c)



(d) Answers will vary. The percentages are fairly close for not graduating high school and just getting a high school diploma. It appears that men are more likely than women to complete bachelors degrees or higher while women are more likely than men to stop at an Associate's degree or after only a little college.

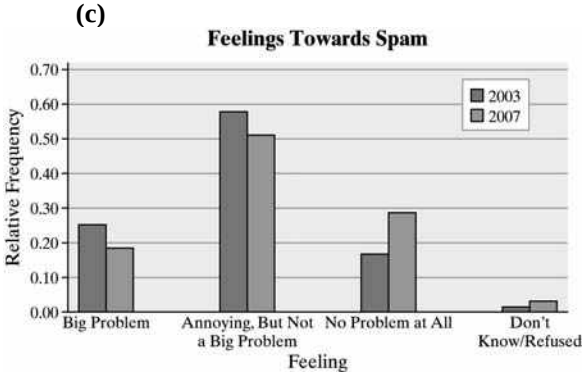
22. (a) Total surveyed in 2003= 373 + 850 + 239 + 15 = 1477
Relative frequency for "Big Problem" = $\frac{373}{1477} \approx 0.2525$ and so on.

Feeling	2003 Rel. Freq.
Big Problem	0.2525
Annoying, not a big problem	0.5755
No problem at all	0.1618
Don't know/Refused	0.0102

(b) Total surveyed in 2007 = 269 + 761 + 418 + 45 = 1493
Relative frequency for "Big Problem" = $\frac{269}{1493} \approx 0.1802$ and so on.

Feeling	2007 Rel. Freq.
Big Problem	0.1802
Annoying, not a big problem	0.5097
No problem at all	0.2800
Don't know/Refused	0.0301

(c)



(d) Answers will vary; the percentage who feel spam is a problem seems to have decreased from 2003 to 2007. Internet users may be more aware of spam and are better able to deal with it using email filters.

23. (a) Total male victims = 791 + 3762 + 3220 + 2977 + 860 = 11,610
Relative frequency for "Less than 17" = $\frac{791}{11,610} \approx 0.0681$

Age	Relative Frequency
Less than 17	0.0681
17-24	0.3240
25-34	0.2773
35-54	0.2564
55 or older	0.0741

(b) Total female victims = 373 + 550 + 599 + 1102 + 465 = 3089
Relative frequency for "Less than 17" = $\frac{373}{3089} \approx 0.1208$

Age	Relative Frequency
Less than 17	0.1208
17-24	0.1781
25-34	0.1939
35-54	0.3567
55 or older	0.1505

(c)

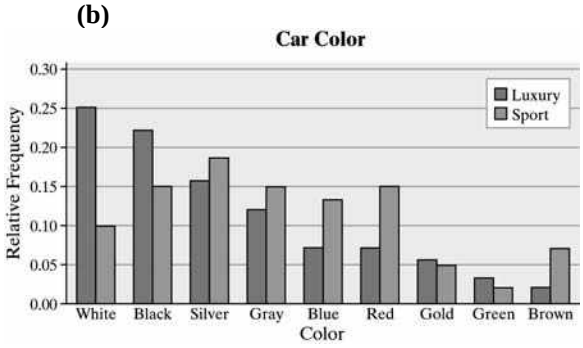
(d) Answers will vary. Children and older victims are more likely to be female. Young adult victims are more likely to be male.

24. (a) Relative frequency for "White" luxury cars = $\frac{25}{100} = 0.25$ and so on.
Relative frequency for "Silver" sport cars = $\frac{10}{100} = 0.10$ and so on.

Relative Frequencies

Color	Luxury Cars	Sport Cars
White	0.25	0.10
Black	0.22	0.15
Silver	0.16	0.18
Bray	0.12	0.15
Blue	0.07	0.13
Red	0.07	0.15
Gold	0.06	0.05
Green	0.03	0.02
Brown	0.02	0.07

(b)

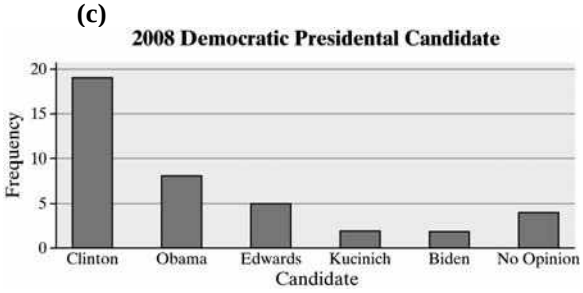


(c) Answers will vary. White is the most popular color for luxury cars, while silver is the most popular for sports cars.

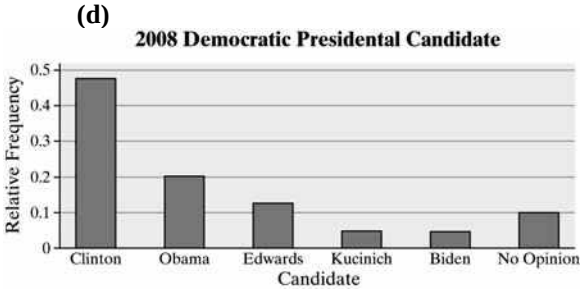
25. (a), (b)
Total number of voters polled = 40
Relative frequency for Clinton voters = $\frac{19}{40} = 0.475$ and so on.

Candidate	Frequency	Rel. Freq.
Clinton	19	0.475
Obama	8	0.2
Edwards	5	0.125
Kucinich	2	0.05
Biden	2	0.05
No Opinion	4	0.1

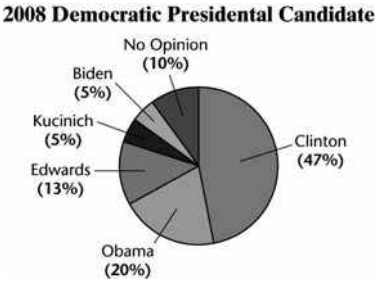
(c)



(d)



(e)



(f) The data indicates that Clinton would win the nomination. The conjecture would be inferential since a conclusion about the entire country is being made from sample data. Your confidence would increase with a larger sample since you have a larger portion of the voters in the country. Note that Obama ultimately won the nomination and went on to win the presidential election with Biden as his running mate.

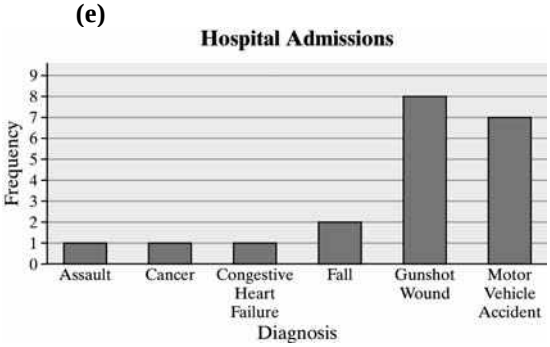
26. **(a)**, **(b)**Total admissions sampled = 20
Relative frequency of "Assault" = $\frac{1}{20} = 0.05$ and so on.

Diagnosis	Frequency	Rel. Freq.
Assault	1	0.05
Cancer	1	0.05
Heart Failure	1	0.05
Fall	2	0.10
Gunshot	8	0.40
M.V. accident	7	0.35

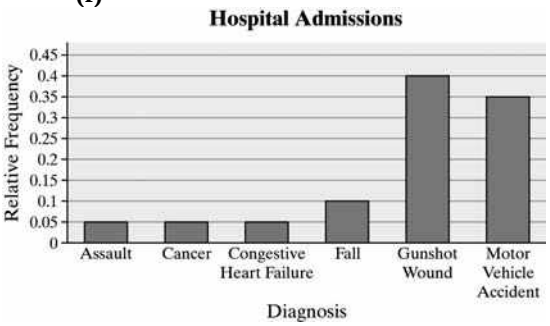
(c) Most admissions were for gunshot wounds.

(d) The relative frequency is 0.35, so the percentage is 35%. That is 35% of the diagnoses were motor vehicle accidents.

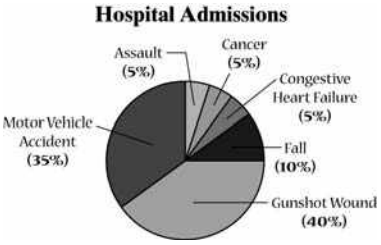
(e)



(f)



(g)



(h) The statement would be inferential because a conclusion about all admissions (a population) is being made based on a sample of admissions.

27. (a), (b)

Total number of players surveyed = 25

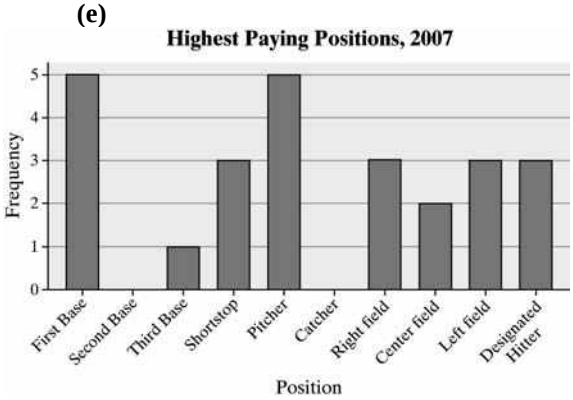
Relative frequency for "First Base" = $\frac{5}{25} = 0.20$ and so on.

Position	Freq.	Rel. Freq.
First Base	5	0.20
Second Base	0	0.00
Third Base	1	0.04
Shortstop	3	0.12
Pitcher	5	0.20
Catcher	0	0.00
Right Field	3	0.12
Center Field	2	0.08
Left Field	3	0.12
Designated Hitter	3	0.12

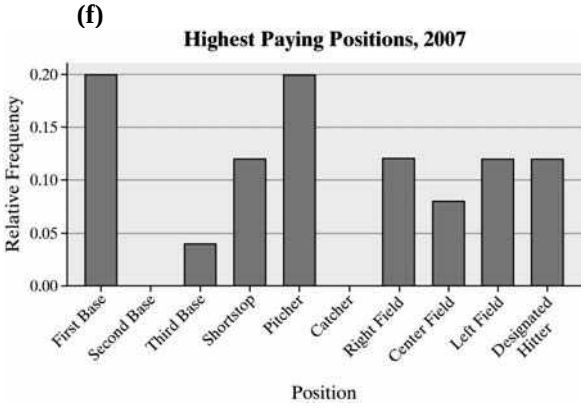
(c) Pitchers and first basemen appear to be more lucrative since they make up the highest percent of the top 25 highest paid players.

(d) Second base and catcher might be avoided since none of the top 25 highest paid players played either position.

(e)



(f)



(g)

28. (a), (b)

Total number of patients = 50

Relative frequency for "Type A" = $\frac{18}{50} = 0.36$ and so on.

Blood Type	Freq.	Rel. Freq.
A	18	0.36
AB	4	0.08
B	6	0.12
O	22	0.44

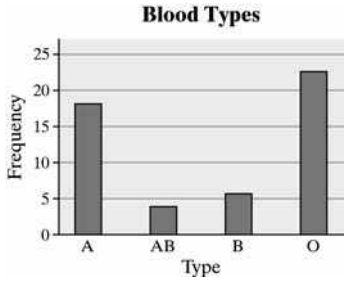
(c) Type O is the most common.

(d) Type AB is the least common.

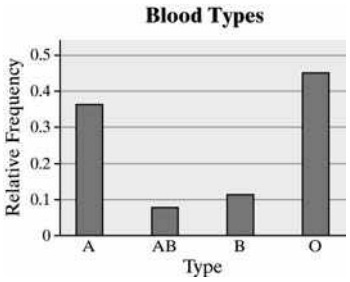
(e) We estimate that 44% of the population has type O blood. This is considered inferential statistics because a conclusion about the population is being drawn based on sample data.

(f) Answers will vary; in 2008 the Red Cross reported that 45% of the population had type O blood (either + or –). Results will differ because of sampling variability.

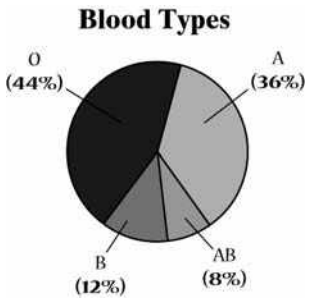
(g)



(h)



(i)



29. (a), (b)

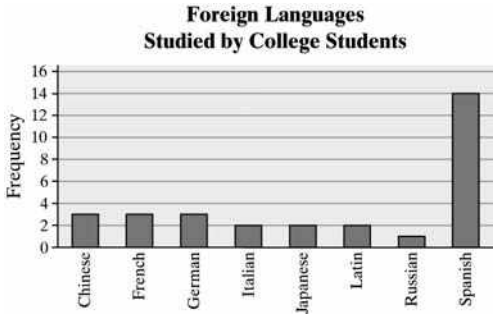
Total number of students = 30

Relative frequency for "Chinese"

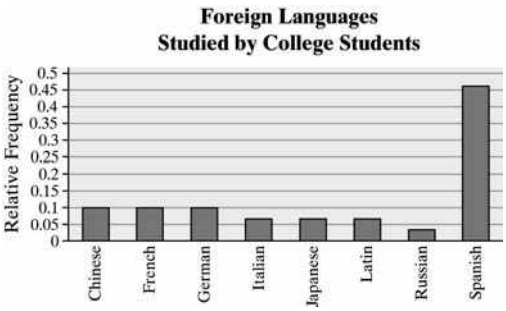
$= \frac{3}{30} = 0.100$ and so on.

Language	Freq.	Rel. Frequency
Chinese	3	0.100
French	3	0.100
German	3	0.100
Italian	2	0.067
Japanese	2	0.067
Latin	2	0.067
Russian	1	0.033
Spanish	14	0.467

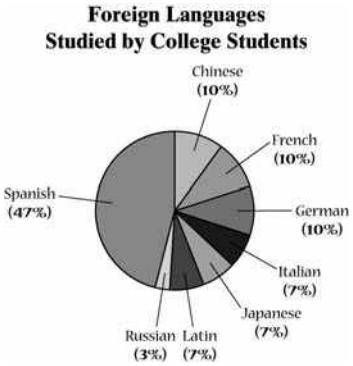
(c)



(d)



(e)



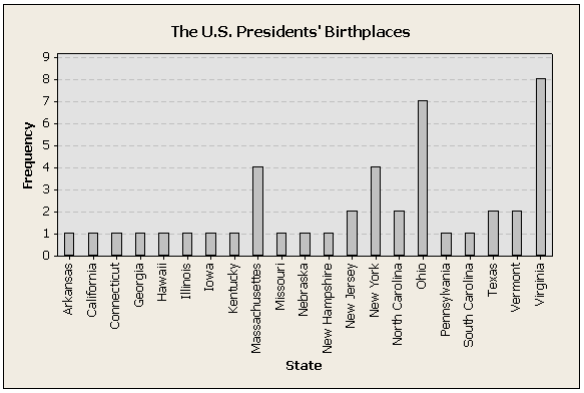
30. (a)

State	AR	CA	CT	GA	HI	IL
Freq.	1	1	1	1	1	1

State	IA	KY	MA	MO	NE
Freq.	1	1	4	1	1

State	NH	NJ	NY	NC	OH
Freq.	1	2	4	2	7

State	PA	SC	TX	VT	VA
Freq.	1	1	2	2	8

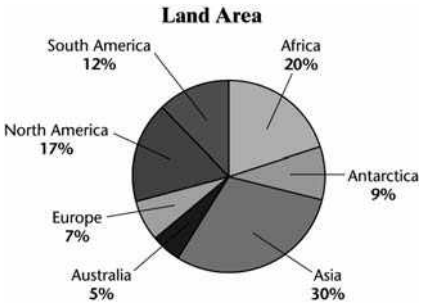


(b) More presidents were born in Virginia than in any other state.

(c) Answers will vary. The data do not take the year of statehood into account. For example, Virginia has been a state for roughly 62 years more than California. The population of the U.S. was more concentrated in the east in the early years so it was more likely that the president would be from that part of the country.

31. (a) It would make sense to draw a pie chart for land area since the 7 continents contain all the land area on Earth.
Total land area is 11,608,000 + 5,100,000 + … + 9,449,000 + 6,879,000 = 57,217,000 square miles
The relative frequency (percentage) for Africa is $\frac{11,608,000}{57,217,000} = 0.2029$.

Continent	Land Area (mi^2)	Rel. Freq.
Africa	11,608,000	0.2029
Antarctica	5,100,000	0.0891
Asia	17,212,000	0.3008
Australia	3,132,000	0.0547
Europe	3,837,000	0.0671
North America	9,449,000	0.1651
South America	6,879,000	0.1202



(b) It would not make sense to draw a pie chart for the highest elevation because there is no whole to which to compare the parts.

32. (a) The researcher wants to determine if online homework improves student learning over traditional pencil-and-paper homework.

(b) This study is an experiment because the researcher is actively imposing treatments (the homework style) on subjects.

(c) Answers will vary. Some examples are same teacher, same semester, and same course.

(d) Assigning different homework methods to entire classes could confound the results because there may be differences between the classes. The instructor may give more instruction to one class than the other. The instructor is not blinded, so he or she may treat one group differently from the other.

(e) *Number of students*: quantitative, discrete
Average age: quantitative, continuous
Average exam score: quantitative, continuous
Type of homework: qualitative
College experience: qualitative

(f) Letter grade is a qualitative variable at the ordinal level of measurement.
Answers will vary. It is possible that ordering the data from A to F is better because it might give more "weight" to the higher grade and the researcher wants to show that a higher percent of students passed using the online homework.

(g) The graph being displayed is a side-by-side bar graph.

(h) Yes; the 'whole' is the set of students who received a grade for the course for each homework method.

(i) The table shows that the two groups with no prior college experience had roughly the same average exam grade. From the bar graph, we see that the students using online homework had a lower percent for As, but had a higher percent who passed with a C or better.

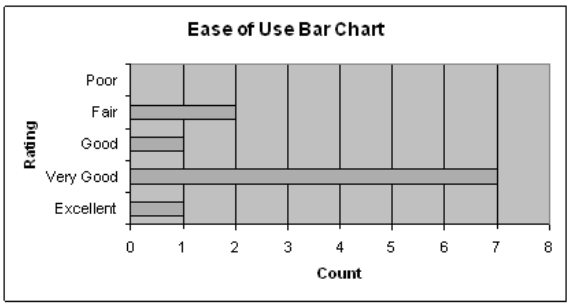
Consumer Reports®: Consumer Reports Rates Treadmills

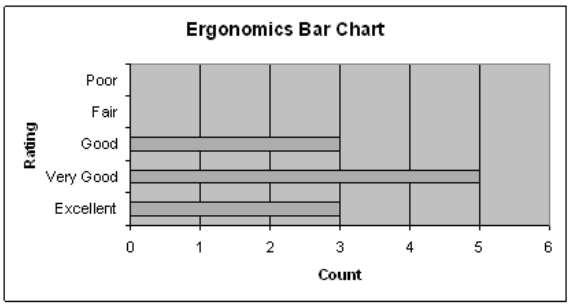
(a) A bar chart is used to display the overall scores. Because the bars are in decreasing order, this is an example of a pareto chart.

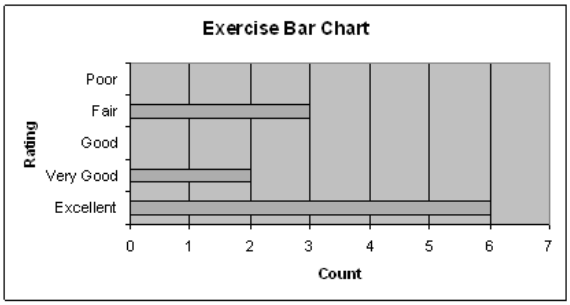
(b) The Precor M9.33 has the highest construction score since it was the only model receiving an excellent rating. Two models, the Tunturi J6F and the ProForm 525E received a Fair rating, making them the models with the lowest ease of use score.

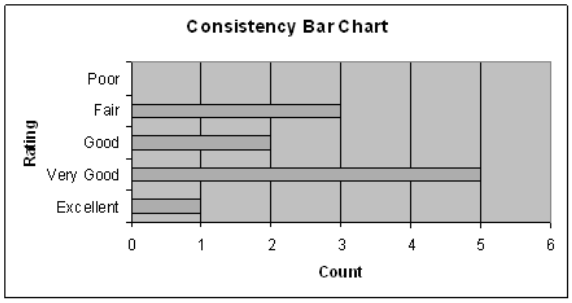
(c) 1 model was rated Excellent, 7 models were rated Very Good, 1 model was rated Good, and 2 models were rated Fair. No models were rated Poor for ease of use.

(d) The following bar charts were created in Microsoft® Excel:









(e) The following scatterplot was obtained by eyeballing the value of the scores from the Overall Score Pareto chart. Although there is a great deal of scatter in the data, even within a similar price range, there appears to be a relationship between score and price. The more expensive models tested by Consumer Reports in March 2002 tended to score higher in overall performance. (One should be cautious about generalizing the conclusions to the universe of treadmills since only a small sample of treadmills have been tested here.)

2.2 Organizing Quantitative Data – The Popular Displays

1. Answers will vary. We have already seen that the pattern of a distribution is the same, whether we look at frequencies or relative frequencies. However, in statistics we will often use data from a sample to guide us to conclusions about the larger population from which the sample is drawn. (This is called inferential statistics.) The actual frequencies for a sample do not, by themselves, give much useful information about the population, but the relative frequencies for the sample data will usually be similar to the relative frequencies for the population.

2. Classes should not overlap because either some observations will be counted more than once or it will not be clear to which class an observation has been assigned.

3. A stem-and-leaf plot has the advantage that the raw data can be recovered from the plot whereas it cannot be from a histogram. On the other hand, stem-and-leaf plots are not well suited to large data sets and there is also more flexibility in the choice of classes for a histogram.

4. Answers will vary. Some possibilities follow: Histograms represent quantitative data, while bar graphs represent qualitative data. The bars in a histogram touch, but the bars in a bar graph do not touch. The width of the bars of a histogram has meaning, but the width of the bars in a bar graph is meaningless.

5. Classes

6. Not all the class widths are the same and the bars in a histogram should touch (unless there is a 0 frequency between two classes).

7. True; for example, if the class width is too large, the data may be lumped into only one or two classes making interpretation difficult.

8. False; histograms are preferred over stem-and-leaf plots for large data sets. Since stem-and-leaf plots require each data value to be represented, they can become cumbersome when the number of values is large.

9. False. The distribution shape shown has a longer tail to the right so it is skewed right.

10. False. The distribution shape shown is symmetric and bell-shaped.

11. **(a)** 8 **(b)** 2 **(c)** 15

(d) $11 - 7 = 4$ **(e)** $\frac{15}{100} = 0.15$ or 15%

(f) The distribution is roughly symmetric.

12. **(a)** 4 cars

(b) There were 9 weeks in which 2 cars sold.

(c) Total frequency = 4 + 2 + 9 + 8 + 12 + 8 + 5 + 2 + 1 + 1 = 52 (as required)
Percentage of time two cars are sold
$= \frac{9}{52} \cdot 100 = 17.3\%$

(d) Slightly skewed to the right.

13. **(a)** Total frequency = 2 + 3 + 13 + 42 + 58 + 40 + 31 + 8 +2 + 1 = 200

(b) 10 (e.g. 70 – 60 = 10)

(c)

IQ Score (class)	Frequency
60–69	2
70–79	3
80–89	13
90–99	42
100–109	58
110–119	40
120–129	31
130–139	8
140–149	2
150–159	1

(d) The class '100 – 109' has the highest frequency.

(e) The class '150 – 159' has the lowest frequency.

14. **(a)** 200 (e.g. 200 – 0 = 200)

(b) 0-199, 200-399, 400-599, 600-799, 800-999, 1000-1199, 1200-1399, 1400-1599, 1600-1799

(c) The highest frequency is in class 0 – 199.

(d) The distribution is skewed right.

(e) Answers will vary. The statement is incorrect because they are comparing counts from populations of different size. To make a fair comparison, the reporter should use rates of fatalities such as the number of fatalities per 1000 residents.

15. (a) Likely skewed right. Most household incomes will be to the left (perhaps in the $50,000 to $150,000 range), with fewer higher incomes to the right (in the millions).

(b) Likely bell-shaped. Most scores will occur near the middle range, with scores tapering off equally in both directions.

(c) Likely skewed right. Most households will have, say, 1 to 4 occupants, with fewer households having a higher number of occupants.

(d) Likely skewed left. Most Alzheimer's patients will fall in older-aged categories, with fewer patients being younger.

16. (a) Likely skewed right. More individuals would consume fewer alcoholic drinks per week, while less individuals would consume more alcoholic drinks per week.

(b) Likely uniform. There will be approximately an equal number of students in each age category.

(c) Likely skewed left. Most hearing-aid patients will fall in older-aged categories, with fewer patients being younger.

(d) Likely bell-shaped. Most heights will occur, say, in the 66- to 70-inch range, with heights tapering off equally in both directions.

17. (a) The closing price at the end of May 2006 was about $45.

(b) The closing price at the end of December 2007 was about $215.

(c) $\frac{215-45}{45} = \frac{170}{45} \approx 3.78$
The price of the stock increased by about 378%.

(d) $\frac{140-168}{168} = \frac{-28}{168} \approx -0.17$
The price of the stock decreased by about 17%.

18. (a) About 8.8 million motor vehicles were produced in the United States in 1991.

(b) About 13.0 million motor vehicles were produced in the United States in 1999.

(c) $\frac{13.0-8.8}{8.8} = \frac{4.2}{8.8} \approx 0.477$
The number of vehicles produced increased by about 47.7% between 1991 and 1999.

(d) $\frac{11.3-13.0}{13.0} = \frac{-1.7}{13.0} \approx -0.131$
The number of vehicles produced decreased by about 13.1% between 1999 and 2006.

19. (a) For 1992, the unemployment rate was about 7.5% and the inflation rate was about 3.0%.

(b) For 2006, the unemployment rate was about 4.6% and the inflation rate was about 3.4%.

(c) $7.5\% + 3.0\% = 10.5\%$
The misery index for 1992 was 10.5%.
$4.6\% + 3.4\% = 8.0\%$
The misery index for 2006 was 8.0%.

(d) Answers may vary. One possibility: An increase in the inflation rate seems to be followed by an increase in the unemployment rate. Likewise, a decrease in the inflation rate seems to be followed by a decrease in the unemployment rate.

20. (a) In 1990, the men's prize money was £230,000 and the ladies' prize money was £207,000.

(b) In 2006, the men's prize money was £655,000 and the ladies' prize money was £625,000.

(c) Answers may vary. One possibility: Until 2007, the prize money for men's singles is higher than the prize money for ladies' singles. Both prizes increase over time at similar rates.

(d) In 2007, the prize money for men's and ladies' singles was the same for the first time. The prize money for each was £700,000.

(e) $\frac{700,000-655,000}{655,000} \approx 0.069$

The men's prize money increased by about 6.9% from 2006 to 2007.

$\frac{700,000-625,000}{625,000} = 0.12$

The ladies' prize money increased by 12% from 2006 to 2007.

21. (a) Total frequency $= 16+18+12+3+1$
$= 50$

Relative frequency of 0 children = 16/50 = 0.32, and so on.

Number of Children Under Five	Relative Frequency
0	0.32
1	0.36
2	0.24
3	0.06
4	0.02

(b) $\frac{12}{50} = 0.24$ or 24% of households have two children under the age of 5.

(c) $\frac{18+12}{50} = \frac{30}{50} = 0.6$ or 60% of households have one or two children under the age of 5.

22. (a) Total frequency = 50.

Relative frequency of 1 throw until a miss = 16/50 = 0.32, and so on.

Number of Free Throws Until a Miss	Relative Frequency
1	0.32
2	0.22
3	0.18
4	0.14
5	0.04
6	0.06
7	0.00
8	0.02
9	0.00
10	0.02

(b) $\frac{7}{50} = 0.14$; 14% of the time she first missed on the fourth try.

(c) $\frac{1}{50} = 0.02$; 2% of the time she first missed on the tenth try.

(d) 'at least 5' means that the basketball player misses on the 6^{th} shot or 7^{th} shot or 8^{th}, etc. $\frac{3+0+1+0+1}{50} = \frac{5}{50} = 0.10$ or 10% of the time.

23. From the legend, 1|0 represents 10, so the original data set is:
10, 11, 14, 21, 24, 24, 27, 29, 33, 35, 35, 35, 37, 37, 38, 40, 40, 41, 42, 46, 46, 48, 49, 49, 53, 53, 55, 58, 61, 62

24. From the legend, 24|0 represents 240, so the original data set is:
240, 244, 247, 252, 252, 253, 259, 259, 263, 264, 265, 268, 268, 269, 270, 271, 271, 273, 276, 276, 282, 283, 288

25. From the legend, 1|2 represents 1.2, so the original data set is:
1.2, 1.4, 1.6, 2.1, 2.4, 2.7, 2.7, 2.9, 3.3, 3.3, 3.3, 3.5, 3.7, 3.7, 3.8, 4.0, 4.1, 4.1, 4.3, 4.6, 4.6, 4.8, 4.8, 4.9, 5.3, 5.4, 5.5, 5.8, 6.2, 6.4

26. From the legend, 12|3 represents 12.3, so the original data set is:
12.3, 12.7, 12.9, 12.9, 13.0, 13.4, 13.5, 13.7, 13.8, 13.9, 13.9, 14.2, 14.4, 14.4, 14.7, 14.7, 14.8, 14.9, 15.1, 15.2, 15.2, 15.5, 15.6, 16.0, 16.3

27. (a) 19 classes

(b) Lower class limits: 0, 2000, 4000, 6000, 8000, 10,000, 12,000, 14,000, 16,000, 18,000, 20,000, 22,000, 24,000, 26,000, 28,000, 30,000, 32,000, 34,000, 36,000
Upper class limits: 1999, 3999, 5999, 7999, 9999, 11,999, 13,999, 15,999, 17,999, 19,999, 21,999, 23,999, 25,999, 27,999, 29,999, 31,999, 33,999, 35,999, 37,999

(c) The class width is found by subtracting consecutive lower class limits. For example, 4,000 – 2,000 = 2,000. Therefore, the class width is 2,000 (dollars).

28. **(a)** 8 classes

(b) Lower class limits: 0, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0
Upper class limits: 0.9, 1.9, 2.9, 3.9, 4.9, 5.9, 6.9, 7.9

(c) The class width is found by subtracting consecutive lower class limits. For example, $2.0 - 1.0 = 1.0$. Therefore, the class width is 1.0.

29. **(a)** 6 classes

(b) Lower class limits: 15, 20, 25, 30, 35, 40;
Upper class limits: 19, 24, 29, 34, 39, 44

(c) The class width is found by subtracting consecutive lower class limits. For example, $20 - 15 = 5$. Therefore, the class width is 5 (years).

30. **(a)** 6 classes

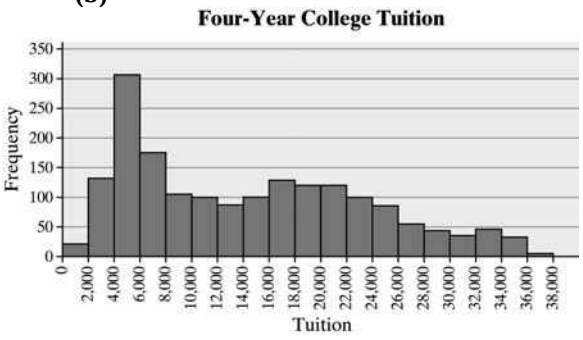
(b) Lower class limits: 0, 5000, 10,000, 15,000, 20,000, 25,000
Upper class limits: 4999, 9999, 19,999, 24,999, 25,999, 29,999

(c) The class width is found by subtracting consecutive lower class limits. For example, $10,000 - 5000 = 5000$. Therefore, the class width is 5000 (students).

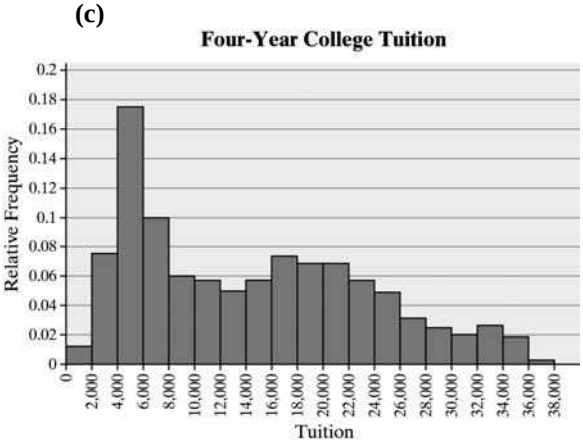
31. **(a)** Total frequency = 19 + 132 + 308 + … + 46 + 32 + 6 = 1790
Relative frequency for 0 – 1999 = 19/1790 = 0.0106 and so on.

Tuition ($)	Relative Frequency
0-1999	0.0106
2000-3999	0.0737
4000-5999	0.1721
6000-7999	0.1006
8000-9999	0.0592
10,000-11,999	0.0553
12,000-13,999	0.0497
14,000-15,999	0.0547
16,000-17,999	0.0698
18,000-19,999	0.0648
20,000-21,999	0.0654
22,000-23,999	0.0559
24,000-25,999	0.0464
26,000-27,999	0.0296
28,000-29,999	0.0251
30,000-31,999	0.0201
32,000-33,999	0.0257
34,000-35,999	0.0179
36,000-37,999	0.0034

(b)



(c)



Total number of colleges with tuition less than $4000 = 19 + 132 = 151

$\frac{151}{1790} \cdot 100 \approx 8.44\%$ of colleges had tuition of less than $4000.

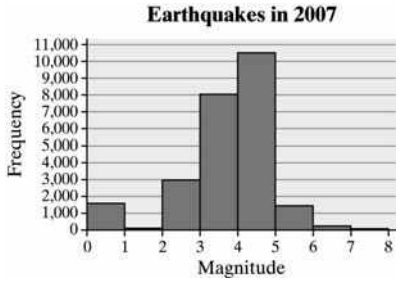
Total number of colleges with tuition of $30,000 or more = 36 + 46 + 32 + 6 = 120

$\frac{120}{1790} \cdot 100 \approx 6.70\%$ of colleges had tuition of $30,000 or more.

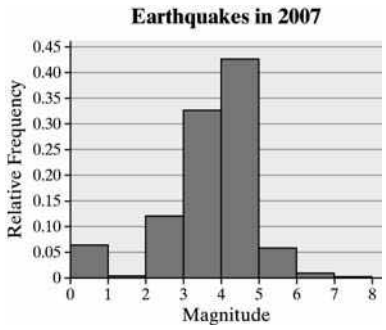
32. (a) Total earthquakes = 1600 + 33 + 2983 + 8028 + 10,465 + 1466 + 151 + 11 = 24,737
Relative frequency for 0-0.9 = 1600/24,737 = 0.0647 and so on.

Magnitude	Relative Frequency
0-0.9	0.0647
1.0-1.9	0.0013
2.0-2.9	0.1206
3.0-3.9	0.3245
4.0-4.9	0.4231
5.0-5.9	0.0593
6.0-6.9	0.0061
7.0-7.9	0.0004

(b)



(c)



$\frac{10,465}{24,737} \cdot 100 = 42.31\%$ of earthquakes in 2007 registered 4.0-4.9 in magnitude, and

$$\frac{10,465 + 8028 + 2983 + 33 + 1600}{24,737} \cdot 100$$

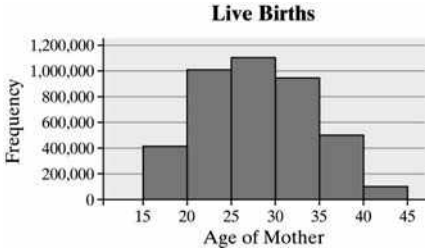
$$= \frac{23,109}{24,737} \cdot 100 = 93.42\%$$

of earthquakes in 2007 registered 4.9 or less in magnitude.

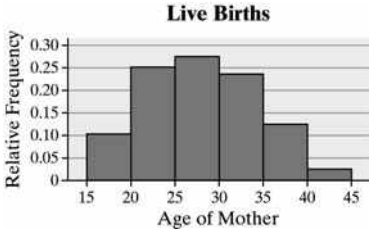
33. (a) Total births = 414,406 + 1,040,399 + 1,132,293 + 952,013 + 483,401 + 104,644 = 4,127,156
Relative frequency for 15-19 = 414,406/4,127,156 = 0.1004 and so on.

Age	Relative Frequency
15-19	0.1004
20-24	0.2521
25-29	0.2744
30-34	0.2307
35-39	0.1171
40-44	0.0254

(b)



(c)

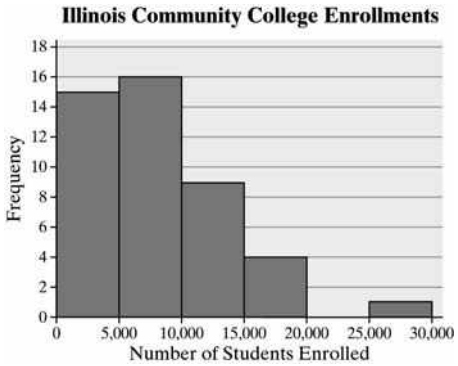


The relative frequency is 0.0254, so 2.54% of live births were to women 40-44 years of age.
$0.1004 + 0.2521 = 0.3525$ so 35.25% of live births were to women 24 years of age or younger.

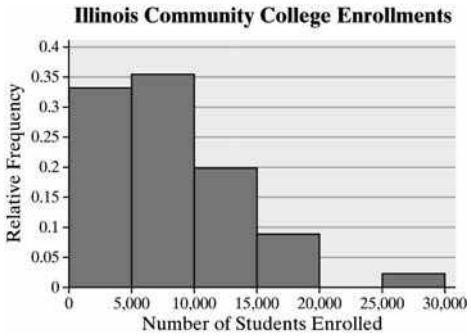
34. (a) Total = 15 + 16 + 9 + 4 + 0 + 1 = 45
Relative frequency for 0-4999 = 15/45 = 0.3333 and so on.

Enrollment	Relative Frequency
0-4999	0.3333
5000-9999	0.3556
10,000-14,999	0.2000
15,000-19,999	0.0889
20,000-24,999	0.0000
25,000-29,999	0.0222

(b)



(c)



The relative frequency is 0.3556, so 35.56% of public Illinois community colleges enrolled between 5000 and 9999 students.

$0.0889 + 0 + 0.0222 = 0.1111$ so 11.11% of public Illinois community colleges enrolled 15,000 or more students.

35. (a) The data are discrete. The possible values for the number of color televisions in a household are countable.

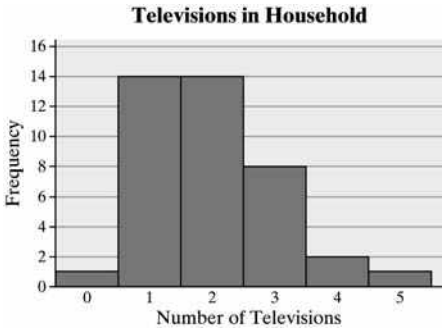
(b), (c)
The relative frequency for 0 color televisions is 1/40 = 0.025, and so on.

Number of Color TVs	Frequency	Relative Frequency
0	1	0.025
1	14	0.350
2	14	0.350
3	8	0.200
4	2	0.050
5	1	0.025

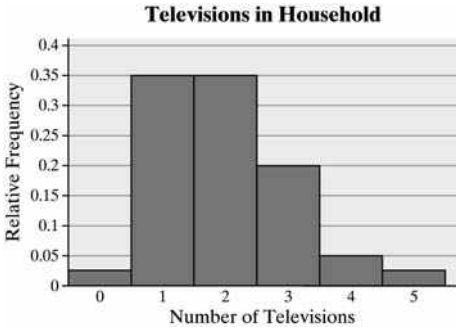
(d) The relative frequency is 0.2 so 20% of the households surveyed had 3 color televisions.

(e) $0.05 + 0.025 = 0.075$
7.5% of the households in the survey had 4 or more color televisions.

(f)



(g)



(h) The distribution is skewed right.

36. (a) The data are discrete. The possible values for the number of customers waiting for a table are countable.

(b) and **(c)**
Relative frequency of 3 customers waiting = 2/40 = 0.05, and so on.

Number of Customers	Freq.	Rel. Freq.
3	2	0.050
4	3	0.075
5	3	0.075
6	5	0.125
7	4	0.100
8	8	0.200
9	4	0.100
10	4	0.100
11	4	0.100
12	0	0.000
13	2	0.050
14	1	0.025

(d) $10.0 + 10.0 + 0.0 + 5.0 + 2.5 = 27.5\%$ of the Saturdays had 10 or more customers waiting for a table at 6 p.m.

(e) $5.0 + 7.5 + 7.5 = 20.0\%$ of the Saturdays had 5 or fewer customers waiting for a table at 6 p.m.

(f)

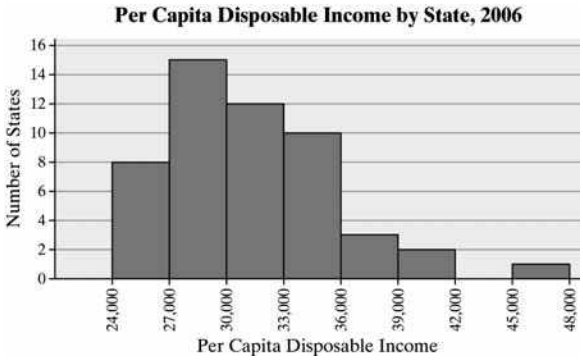
(g)

(h) The distribution is more or less symmetric.

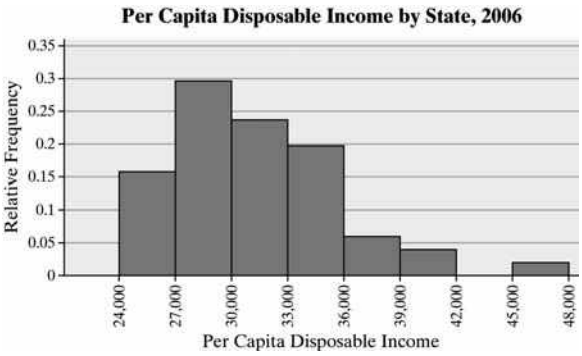
37. **(a)** and **(b)**
Relative frequency for 24,000-26,999 = 8/51 = 0.1569 and so on.

Disposable Income ($)	Freq.	Rel. Freq.
24,000 – 26,999	8	0.1569
27,000 – 29,999	15	0.2941
30,000 – 32,999	12	0.2353
33,000 – 35,999	10	0.1961
36,000 – 38,999	3	0.0588
39,000 – 41,999	2	0.0392
42,000 – 44,999	0	0.0000
45,000 – 47,999	1	0.0196

(c)



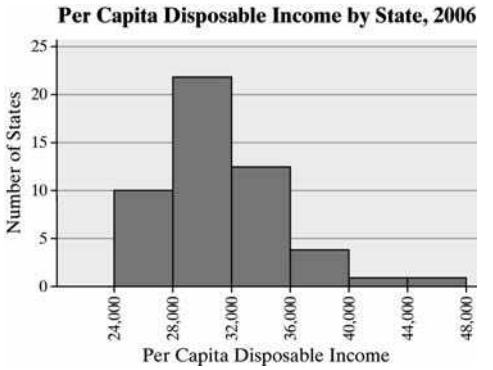
(d)

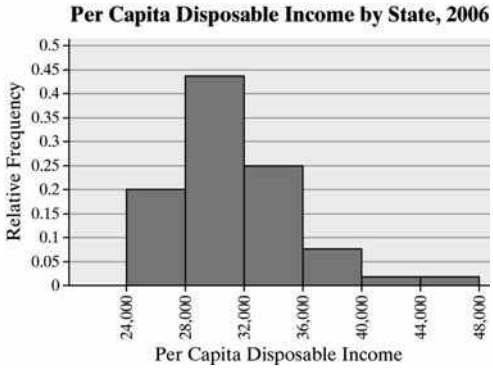


(e) The distribution appears to be skewed right.

(f) Relative frequency for 24,000-27,999 = 10/51 = 0.1961 and so on.

Disposable Income ($)	Freq.	Rel. Freq.
24,000 – 27,999	10	0.1961
28,000 – 31,999	22	0.4314
32,000 – 35,999	13	0.2549
36,000 – 39,999	4	0.0784
40,000 – 43,999	1	0.0196
44,000 – 47,999	1	0.0196





The distribution appears to be skewed right.

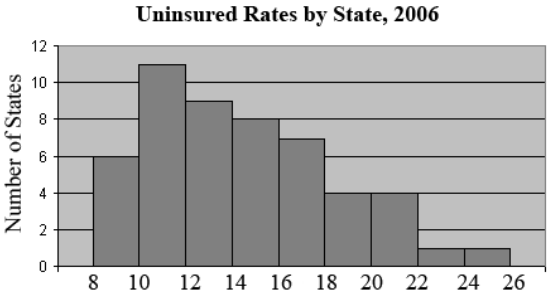
(g) Answers will vary. While both distributions indicate the data are skewed right, the first distribution provides a more detailed look at the data. The second distribution is a bit coarser.

38. (a) and **(b)**

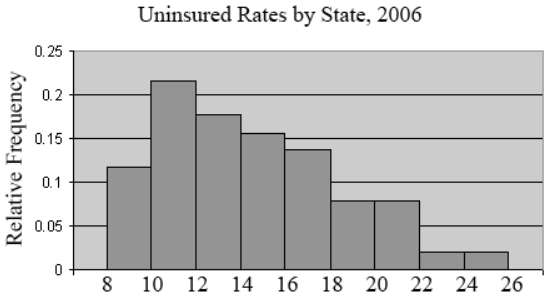
Relative frequency for 8.0-9.9 = 7/51 = 0.1373 and so on.

% Uninsured	Freq.	Rel. Freq.
8.0 – 9.9	6	0.1176
10.0 – 11.9	11	0.2157
12.0 – 13.9	9	0.1765
14.0 – 15.9	8	0.1569
16.0 – 17.9	7	0.1373
18.0 – 19.9	4	0.0784
20.0 – 21.9	4	0.0784
22.0 – 23.9	1	0.0196
24.0 – 25.9	1	0.0196

(c)



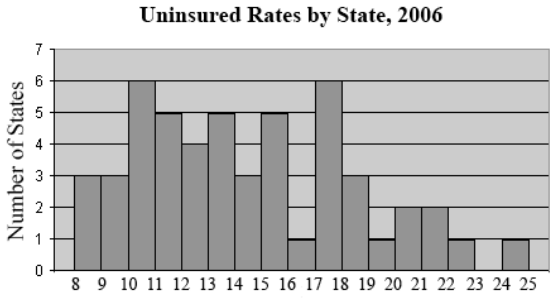
(d)

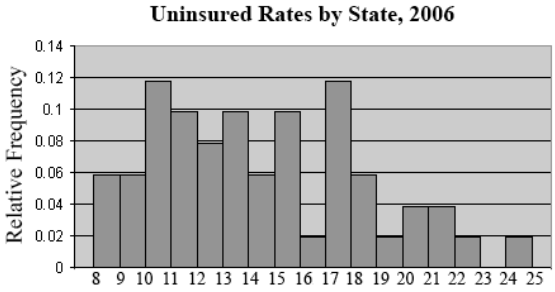


(e) The distribution is skewed right.

(f) Relative frequency for 8.0-8.9 = 3/51 = 0.0588 and so on.

% Uninsured	Freq.	Rel. Freq.
8.0 – 8.9	3	0.0588
9.0 – 9.9	3	0.0588
10.0 – 10.9	6	0.1176
11.0 – 11.9	5	0.0980
12.0 – 12.9	4	0.0784
13.0 – 13.9	5	0.0980
14.0 – 14.9	3	0.0588
15.0 – 15.9	5	0.0980
16.0 – 16.9	1	0.0196
17.0 – 17.9	6	0.1176
18.0 – 18.9	3	0.0588
19.0 – 19.9	1	0.0196
20.0 – 20.9	2	0.0392
21.0 – 21.9	2	0.0392
22.0 – 22.9	1	0.0196
23.0 – 23.9	0	0.0000
24.0 – 24.9	1	0.0196





The distribution is skewed right.

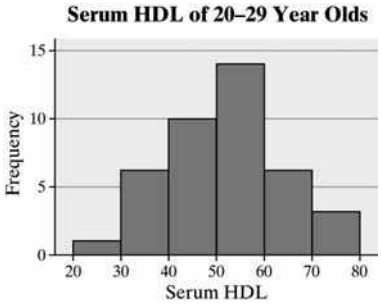
(g) Answers will vary. Both frequency distributions show the data are skewed right. The number of classes in the second distribution gives more detail, but this makes the graph a bit more jagged. The first distribution gives a cleaner view of the data.

39. **(a)** and **(b)**

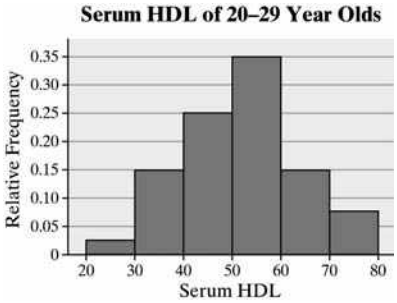
Total number of data points = 40
Relative frequency of 20 – 29 = 1/40 = 0.025, and so on.

HDL Cholesterol	Frequency	Relative Frequency
20–29	1	0.025
30–39	6	0.150
40–49	10	0.250
50–59	14	0.350
60–69	6	0.150
70–79	3	0.075

(c)



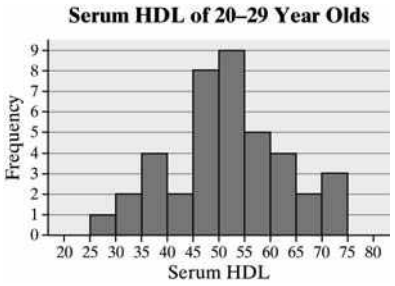
(d)

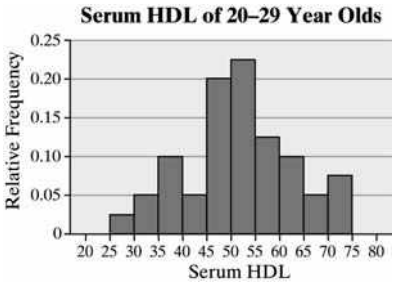


(e) The distribution appears to be roughly bell-shaped.

(f) Relative frequency of 25 – 29 = 1/40 = 0.025, and so on.

HDL Cholesterol	Frequency	Relative Frequency
20–24	0	0.000
25–29	1	0.025
30–34	2	0.050
35–39	4	0.100
40–44	2	0.050
45–49	8	0.200
50–54	9	0.225
55–59	5	0.125
60–64	4	0.100
65–69	2	0.050
70–74	3	0.075





The distribution is roughly bell-shaped.

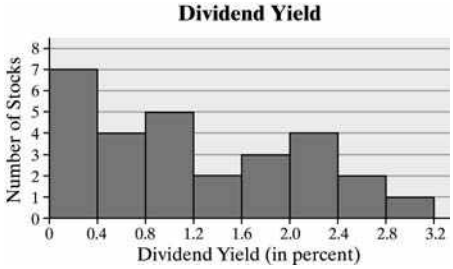
(g) Answers will vary. The first distribution gives a smoother pattern. The additional detail in the second case does not provide much more information.

40. **(a)** and **(b)**

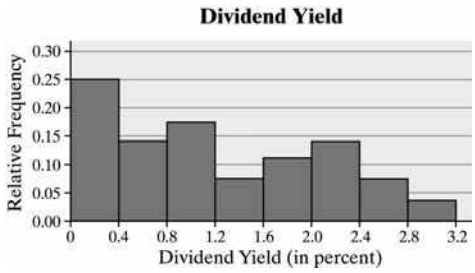
Relative frequency for 0.00-0.39 = 7/28 = 0.2500, and so on.

Dividend	Freq.	Rel. Freq.
0.00 – 0.39	7	0.2500
0.40 – 0.79	4	0.1429
0.80 – 1.19	5	0.1786
1.20 – 1.59	2	0.0714
1.60 – 1.99	3	0.1071
2.00 – 2.39	4	0.1429
2.40 – 2.79	2	0.0714
2.80 – 3.19	1	0.0357

(c)



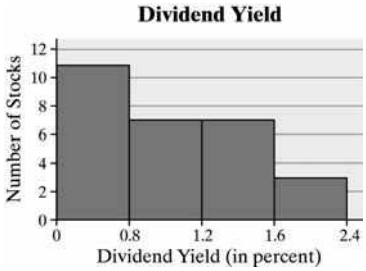
(d)

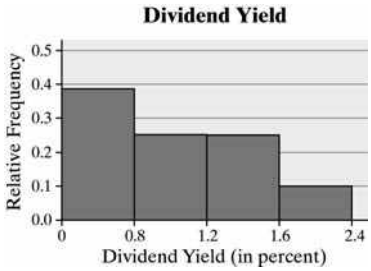


(e) The distribution is skewed right.

(f) Relative frequency for 0.00-0.79 = 11/28 = 0.3929 and so on.

Dividend	Freq.	Rel. Freq.
0.00 – 0.79	11	0.3929
0.80 – 1.59	7	0.2500
1.60 – 2.39	7	0.2500
2.40 – 3.19	3	0.1071





The distribution is skewed right.

(g) Answers will vary. Both distributions indicate the data are skewed right. The first graph is preferred because it gives more detailed information. The second graph is a little too compressed to get a complete view of what is happening with the data.

41. Answers will vary. One possibility follows.

(a) We can determine a class width by subtracting the smallest value from the largest, dividing by the desired number of classes, then rounding up. For example,

$$\frac{23.59 - 6.37}{6} = 2.87 \rightarrow 3$$

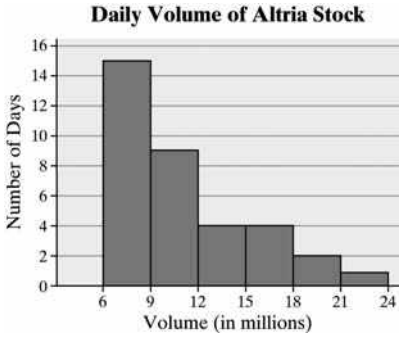
Our first lower class limit should be a nice number below the smallest data value. In this case, 6 is a good first lower limit since it is the nearest whole number below the smallest data value. Thus, we will have a class width of 3 and the first class will have a lower limit of 6.

(b), (c)

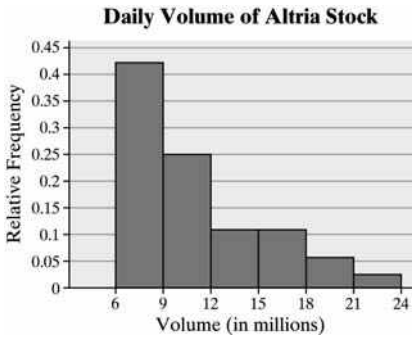
Relative frequency for 6-8.99 = 15/35 = 0.4286 and so on.

Volume	Freq.	Rel. Freq.
6 – 8.99	15	0.4286
9 – 11.99	9	0.2571
12 – 14.99	4	0.1143
15 – 17.99	4	0.1143
18 – 20.99	2	0.0571
21 – 23.99	1	0.0286

(d)



(e)



(f) The distribution is skewed right.

42. Answers will vary. One possibility follows.

(a) We can determine a class width by subtracting the smallest value from the largest, dividing by the desired number of classes, then rounding up. For example,

$$\frac{1459-98}{8} = 170.125 \rightarrow 200$$

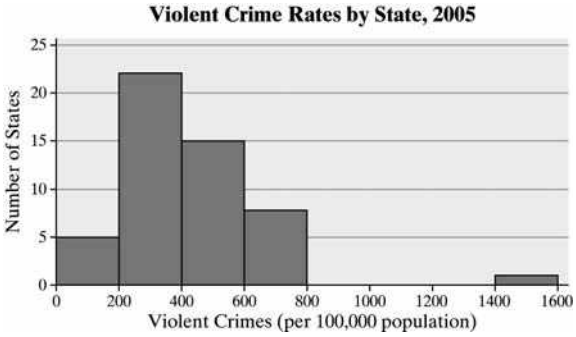
Our first lower class limit should be a nice number below the smallest data value. In this case, 0 is a good first lower limit. Thus, we will have a class width of 200 and a first class lower limit of 0

(b), (c)

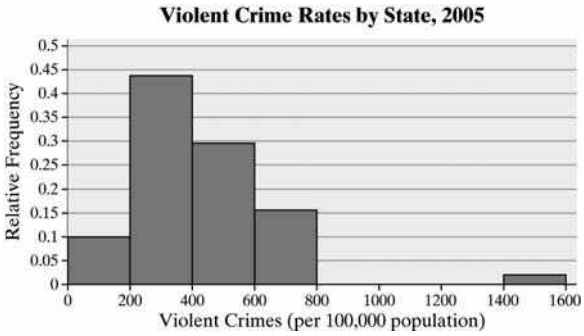
Relative frequency for 0-199 = 5/51 = 0.0980, and so on.

Violent Crimes per 100,000	Freq.	Rel. Freq.
0 – 199	5	0.0980
200 – 399	22	0.4314
400 – 599	15	0.2941
600 – 799	8	0.1569
800 – 999	0	0.0000
1000 – 1199	0	0.0000
1200 – 1399	0	0.0000
1400 – 1599	1	0.0196

(d)



(e)



(f) The distribution is skewed right.

43. **(a)**

President Ages at Inauguration

```
4 | 23
4 | 667899
5 | 0011112244444
5 | 555566677778
6 | 0111244
6 | 589
```

Legend: 4 | 2 represents 42 years.

(b) The distribution appears to be roughly symmetric and bell shaped.

44. **(a)**

Divorce Rates by State, 2004

```
1 | 7
2 | 2
2 | 568889
3 | 00011223
3 | 56667788999
4 | 011244
4 | 5677889
5 | 013
5 |
6 | 34
```

Legend: 1 | 7 represents 1.7 divorces per 1000 population.

(b) The distribution appears to be roughly symmetric and bell shaped.

45. **(a)**

Fat in McDonald's Breakfast

```
0 | 39
1 | 1266
2 | 1224577
3 | 0012267
4 | 6
5 | 159
```

Legend: 5 | 1 represents 51 grams of fat.

(b) The distribution appears to be roughly symmetric and bell shaped.

46. **(a)**

Gasoline Mileages

```
2 | 34
2 | 56777888899999999
3 | 00111111222222233333333444
3 | 5555667
4 |
4 | 5
```

Legend: 2 | 3 represents 23 miles per gallon.

(b) The distribution appears to be slightly skewed right (though possibly bell shaped).

47. (a) Rounded data:

7.1	12.8	8.2	7.0	12.8	7.6
14.8	10.1	11.1	10.5	7.6	20.7
4.9	7.1	6.5	7.0	6.9	5.4
8.3	11.8	10.0	15.5	8.1	7.0
8.3	6.3	6.9	6.1	9.6	13.8
11.9	7.4	15.3	7.5	6.2	7.7
7.3	6.5	8.7	14.0	7.0	6.7
7.0	10.3	6.0	11.4	6.9	6.1
5.0	8.1	5.3			

Average Electric Rates by State, 2006

4	9
5	034
6	01123557999
7	0000011345667
8	112337
9	6
10	0135
11	1489
12	88
13	8
14	08
15	35
16	
17	
18	
19	
20	7

Legend: 4 | 9 represents 4.9 cents/kWh.

(b) The distribution is skewed right.

(c) Hawaii's average retail price is 20.7 cents/kWh. Hawaii's rate is so much higher because it is an island far away from the mainland. Resources on the island are limited and importing resources increases the overall cost.

48. (a) Percentage change in housing prices, rounded to the nearest whole percent:

12	19	14	42	78
40	40	114	56	20
36	87	73	28	42
37	16	24	92	29
14	18	24	41	26
94	63	81	57	41
74	72	29	76	26
19	39	81	19	31

Change in HPI, 2002–2007

1	24468999
2	04466899
3	1679
4	001122
5	67
6	3
7	23468
8	117
9	24
10	
11	4

Legend: 1 | 2 represents a 12% increase.

(b) The distribution is skewed right.

49. (a) **Problems per 100 Vehicles, 2004**

16	2
17	
18	79
19	46
20	9
21	26
22	4
23	
24	0
25	
26	22457
27	6
28	05589
29	578
30	
31	044
32	77
33	
34	6
35	
36	5
37	5
38	6
39	3
40	
41	1
42	
43	2
44	
45	
46	
47	2

Legend: 16 | 2 represents 162 problems.

(b) Number of problems per 100 vehicles, rounded to the nearest tens:

160	190	190	190	200	210
210	220	220	240	260	260
260	270	270	280	280	290
290	290	290	300	300	300
310	310	310	330	330	350
370	380	390	390	410	430
470					

(c) **Problems per 100 Vehicles, 2004**

1	6999
2	01122466677889999
3	0001113357899
4	137

Legend: 1 | 6 represents 160 problems.

(d)

Problems per 100 Vehicles, 2004

1	6999
2	011224
2	66677889999
3	00011133
3	57899
4	13
4	7

Legend: 1 | 6 represents 160 problems.

(e) Answers will vary. The third display (with split stems) seems to be the best. The first display spreads the data too thin, while the second compresses the data too much. In both those cases, it is hard to get a good feel for what the data look like. The third display spreads the data enough to see the main features without diluting the details.

50. (a) Violent crime rates rounded to the nearest tens:

430	710	590	290	350	530
630	450	110	610	510	230
510	260	700	130	290	120
530	260	460	360	430	280
530	550	550	700	250	350
400	320	300	450	760	270
280	290	280	470	180	240
630	390	530	100	750	230
1460	270	280			

(b)

Violent Crime Rates by State, 2005

1	01238
2	334566778888999
3	025569
4	0335567
5	113333559
6	133
7	00156
8	
9	
10	
11	
12	
13	
14	6

Legend: 1 | 0 represents 100 violent crimes.

(c)

Violent Crime Rates by State, 2005

1	0123
1	8
2	334
2	566778888999
3	02
3	5569
4	033
4	5567
5	113333
5	559
6	133
6	
7	001
7	56
8	
8	
9	
9	
10	
10	
11	
11	
12	
12	
13	
13	
14	
14	6

Legend: 1 | 0 represents 100 violent crimes.

(d) Answers will vary. The first display is decent. It clearly shows that the distribution is skewed right and has an outlier. The second display is not as good as the first. Splitting the stems did not reveal any additional information and has made the display more cluttered and cumbersome.

51. (a)

Ages of Academy Award Winners

Best Actor		Best Actress
9	2	1566899
98877766220	3	0123333455689
76555332200	4	112599
4321	5	
20	6	11
6	7	4
	8	0

(b) Answers will vary. It appears that Academy Award winners for best actor tend to be older on the whole than winners for best actress.

52. (a)

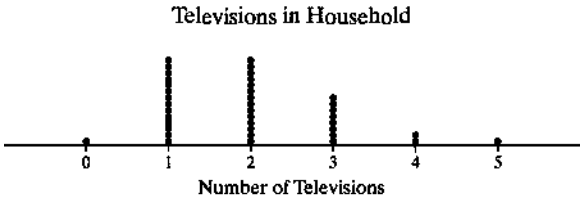
Home Run Distances

McGwire		Bonds
	32	00
	33	
10	34	7
00	35	0
9000	36	00015
70000	37	005555
85500000	38	000005
80000	39	00146
900	40	000045
00000	41	0000000000155677
5300000	42	000000009
0000000	43	00000556
000	44	00002
820000	45	04
100	46	
8000	47	
0	48	8
	49	
0	50	
00	51	
7	52	
	53	
	54	
0	55	

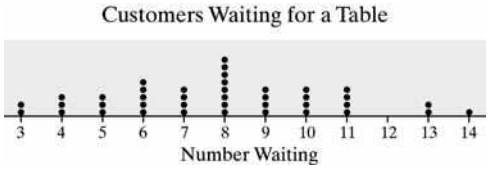
Legend: 0 | 34 | 7 represents 340 feet for McGwire and 347 feet for Bonds.

(b) Answers will vary. For both players, the distances of homeruns mainly fall from 360 to 450 feet. McGwire has quite a few extremely long distances.

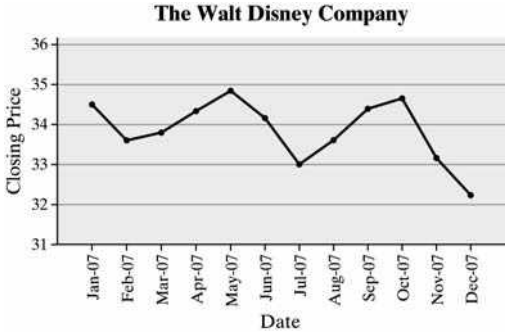
53.



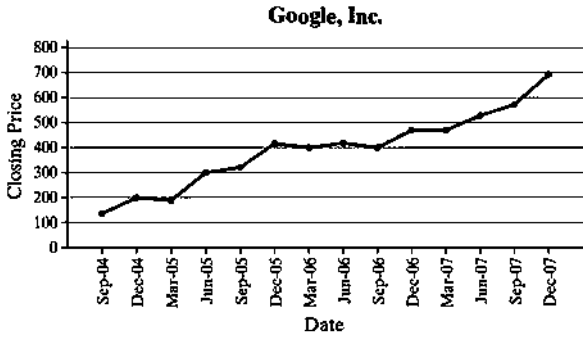
54.



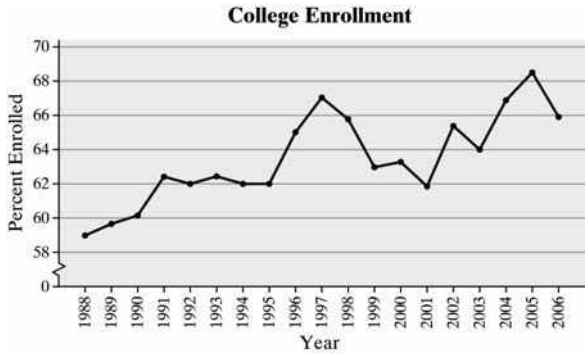
55. The price of Disney stock over the year seemed to fluctuate somewhat, with a general downward trend.



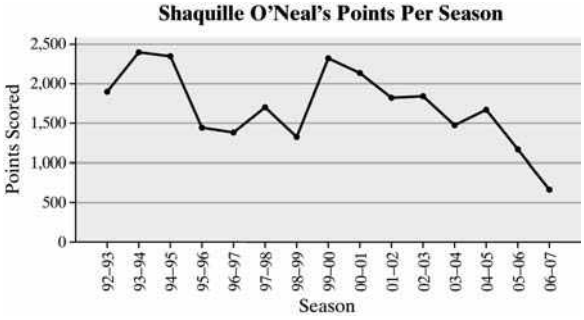
56. Since going public, the price of Google stock has increased steadily and rapidly.



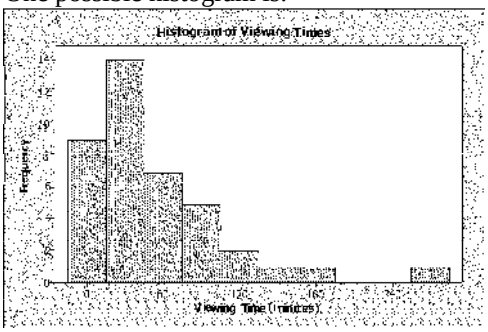
57. The percent of recent high school graduates enrolling in college seems to have increased slightly over the given time period amid a variety of fluctuations. The early 1990s showed no increase, but this was followed by an unusual jump and decline in the mid-to-late 1990s.



58. Between 1995 and 1999, O'Neal had significantly lower total points than the previous three seasons. His points increased strongly for the 1999-2000 season, but have steadily decreased since then.



59. Because the data are quantitative, a stem-and-leaf plot or a histogram would be appropriate. One possible histogram is:



The data appear to be skewed right with a gap and one potential outlier. It seems as if the majority of surfers spent less than one minute viewing the page, while a few surfers spent several minutes viewing the page.

60. Answers will vary. Reports should address the fact that the number of people going to the beach and participating in underwater activities (e.g. scuba diving, snorkeling) has also increased, so an increase in shark attacks is not unexpected. A better comparison would be the rate of attacks per 100,000 beach visitors. The number of fatalities could decrease due to better safety equipment (e.g. bite resistant suits) and better medical care.

2.3 Graphical Misrepresentations of Data

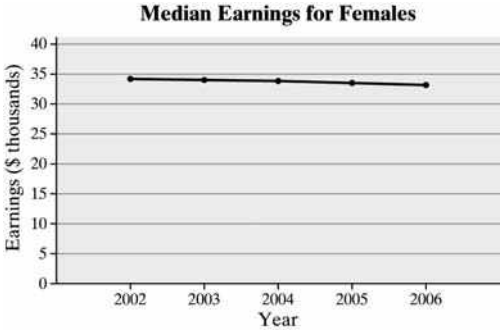
1. The lengths of the bars are not proportional. For example, the bar representing the cost of Clinton's inauguration should be slightly more than 9 times as long as the one for Carter's cost, and twice as long as the bar representing Reagan's cost.

2. (a) The lengths of the bars are not proportional. For example, the bar for soda is 1/3 the size of the bar for cheeseburger, but the number of steps for a cheeseburger is just over twice that for the soda. In addition, it is unclear where the graph begins: at the base of each figure or the bottom of the platform.

(b) Answers will vary. The pictures could be replaced by simple bars (of the same width) that are proportional in area.

3. (a) The vertical axis starts at 31.5 instead of 0. This tends to indicate that the median earnings for females decreased at a faster rate than actually occurred.

(b) This graph indicates that the median earnings for females has decreased slightly over the given time period.



4. (a) The vertical axis starts at 0.08 instead of 0. This might cause the reader to conclude, for example, that the percentage of employed people aged 25-34 who are members of a union is less than half the percentage of those aged 45-54 or 55-64.

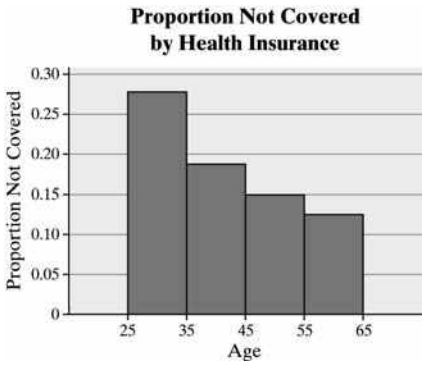
(b)

5. The bar for 12p-6p covers twice as many hours as the other bars. By combining two 3-hour periods, this bar looks larger compared to the others, making afternoon hours look more dangerous. If this bar were split into two periods, the graph may give a different impression. For example, the graph may show that daylight hours are safer.

6. The article is basing its conclusion on a comparison of categories that do not cover the same number of years. A better comparison is the incidence rate (number of accidents per 100,000 licensed drivers). [Note: only about 14% of licensed drivers in 2005 were aged 24 years or younger.]

7. **(a)** The vertical axis starts at 0.1 instead of 0. This might cause the reader to conclude, for example, that the proportion of people aged 25-34 years old who are not covered by any type of health insurance is more than 4 times the proportion for those aged 55-64 years old.

(b)



8. **(a)** The vertical axis starts at 150 without indicating a gap.

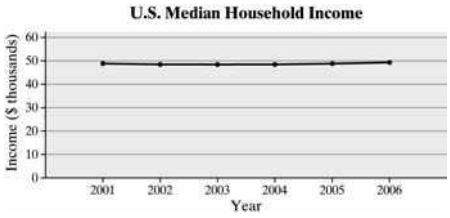
(b) The graph may be trying to convey that new home construction in the Midwest increased significantly between 2000 and 2004, then declined at roughly twice the rate between 2004 and 2006.

(c) The graph does support this view. The sharp drop after 2005 indicates a market cooling and less demand for new construction.

9. **(a)** The vertical axis starts at 47 without including a gap.

(b) The graph may be trying to convey that median household income is on the rise after a period of decline.

(c)



10. **(a)** $\dfrac{696.3}{7.5} = 92.84 \rightarrow 93$

Since the oil reserve in 2007 was about 93 times the oil reserve in 1977, the graphic for 2007 should be 93 times the size (i.e. area) of the graphic for 1977.

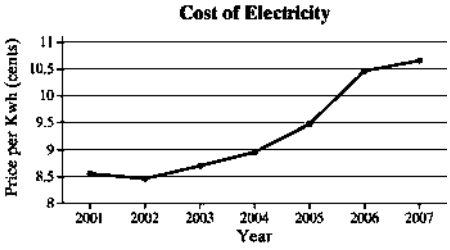
(b) $\dfrac{696.3 \text{ million}}{10.1 \text{ million}} = \dfrac{696.3}{10.1} \approx 68.9$

Assuming no change in production, the 2007 oil reserves would last about 69 days.

11. **(a)** The graphic is misleading because the bars are not proportional. The bar for housing should be a little more than twice the length of the bar for transportation, but it is not.

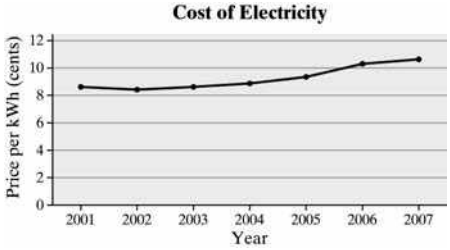
(b) The graphic could be improved by adjusting the bars so that their lengths are proportional.

12. **(a)**

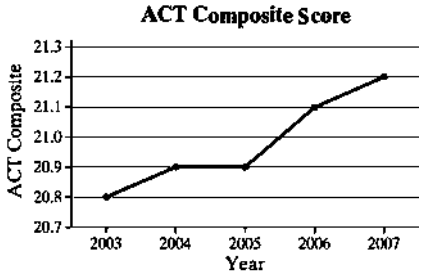


The graphic is misleading because it starts at 8 instead of 0 without indicating a gap. This might cause the reader to conclude that cost of electricity has risen more sharply than actually occurred.

(b)

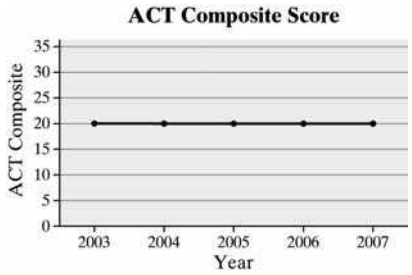


13. (a)



This graphic is misleading because the vertical scale starts at 20.7 instead of 0 without indicating a gap. This might cause the reader to think that ACT composite scores are increasing more quickly than they really are.

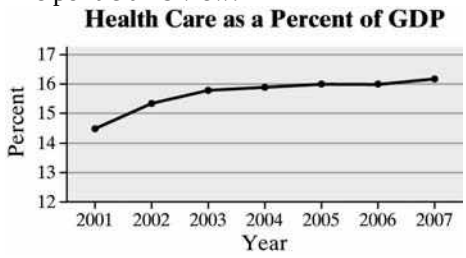
(b)



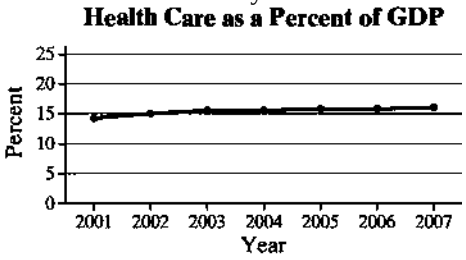
(c) The graph in part (a) is preferred because the trend can be seen. ACT composite scores have been increasing, though not as sharply as indicated in the graph.

14. The graphic does not support the safety manager's claim. The vertical scale starts at 0.17 instead of 0, so the difference between the bars is distorted. While there has been a decrease in the proportion of workers injured, it appears that the decrease is only about 10% of the 1992 rate.

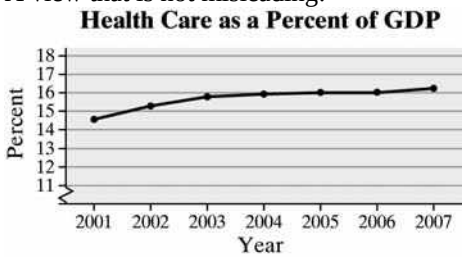
15. (a) The politician's view:



(b) The health care industry's view:

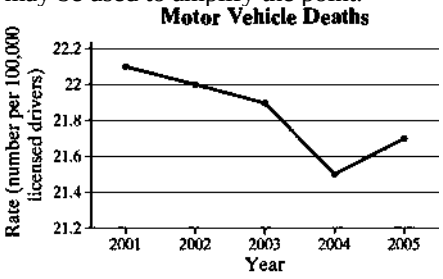


(c) A view that is not misleading:



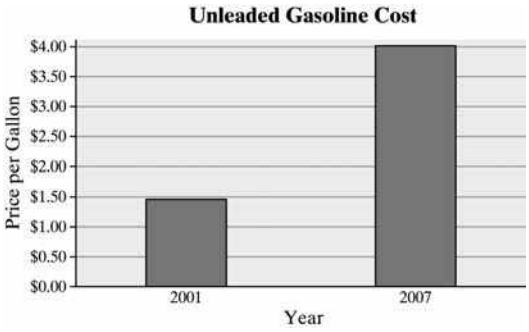
16. (a) Someone wanting to demonstrate that roads are getting less safe would show a plot of the number of motor vehicle deaths. This number is likely to rise each year simply because there are more drivers on the road which increases the likelihood of a serious accident. In addition, a misleading scale on the y-axis may be used to amplify the point.

(b) Someone wanting to demonstrate that roads are becoming safer would want to consider the rate of motor vehicle deaths. For example, the number per 100,000 licensed drivers. While the number of deaths may increase, the number of licensed drivers will also likely increase. Therefore, it is possible for the rate of occurrence to decrease even though the actual number of deaths increases. In addition, a misleading scale on the y-axis may be used to amplify the point.

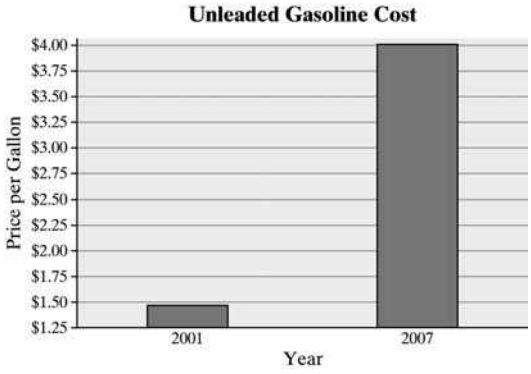


(c) The graph in part (b) is a better representation. As a population grows, an increase in the number of licensed drivers will likely lead to an increase in motor vehicle deaths. The incident rate takes this population growth into account.

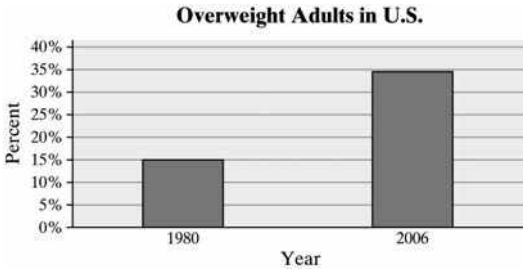
17. (a) A graph that is not misleading will use a vertical scale starting at $0 and bars of equal width. One example:



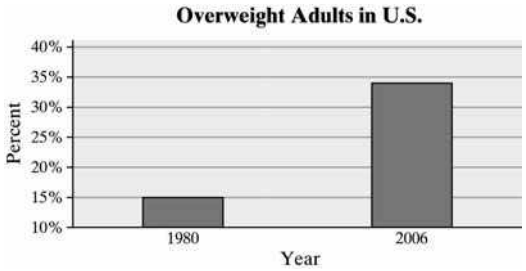
(b) A graph that is misleading might use bars of unequal width or will use a vertical scale that does not start at $0. One example, as follows, indicates that the average price for regular unleaded gasoline has increased ten-fold from 2001 to 2007.



18. (a) A graph that is not misleading will use a vertical scale starting at 0% and bars of equal width. One example:



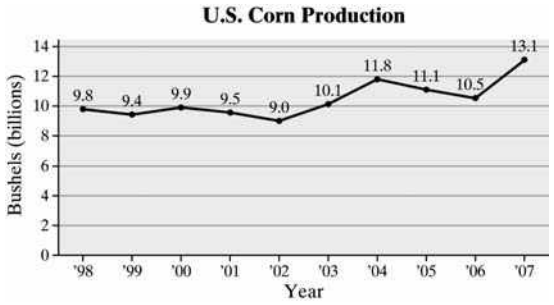
(b) A graph that is misleading might use bars of unequal width or will use a vertical scale that does not start at 0%. One example, as follows, indicates that the percent of overweight adults in the U.S. has more than quadrupled between 1980 and 2006.



19. (a) The graph is a time series plot because the data are plotted in time order.

(b) The graph is too cluttered causing the grid and background to stand out more than the actual data; the axes are not labeled so it is not clear what production measure is being displayed.

(c) The following graph presents the same data, but makes the data stand out and clearly labels the axes.



20. (a) The two variables being graphed are Cornell's tuition and Cornell's ranking.
Tuition: quantitative, discrete, interval level of measurement
Ranking: qualitative, ordinal level of measurement

(b) The graph was likely generated from existing data sources such as the U.S. Department of Education.

(c) The graph is a time series plot because the data are plotted in time order.

(d) Answers may vary. The graph appears to show that the cost to attend Cornell is increasing while its ranking is decreasing. The graph is misleading because the ranking graph may imply that ranking is getting worse (the graph goes down) when in fact a lower value means a better ranking.

(e) Answers may vary. No horizontal or vertical scale is provided; the horizontal scale is inconsistent; the vertical scale is inconsistent between graphs; the reader's attention is drawn more to the picture than the data.

Chapter 2 Review Exercises

1. (a) The bar for natural gas appears to be halfway between 20 and 25, so it appears that the U.S. consumed about 22.5 quadrillion Btu in energy from natural gas during 2006.

(b) The corresponding bar in the graph is more than halfway between 0 and 5, but not quite to 5. Therefore, we might estimate that the U.S. consumed about 3 quadrillion Btu in energy from biomass during 2006.

(c) $41 + 22.5 + 22.5 + 7.5 + 3 + 3 + 0.5 = 100$ Total energy consumption in the U.S. during 2006 was about 100 quadrillion Btu.

(d) "Other" (including geothermal, wind, and solar) has the lowest frequency.

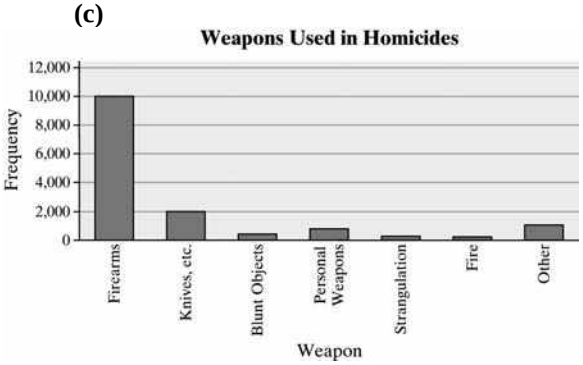
(e) No; the data are qualitative, so order (and thus skewness) is irrelevant.

2. (a) Total homicides = 10,075 + 1,902 + 609 + 892 + 208 + 119 + 1,040 = 14,845
Relative frequency for Firearms = $\frac{10,075}{14,845} \approx 0.6787$ and so on.

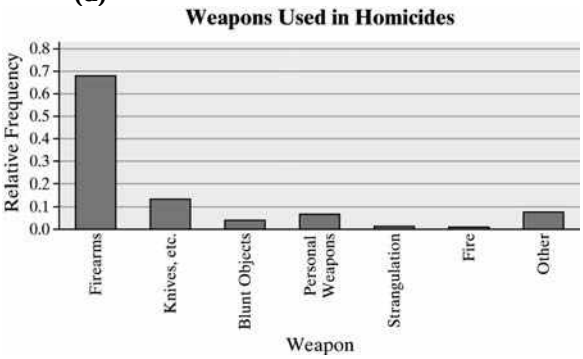
Type of Weapon	Relative Frequency
Firearms	0.6787
Knives or cutting intstruments	0.1281
Blunt objects (clubs, hammers, etc.)	0.0410
Personal weapons (hands, fists, etc.)	0.0601
Strangulation	0.0140
Fire	0.0080
Other weapon or not stated	0.0701

(b) The relative frequency is 0.0410, so 4.1% of the homicides were due to blunt objects.

(c)



(d)

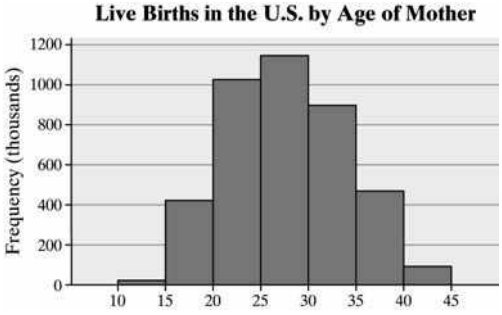


(e)

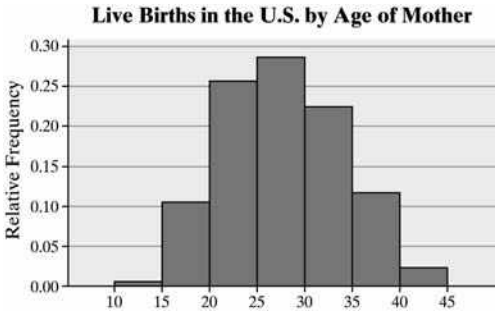
3. (a) Total births (in thousands) = 6 + 435 + 1081 + 1182 + 950 + 499 + 105 = 4258
Relative frequency for 10-14 year old mothers = $6/4258 \approx 0.0014$ and so on.

Age of Mother	Rel. Freq.
10 – 14	0.0014
15 – 19	0.1022
20 – 24	0.2539
25 – 29	0.2776
30 – 34	0.2231
35 – 39	0.1172
40 – 44	0.0247

(b) The distribution is roughly symmetric and bell shaped.



(c)



(d) From the relative frequency table, the relative frequency of 20-24 is 0.2539 and so the percentage is 25.39%.

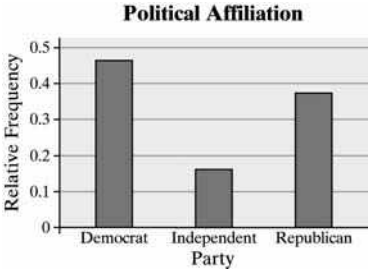
(e) $\dfrac{950+499+105}{4258} = \dfrac{1554}{4258} \approx 0.3650$

36.5% of live births were to mothers aged 30 years or older.

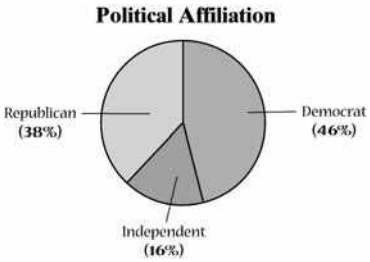
4. (a) and **(b)**

Affiliation	Frequency	Relative Frequency
Democrat	46	0.46
Independent	16	0.16
Republican	38	0.38

(c)



(d)

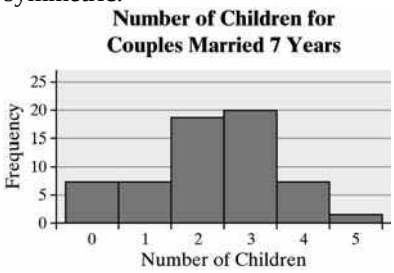


(e) Democrat appears to be the most common affiliation in Naperville.

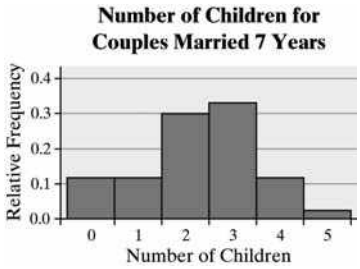
5. (a), **(b)**

Family Size	Freq.	Rel. Freq.
0	7	0.1167
1	7	0.1167
2	18	0.3000
3	20	0.3333
4	7	0.1167
5	1	0.0167

(c) The distribution is more or less symmetric.



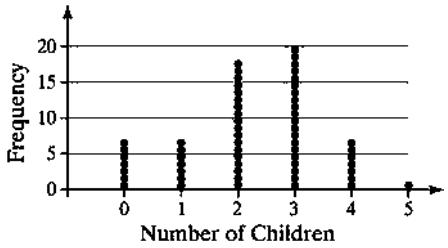
(d)



(e) From the relative frequency table, the relative frequency of two children is 0.3000 so 30% of the couples have two children.

(f) From the frequency table, the relative frequency of at least two children (i.e. two or more) is $0.3000 + 0.3333 + 0.1167 + 0.0167 = 0.7667$ or 76.67%. So, 76.67% of the couples have at least two children.

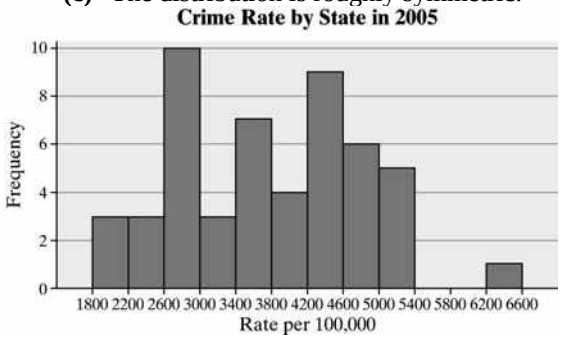
(g)



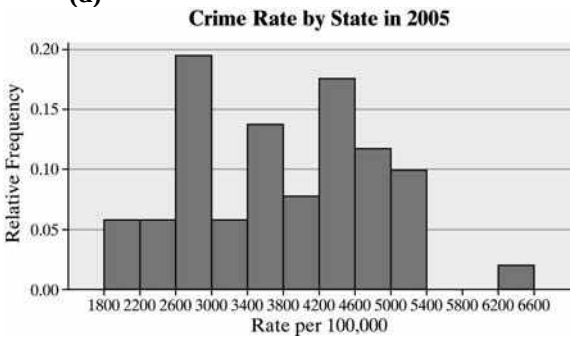
6. (a), (b)

Class	Freq.	Rel. Freq.
1800 – 2199	3	0.0588
2200 – 2599	3	0.0588
2600 – 2999	10	0.1961
3000 – 3399	3	0.0588
3400 – 3799	7	0.1373
3800 – 4199	4	0.0784
4200 – 4599	9	0.1765
4600 – 4999	6	0.1176
5000 – 5399	5	0.0980
5400 – 5799	0	0.0000
5800 – 6199	0	0.0000
6200 – 6599	1	0.0196

(c) The distribution is roughly symmetric.

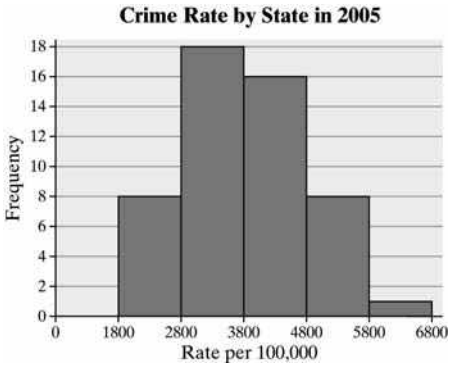


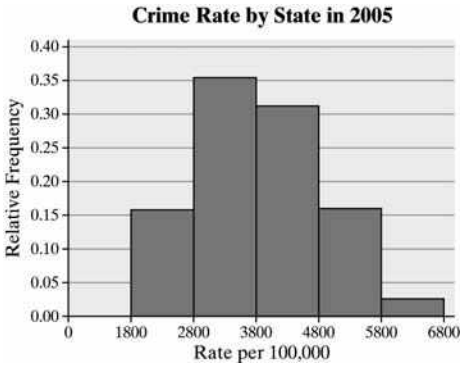
(d)



(e)

Class	Freq.	Rel. Freq.
1800 – 2799	8	0.1569
2800 – 3799	18	0.3529
3800 – 4799	16	0.3137
4800 – 5799	8	0.1569
5800 – 6799	1	0.0196





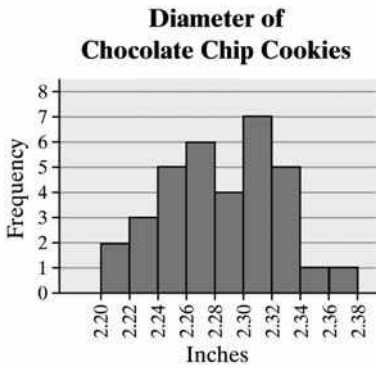
Answers will vary. Both class widths give a good overall picture of the distribution. The first class width provides a little more detail to the graph, but not necessarily enough to be worth the trouble. An intermediate value, say a width of 500, might be a reasonable compromise.

7. (a), (b)

Answers will vary. Using 2.2000 as the lower class limit of the first class and 0.0200 as the class width, we obtain the following.

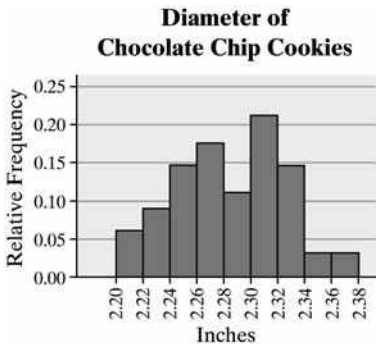
Class	Freq.	Rel. Freq.
2.2000 – 2.2199	2	0.0588
2.2200 – 2.2399	3	0.0882
2.2400 – 2.2599	5	0.1471
2.2600 – 2.2799	6	0.1765
2.2800 – 2.2999	4	0.1176
2.3000 – 2.3199	7	0.2059
2.3200 – 2.3399	5	0.1471
2.3400 – 2.3599	1	0.0294
2.3600 – 2.3799	1	0.0294

(e)



The distribution is roughly symmetric.

(f)



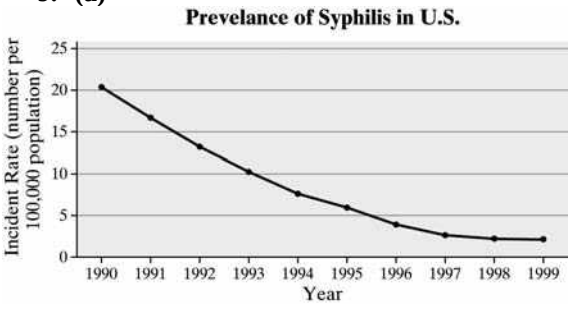
8. **Hours Spent Online**

12	
13	467
14	05578
15	1236
16	456
17	113449
18	066889
19	2
20	168
21	119
22	29
23	48
24	4
25	7
26	

Legend: 13 | 4 = average 13.4 hours per week.

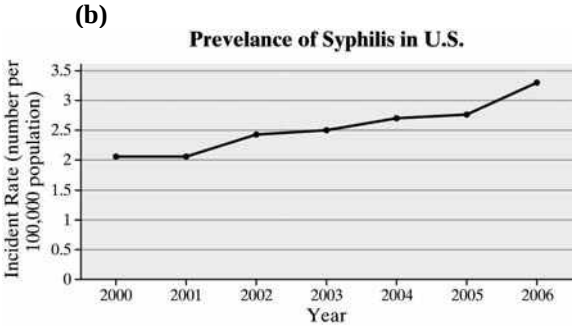
The distribution is slightly skewed right.

9. (a)



The incident rate decreased dramatically over the given time period.

(b)

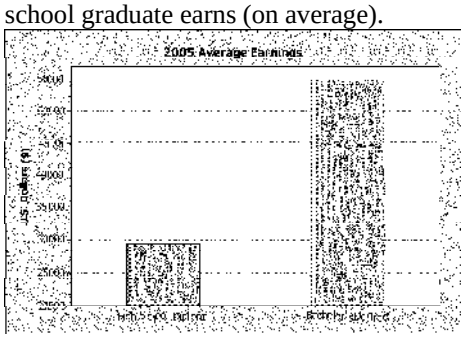


(c) Although the incident rate declined between 1990 and 2000, it increases gradually between 2002 and 2006. The plan needed to be adjusted to deal with this increase.

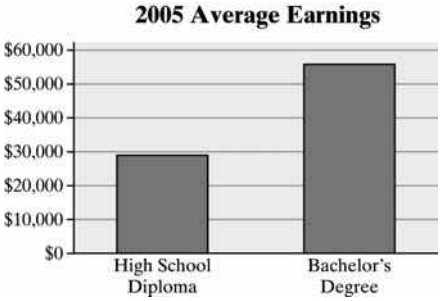
(d) No; a histogram will not allow the observer to see trends over time, only the distribution of yearly incident rates. To see a trend over time, a time series plot should be constructed.

10. The graph is misleading because there is no vertical scale.

11. **(a)** Graphs will vary. One way to mislead would be to start the vertical scale at a value other than 0. For example, starting the vertical scale at $20,000 might make the reader believe that college graduates earn more than three times what a high school graduate earns (on average).



(b) A graph that does not mislead would use equal widths for the bars and would start the vertical scale at $0.



12. **(a)** Flats are preferred the most (40%) and extra-high heels are preferred the least (1%).

(b) The graph is misleading because the bar heights and areas for each category are not proportional.

Chapter 2 Test

1. **(a)** The United States won the most men's singles championships between 1968 and 2007 with 15 wins.

(b) $6 - 4 = 2$
Representatives from Australia have won 2 more championships than representatives from Germany.

(c) $15 + 7 + 6 + 5 + 4 + 1 + 1 + 1 = 40$ championships between 1968 and 2007.
$\frac{7}{40} = 0.175$
Representatives of Sweden won 17.5% of the championships between 1968 and 2007.

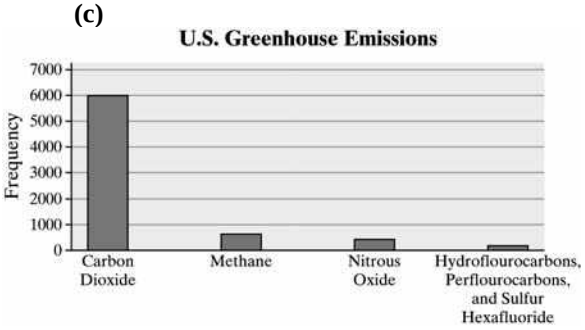
(d) No, it is not appropriate to describe the shape of the distribution as skewed right. The data represented by the graph are qualitative so the bars in the graph could be placed in any order.

2. **(a)** Total emissions = 5,934.4 + 605.1 + 378.6 + 157.6 = 7,075.7 million metric tons
Relative frequency of carbon dioxide = 5934.4/7075.7 ≈ 0.8387, and so on.

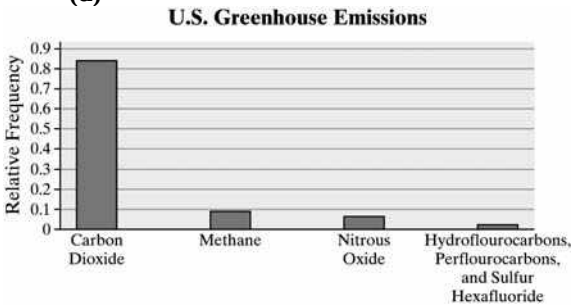
Gas	Relative Frequency
Carbon Dioxide	0.8387
Methane	0.0855
Nitrous Oxide	0.0535
Hydrofluorocarbons, perfluorocarbons, and sulfur hexafluoride	0.0223

(b) The relative frequency is 0.8387 so the percentage of emissions due to carbon dioxide is 83.87%.

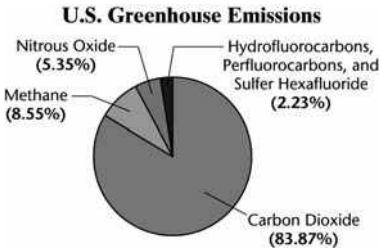
(c)



(d)



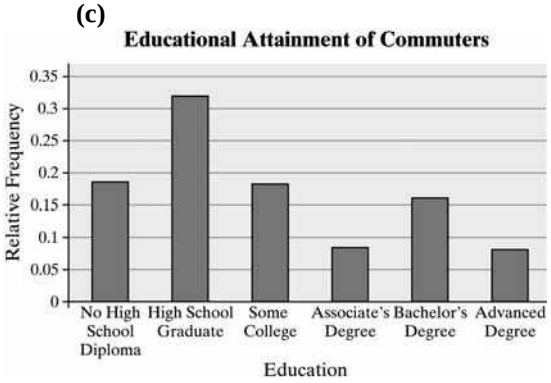
(e)



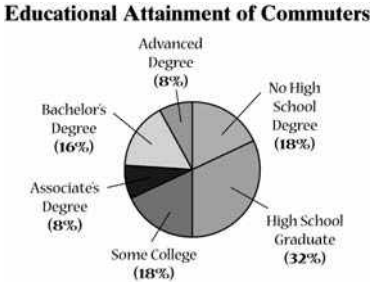
3. (a), (b)

Education	Freq.	Rel. Freq.
No high school diploma	9	0.18
High school graduate	16	0.32
Some college	9	0.18
Associate's degree	4	0.08
Bachelor's degree	8	0.16
Advanced degree	4	0.08

(c)



(d)

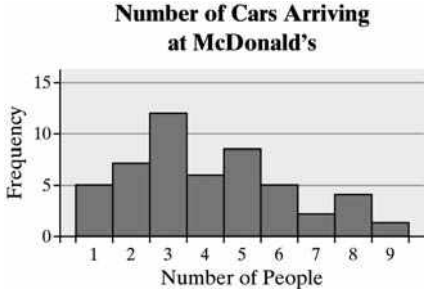


(e) The largest bar (and largest pie segment) corresponds to 'High School Graduate', so high school graduate is the most common educational level of a commuter.

4. (a), (b)

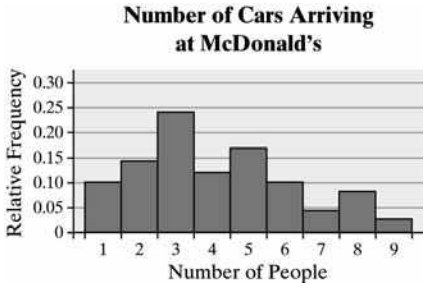
No. of Cars	Freq.	Rel. Freq.
1	5	0.10
2	7	0.14
3	12	0.24
4	6	0.12
5	8	0.16
6	5	0.10
7	2	0.04
8	4	0.08
9	1	0.02

(c)



The distribution is skewed right.

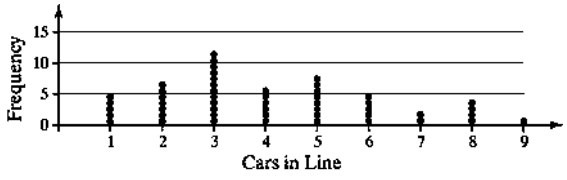
(d)



(e) The relative frequency of exactly 3 cars is 0.24. So, for 24% of the weeks, exactly three cars arrived between 11:50am and 12:00 noon.

(f) The relative frequency of 3 or more cars $= 1 -$ the relative frequency of 1 or 2 cars $= 1 - (0.10 + 0.14) = 0.76$. So, for 76% of the weeks, three or more cars arrived between 11:50am and 12:00 noon.

(g)



5. Answers may vary. One possibility follows:

(a) We can determine a class width by subtracting the smallest value from the largest, dividing by the desired number of classes, then rounding up. For example,

$$\frac{459-100}{8} = \frac{359}{8} = 44.875 \rightarrow 50$$

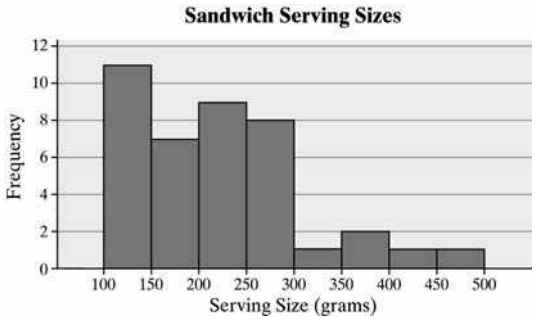
Our first lower class limit should be a nice number below the smallest data value. In this case, 100 is a good first lower limit. Thus, we will have a class width of 50 grams and the first class will have a lower limit of 100 grams.

(b), (c)

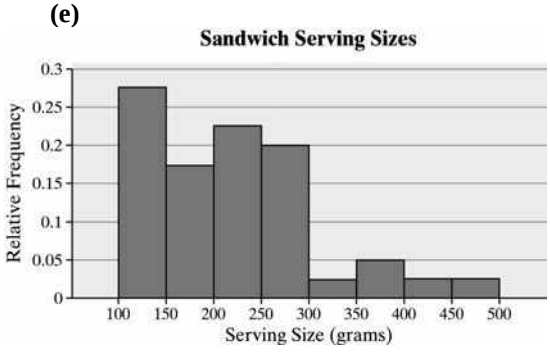
Relative frequency for 100-149 = 11/40 = 0.275 and so on.

Volume	Freq.	Rel. Freq.
100 – 149	11	0.275
150 – 199	7	0.175
200 – 249	9	0.225
250 – 299	8	0.200
300 – 349	1	0.025
350 – 399	2	0.050
400 – 449	1	0.025
450 – 499	1	0.025

(d) The distribution is skewed right.



(e)



6. (a) Rounded values:

67	67	68	57	78	65	68	64
77	59	69	62	53	75	79	76
82	67	73	69	64	66	94	
72	72	67	76	75	61	52	
72	74	71	68	61	57	66	
70	81	55	70	69	70	65	
59	67	64	75	69	78	59	

Fertility Rates by State, 2006

```
5 | 23
5 | 577999
6 | 112444
6 | 5566777778889999
7 | 000122234
7 | 555667889
8 | 12
8 |
9 | 4
```

Legend: 5 | 2 represents 52 births per 1000 women aged 15 to 44.

(b) The distribution is fairly symmetric, but does appear to be slightly skewed right.

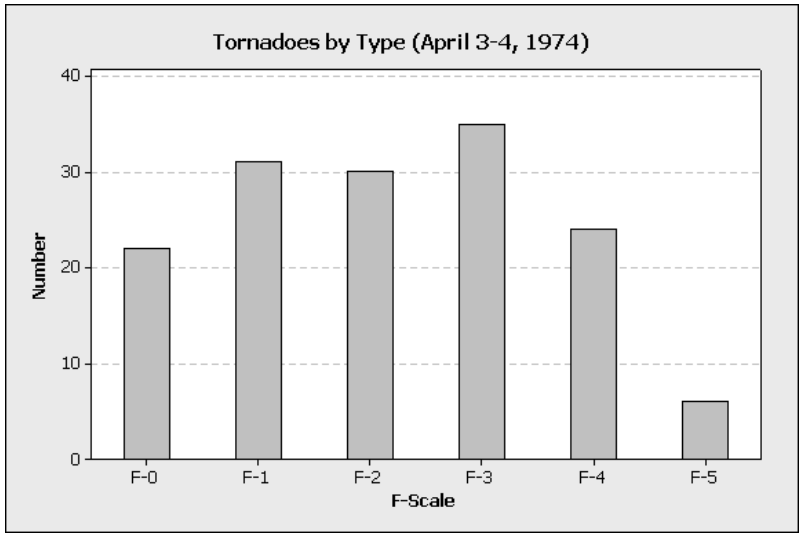
7.

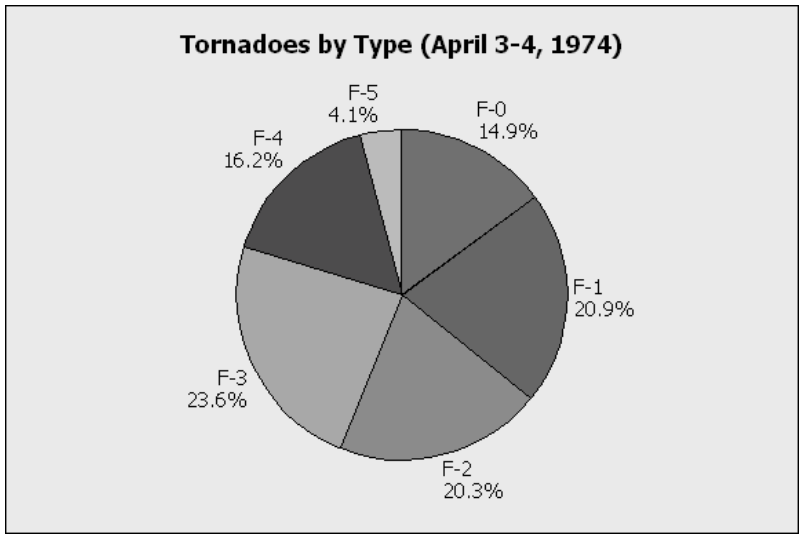
The price of milk decreased slightly from January 2006 to July 2006. It then increased slowly at first, then rapidly over the summer of 2007. Around late summer 2007 it appears that the price is leveled off.

8. Answers may vary. It is difficult to interpret this graph because it is not clear whether the scale is represented by the height of the steps, the width of the steps, or by the graphics above the steps. The graphics are misleading because they must be increased in size both vertically and horizontally to avoid distorting the image. Thus, the resulting areas are not proportionally correct. The graph could be redrawn using bars whose widths are the same and whose heights are proportional based on the given percentages. The use of graphics should be avoided, or a standard size graphic representing a fixed value could be used and repeated as necessary to illustrate the given percentages.

Case Study: The Day the Sky Roared

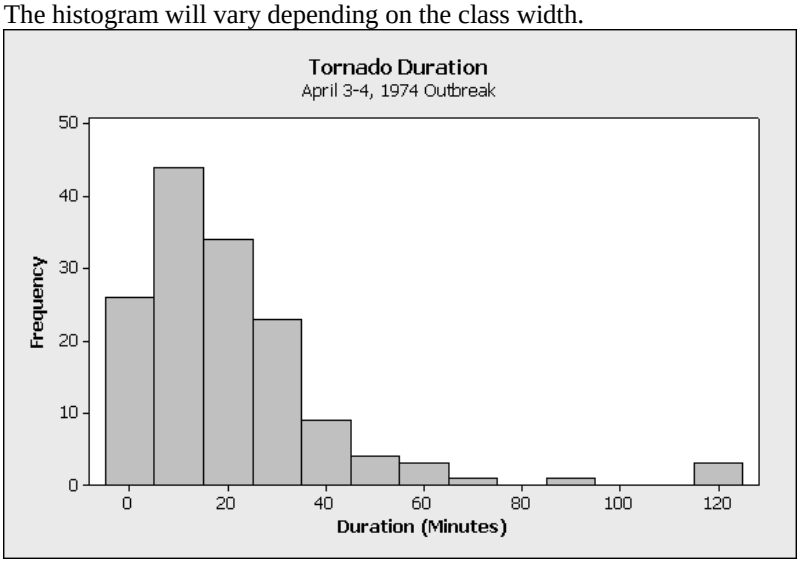
1.



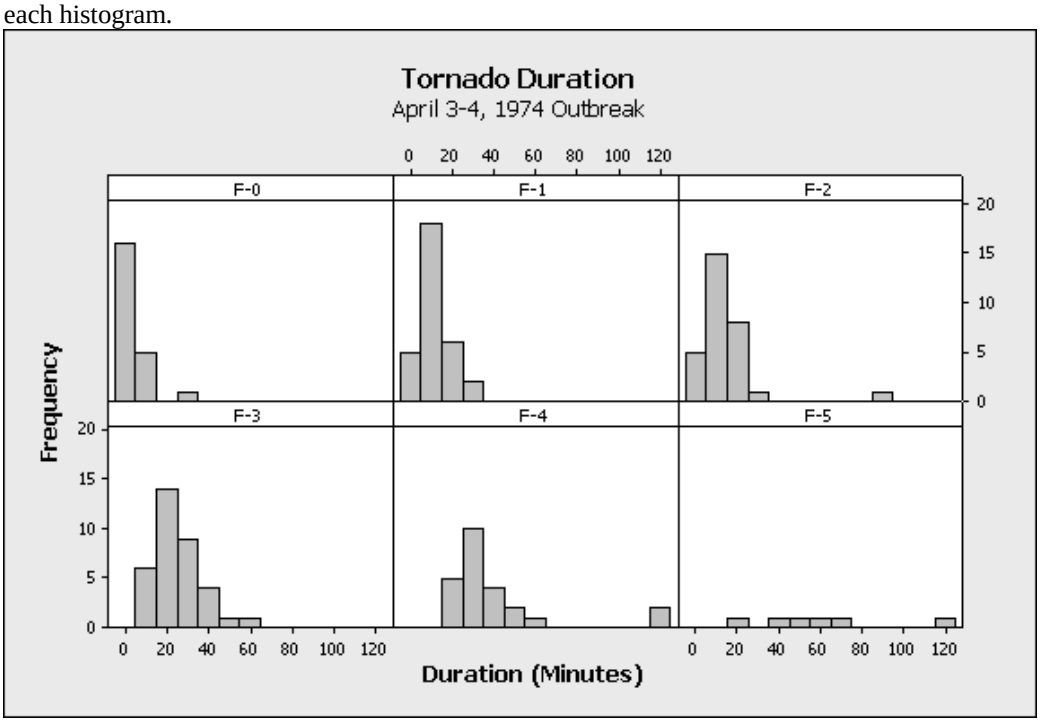


During the April 3-4, 1974 outbreak, 20% of the tornadoes exceeded F-3 on the Fujita Wind Damage Scale. This was much greater than the 1% that typically occurs.

2. The histogram will vary depending on the class width.



3. Histograms may vary depending on class widths. For comparison purposes, the same class width was used for each histogram.

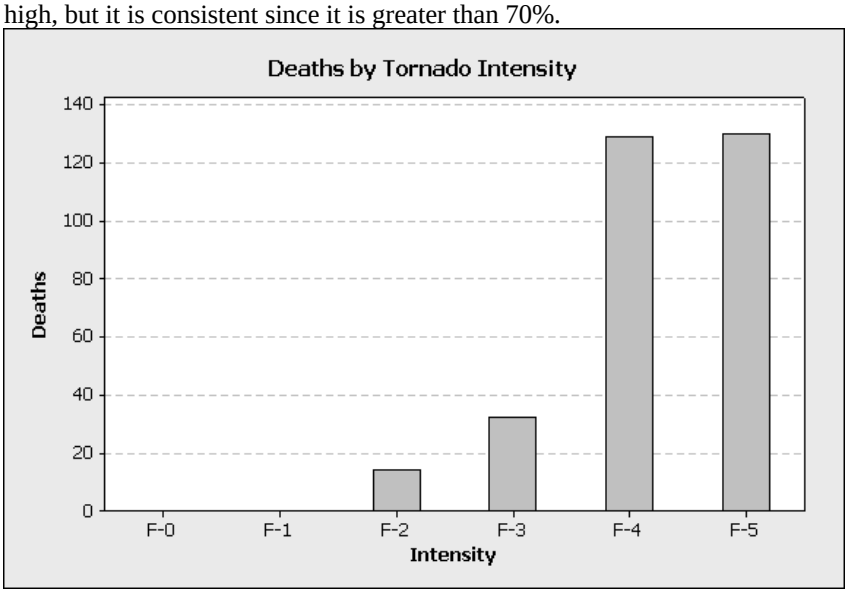


The distributions all appear to be skewed right, though the distribution for F-5 tornadoes is difficult to see due to the low sample size. There is an obvious shift in the distributions. As the strength of the tornado increases, the duration of the tornado increases.

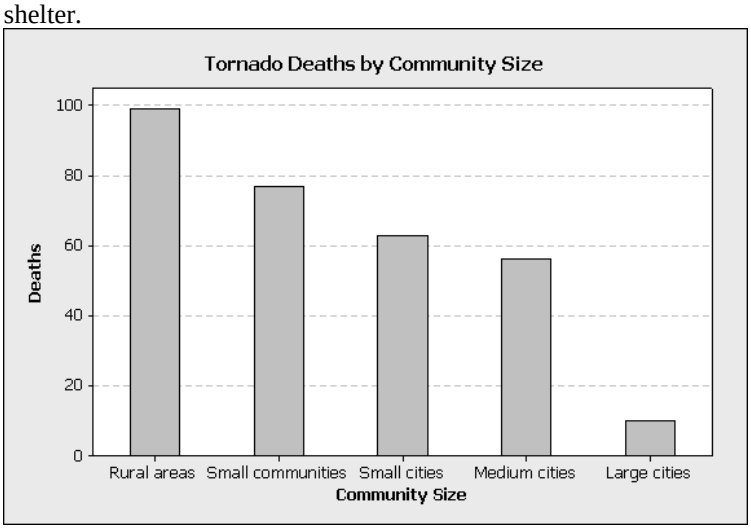
4. There were 305 deaths during the outbreak. Of these, 259 were due to the more severe tornadoes.

$$\frac{259}{305} \approx 0.8492$$

Roughly 85% of the deaths during the outbreak were due to the more severe tornadoes. This may be a little high, but it is consistent since it is greater than 70%.



5. The provided data is not sufficient to determine whether or not tornadoes are more likely to strike rural areas. Some research at Texas A&M University indicates that tornadoes are more likely to occur in urban or suburban areas, possibly due to greater temperature differences. The data does indicate that the number of deaths decreases as the population of the community increases. The higher the population density, the greater the chance that a tornado is detected and reported early, thereby providing more time for residents to take shelter.



6. Answers will vary. The outbreak of April 3-4, 1974 seemed to be more severe in intensity than usual with 20% of the tornadoes being classified as F-4 or F-5. While the shape of the duration distribution was roughly the same for each intensity level, the duration of a tornado increased with its intensity. The number of deaths decreased as the community size increased.