Chapter 2
Exploring Data with Graphs and Numerical Summaries

SECTION 2.1: PRACTICING THE BASICS

2.1 Categorical/quantitative difference
  a) Categorical variables are those in which observations belong to one of a set of categories, whereas quantitative variables are those on which observations are numerical.
  b) An example of a categorical variable is religion. An example of a quantitative variable is temperature.

2.2   U.S. married-couple households
  The variable summarized is categorical. The variable is type of U.S. married-couple households, and there are four types: traditional, dual-income with children, dual-income with no children, and other. These types are the categories.

2.3   Identify the variable type
  a) quantitative
  b) categorical
  c) categorical
  d) quantitative

2.4   Categorical or quantitative?
  a) categorical
  b) quantitative
  c) categorical
  d) quantitative

2.5   Discrete/continuous
  a) A discrete variable is a quantitative variable for which the possible values are separate values such as 0, 1, 2, …. A continuous variable is a quantitative variable for which the possible values form an interval.
  b) Example of a discrete variable: the number of children in a family (a given family can't have 2.43 children).
      Example of a continuous variable: temperature (we <u>can</u> have a temperature of 48.659).

2.6   Discrete or continuous?
  a) continuous
  b) discrete
  c) continuous
  d) discrete

2.7   Discrete or continuous 2
  a) continuous
  b) discrete
  c) discrete
  d) continuous

2.8 **Number of children**
a) The variable, number of children, is quantitative.
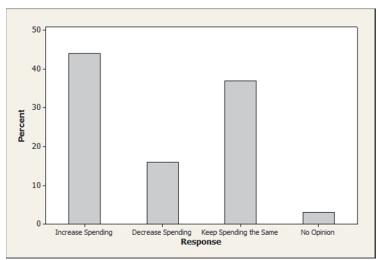b) The variable, number of children, is discrete.
c)

No. children	0	1	2	3	4	5	6	7	8+
Count	521	323	524	344	160	77	30	19	22
Proportion	0.258	0.160	0.259	0.170	0.079	0.038	0.015	0.009	0.011
Percentage	25.8	16.0	25.9	17.0	7.9	3.8	1.5	0.9	1.1

d) The mode – the most common score – is 2.

SECTION 2.2: PRACTICING THE BASICS

2.9 **Federal spending on financial aid**
a)

b) It's much easier to sketch bar charts relatively accurately.
c) The advantage of using one of these visual displays to summarize the results is that we can get a better sense of the data when we can see the sizes of the various categories as opposed to just reading the numbers.

2.10 **What do alligators eat?**
a) Primary food choice is categorical.
b) The modal category is "fish."
c) Approximately 43% of alligators ate fish as their primary food choice.
d) This is an example of a Pareto chart, a chart that is organized from most to least frequent choice.

2.11 **Weather stations**
a) The slices of the pie portray categories of a variable (i.e., regions).
b) The first number is the frequency, the number of weather stations in a given region. The second number is the percentage of all weather stations that are in this region.
c) It is easier to identify the modal category using a bar graph than using a pie chart because we can more easily compare the heights of bars than the slices of a piece of pie. For example, in this case, the slices for Midwest and West look very similar in size, but it would be clear from a bar graph that West was taller in height than Midwest.
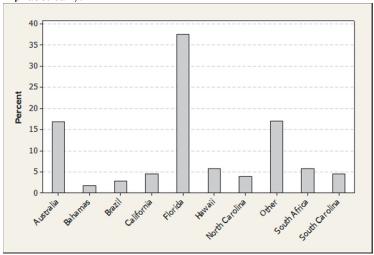
2.12   France is most popular holiday spot
a)   Country visited is categorical.
b)   A Pareto chart would make more sense because it allows the viewer to easily locate the categories with the highest and lowest frequencies.
c)   A dot plot or stem-and-leaf plot do not make sense because the data are categorical; these two types of plots are used with quantitative data (and also with data that have relatively few observations).
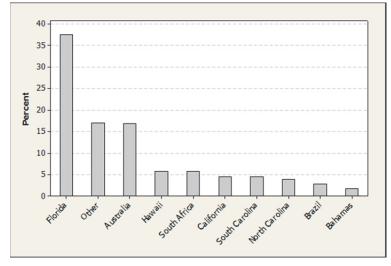
2.13   Shark attacks worldwide
The following charts use percentages.
(i) Alphabetically:

(ii) Pareto chart:

The Pareto chart is more useful than the chart arranged alphabetically because we can easily compare regions and see what outcomes occurred most frequently.

2.14   Sugar dot plot
a)   The minimum sugar value is zero grams, and the maximum is 18 grams.
b)   The sugar outcome that occurs most frequently is called the mode. For this data set there are five modes: three, four, eleven, twelve and fourteen grams.
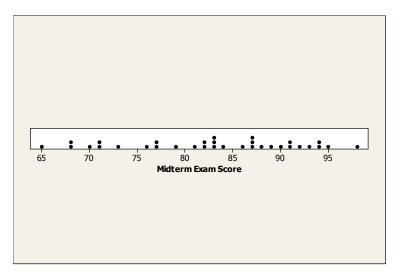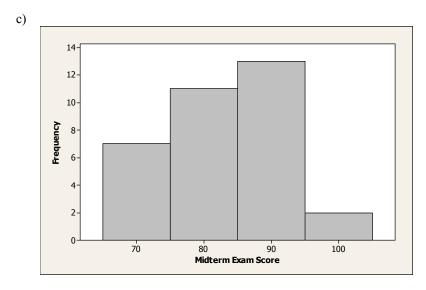
2.15   **Super Bowl tickets**
a)   2 | 123445889
     3 | 0356
     4 | 2
     5 | 0
     6 |
     7 | 5

b)   This plot gives us an overview of all the data. We see clearly that the amounts fall between $2100 and $7600, with more most fans spending between $2100 and $3700 on Super Bowl XLV tickets.

c)   2 | 12344
     2 | 5889
     3 | 03
     3 | 56
     4 | 2
     4 |
     5 | 0
     5 |
     6 |
     6 |
     7 |
     7 | 5

In the second plot, we are able to see the long right tail of the distribution of amounts more clearly.

2.16   **Graphing exam scores**
a)   There are 33 students in the class; the minimum score is 65 and the maximum is 98.
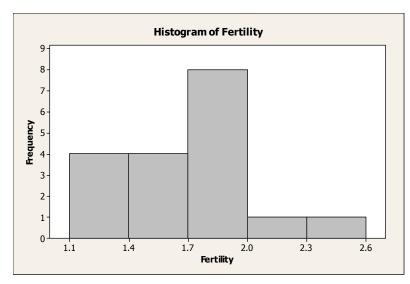b)

c)

2.17 **Fertility rates**
   a)   1 | 3333445677778899
       2 | 04
       A disadvantage of this plot is that it is too compact making it difficult to visualize where the data fall.
   b)   1 | 333344
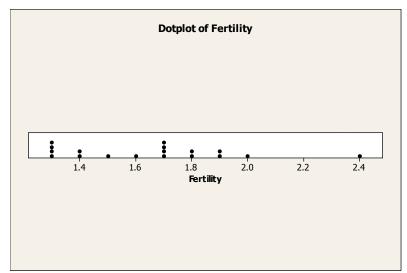   1 | 5677778899
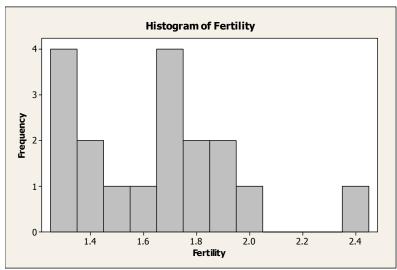   2 | 04
   c)

2.18 **Fertility plotted**
   a)   Since the leaf unit for this example is equal to 0.010, i.e. the hundredths place, the stem digits represent the ones and tenths. Thus, a stem of 13 with a leaf of 0 is to be read 1.30.
   b)   The largest value, 2.40, stands out from the others.

c)

As in the stem-and-leaf plot, the largest value, 2.40, stands out from the others.

d)

The value 2.4 stands out in the histogram as well since it is represented by a bar that is separated from the remaining cluster of bars.

2.19   Leaf unit
a)   The observations in the first row of the plot are 0 milligrams and 1000 milligrams.
b)   The sugar outcomes that occur most frequently are 3000 mg, 4000 mg, 11,000 mg, 12,000 mg and 14,000 mg.

2.20   Truncated data and split stems
a)   The range of data goes from 0 to 34 because the last digit has been removed.
b)   The sodium value of Special K, 220, is the last value on the plot in the row 2 | 00112. The last digit, 2, refers to this value.
c)   We no longer see the original sodium count. We don't know, for example, if two cereals have sodium levels of exactly 200, or if they are different; among the many possibilities, one could be 201 and one could be 209. Also, the gaps between values appear less than they actually are. For example, an actual gap between 220 and 290 is shown as a gap between 22 and 29.
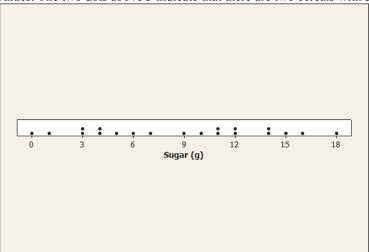
2.21 **Histogram for sugar**
  a) The intervals are 0 to 1, 1 to 3, 3 to 5, 5 to 7, 7 to 9, 9 to 11, 11 to 13, 13 to 15, 15 to 17 and 17 to 19.
  b) The distribution is bimodal. This is likely due to the fact that grams of sugar for both adult and child cereals are included. Child cereals, on average, have more sugar than adult cereals have.
  c) The dot and stem-and-leaf plots allow us to see all the individual data points; we cannot see the individual values in the histogram.
  d) The label of the vertical axis would change from Frequency to Percentage, but the relative differences among bars would remain the same. For example, the frequency for the lowest interval is 1; that would convert to 5% of all data points. The frequency for the third interval is 4, which would convert to 20%. In both cases, the second interval is four times as high as the first.

⌨**2.22** **Sugar plots**
  a) One can interpret this plot by looking at the dots above each amount of grams; the dots indicate observed values. The two dots above 3 indicate that there are two cereals with 3 grams of sugar.
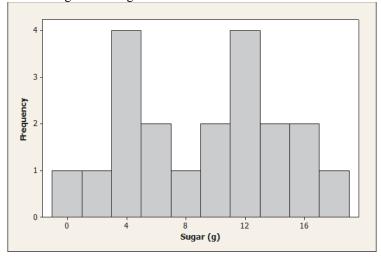
  b) One can interpret this plot by looking at the leaves next to each stem; the leaves indicate observed values. For example, there are a zero and two ones next to the first "one." This indicates that there was one cereal with 10 grams of sugar and two cereals with 11 grams of sugar.

```
Stem-and-leaf of SUGAR(g)   N = 20
Leaf Unit = 1.0
  2    0   01
  4    0   33
  7    0   445
  9    0   67
 10    0   9
 10    1   011
  7    1   22
  5    1   445
  2    1   6
  1    1   8
```

c)   One can interpret this histogram by looking at the frequencies above each interval. If one looks at the frequency above the number 4, one can see that there are four cereals that have at least 3 grams of sugar but less than 5 grams of sugar

2.23   Shape of the histogram
  a)   Assessed value houses in a large city – skewed to the right (a long right tail) because of some very expensive homes
  b)   Number of times checking account overdrawn in the past year for the faculty at the local university – skewed to the right because of the few faculty who overdraw frequently
  c)   IQ for the general population – symmetric because most would be in the middle, with some higher and some lower; there's no reason to expect more to be higher or lower (particularly because IQ is constructed as a comparison to the general population's "norms")
  d)   The height of female college students – symmetric because most would fall in the middle, going down to a few short students and up to a few tall students

2.24   More shapes of histograms
  a)   The scores of students (out of 100 points) on a very easy exam in which most score perfectly or nearly so, but a few score very poorly – skewed to the left because of the few who score poorly
  b)   The weekly church contribution for all members of a congregation, in which the three wealthiest members contribute generously each week – skewed to the right because of the few wealthy members' contributions
  c)   Time needed to complete a difficult exam (maximum time is 1 hour) – skewed to the left because most take almost or all of the whole time, whereas a few finish very quickly
  d)   Number of music CD's (compact discs) owned, for each student in your school – skewed to the right because of a few students' huge CD collections

2.25   How often do students read the newspaper?
  a)   This is a discrete variable because the value for each person would be a whole number. One could not read a newspaper 5.76 times per week, for example.
  b)   (i)   The minimum response is zero.
       (ii) The maximum response is nine.
       (iii) Two students did not read the newspaper at all.
       (iv) The mode is three.
  c)   This distribution is unimodal and somewhat skewed to the right.

2.26   Blossom widths
  a)   The distribution is skewed to the right; that is, it has a long right tail.
  b)   16/50 = 0.32. 32% of the blossoms in the data set have a petal width of more than 0.25 cm.
  c)   It is not possible to accurately determine this percentage since 0.3 cm is the midpoint of an interval not an endpoint. We are only able to accurately determine the percentage greater than or less than a value when the value is an endpoint of an interval.
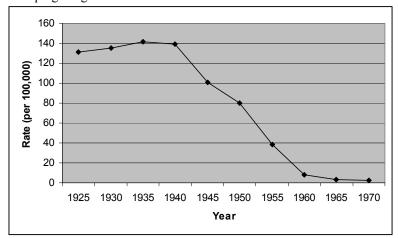
2.27 **Central Park temperatures**
  a) The distribution is somewhat skewed to the left; that is, it has a tail to the left.
  b) A time plot connects the data points over time in order to show time trends – typically increases or decreases over time. We cannot see these changes over time in a histogram.
  c) A histogram shows us the number of observations at each level; it is more difficult to see how many years had a given average temperature in a time plot. We also can see the shape of the distribution of temperatures from the histogram but not from the time plot.

2.28 **Is whooping cough close to being eradicated?**
  a) One can see in the time plot below that after an initial slight increase, there was a sharp and steady decrease in incidence of whooping cough starting around 1940. The decrease leveled off starting around 1960. These data suggest that the whooping cough vaccination was proving effective in reducing the incidence of whooping cough.
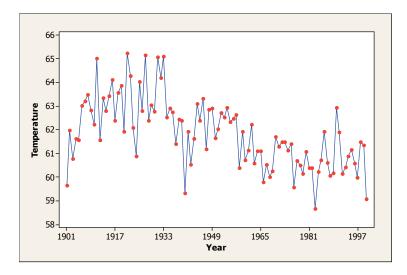
  b) Since 1993, the incidence rate has been about 2. This is similar to the incidence rate of 2.7 in 1970. It seems that the U.S. has diminished the incidence of whooping cough greatly, but has reached something of a plateau in the quest toward eradication.
  c) A histogram would not address this question because it does not show the rates for each year; we would not be able to see changes over time.

2.29 **Warming in Newnan, GA?**
  Overall, the time plot (below) does seem to show a decrease in temperature over time.

SECTION 2.3: PRACTICING THE BASICS

2.30   Median versus mean
   a)   The median is preferred when a distribution is highly skewed. (An outlier pulls the mean in that direction, but the size of the outlier does not affect the median). The median better represents what is typical. An example is annual income in most societies.
   b)   The mean is preferred when a distribution is close to symmetric or only slightly skewed, or if the variable is discrete with only a few distinct values. An example is number of children in a family.

2.31   More on emissions
   a)   Mean: $\bar{x} = \dfrac{\sum x}{n} = (6534 + 5833 + 1729 + 1495 + 1214 + 829 + 574)/7 = 2601.1$

   Median: Find the middle value [(n+1)/2 = (7+1)/2 = 4]
       574   829   1214   1495   1729   5833   6534
       The median is 1495.
   b)   Some of these countries, including the U.S. and China, have much larger populations than others. One would expect total CO_2 emissions to increase as population size increased. It would be more useful to know the level of CO_2 emissions per person.

2.32   Resistance to an outlier
   a)   The median for all three data sets is ten. The values for all three sets of observations are already arranged in numerical order, and the middle number for each is 10.
   b) $\bar{x} = \dfrac{\sum x}{n}$

   Set 1: $\bar{x}$ =(8+9+10+11+12)/5=10
   Set 2: $\bar{x}$ =(8+9+10+11+100)/5=27.6
   Set 3: $\bar{x}$ =(8+9+10+11+1000)/5=207.6
   c)   As the highest value becomes more and more of an extreme outlier, the median is unaffected, whereas the mean gets higher and higher as the data become more skewed.

2.33   Income and race
   For both blacks and whites, the income distributions are skewed to the right. When this happens, the means are higher than are the medians, because they are affected by the size of the outliers (a few very rich individuals).

2.34   Labor dispute
   Management would want to use the mean because it would be skewed right by the outliers – the few members of management who make a whole lot of money. The mean income would be higher because of the outliers. The workers would prefer the median because it is not affected by the large outliers. It is a more accurate measure of the actual typical income.

2.35   Cereal sodium
   The sodium value of zero is an outlier. It causes the mean to be skewed to the left, but its actual value does not affect the median.

2.36   Center of plots
   a)   The mean and median would be the same for the dot plots to the middle and to the right because the distributions are symmetric.
   b)   The distribution to the left is skewed to the right, and the mean would be higher than the median would. The mean would be pulled toward the higher, atypical values.

2.37   **Public transportation – center**
   a)   The mean is 2, the median is 0, and the mode is zero. Thus, the average score is 2, the middle score is zero
        (indicating that the mean is skewed by outliers), and the most common score also is zero.

   Mean: $\bar{x} = \dfrac{\sum x}{n}$ =(0+ 0+4+ 0+ 0+0+10+0+6+0)/10=2

   Median: middle score of 0, 0, 0, 0, 0, 0, 0, 4, 6, and 10
   Mode: the most common score is zero.
   b)   Now the mean is 10, but the median is still 0.

   Mean: $\bar{x} = \dfrac{\sum x}{n}$ =(0+ 0+4+ 0+ 0+0+10+0+6+0+90)/11=2

   Median: middle score of 0, 0, 0, 0, 0, 0, 0, 4, 6, 10, and 90
   The median is not affected by the magnitude of the highest score, the outlier. Because there are so many
   zeros, even though we've added one score, the median remains zero. The mean, however, is affected by the
   magnitude of this new score, an extreme outlier.

⌨**2.38** **Public transportation - outlier**
   a)   The *mean versus median* applet confirm that the median is not affected by the magnitude of the highest
        score. Because there are so many zeros, even though we've added one score, the median remains zero. The
        mean, however, is affected by the magnitude of this new score, an extreme outlier.
   b)   The applet demonstrates that the outlier has a weaker effect when there are more scores near the original
        mean.

2.39   **Student hospital costs**
   a)   The vast majority of students did not have any hospital stays last year. Thus, the most common score
        (mode), and the middle score (median), will be zero. There will be several students with hospital stays,
        however, and there likely will be a few with very expensive hospital stays. The money spent by these
        students will make the mean positive.
   b)   Another variable that would have this property is the number of aquarium fish in one's home. Most people
        don't have any, but there are a few people with fish-filled tanks. Another example is the number of times
        arrested in the previous year.

2.40   **Net worth by degree**
   a)   Some people who make a huge amount of money inflate the mean; these inflated values (outliers) do not
        affect the median.
   b)   The median is a much more realistic measure of a typical net worth because it is not affected by the inflated
        incomes of relatively few people.

2.41   **Canadian income**
   This distribution is skewed to the right. The fact that the mean is higher than the median indicates that there are
   extremely high incomes that are affecting the mean, but not the median.

2.42   **Baseball salaries**
   There are a few valuable players who receive exorbitant salaries, whereas the typical player is paid much less
   (although still a lot by most people's standards!). The very high salaries of the few affect the mean, but not the
   median.

2.43 **European fertility**
  a) The median fertility rate is 1.7. Thus, about half of the countries listed have mean fertility rates at or below 1.7 with the remaining countries having fertility rates above 1.7.
  b) The mean of the fertility rates is 1.65.
  c) Since the population of adult women can vary greatly among the countries, it is necessary to calculate an overall fertility rate for the country in order to make comparisons. This rate is found by calculating the mean number of children per adult woman. The mean for a variable need not be one of the possible values for the variable. Although the number of children born to each adult woman is a whole number, the mean number of children born per adult woman need not be a whole number. For example, the mean number of children per adult woman is considerably higher in Mexico than in Canada.

2.44 **Sex partners**
  a)

Number of partners	Number of respondents
0	102
1	233
2	18
3	9
4	2
5	1
Total	365

  If the data is sorted from smallest to largest, the median is the number in the 183rd position. Since 102 respondents answered 0 and 233 answered 1, the median is 1.

  b) Mean: $\bar{x} = \dfrac{\sum x_i}{n} = \left(0*102 + 1*233 + 2*18 + 3*9 + 4*2 + 5*1\right)/365 = 0.85$.

  c) Since the total number of respondents is 365, the median is still the value in the 183rd place when the data are sorted from smallest to largest. Since 233 respondents gave an answer of 1, the median is still 1. However, the value of the mean changes:

  Mean: $\bar{x} = \dfrac{\sum x_i}{n} = \left(1*233 + 2*18 + 3*9 + 4*2 + 5*103\right)/365 = 2.2$.

2.45 **Knowing homicide victims**
  a) The mean is 0.16.
  $$\sum x/n = (3944(0) + 279(1) + 97(2) + 40(3) + 23(4.5))/4383 = 0.16.$$
  b) The median is the middle score. With 4383 scores, the median is the score in the 2192nd position. Thus, the median is 0.
  c) The median would still be 0, because there are still 2200 people who gave 0 as a response. The mean would now be 1.95.
  $$\sum x/n = (2200(0) + 279(1) + 97(2) + 40(3) + 1767(4.5))/4383 = 1.95.$$
  d) The median is the same for both because the median ignores much of the data. The data are highly discrete; hence, a high proportion of the data falls at only one or two values. The mean is better in this case because it uses the numerical values of all of the observations, not just the ordering.

2.46 **Accidents**
  In this case, the mean is likely to be more useful because it uses the numerical values of all of the observations, not just the ordering. Because so many people would report 0 motor accidents, the median is not very useful. It ignores too much of the data.

SECTION 2.4: PRACTICING THE BASICS

2.47   Sick leave
   a)   The range is six; this is the distance from the smallest to the largest observation. In this case, there are six days separating the fewest and most sick days taken (6-0=6).
   b)   The standard deviation is the typical distance of an observation from the mean (which is 1.25).

$$s^2 = \frac{\sum(x-\bar{x})^2}{n-1} = ((0\text{-}1.25)^2+(0\text{-}1.25)^2+(0\text{-}1.25)^2+(0\text{-}1.25)^2+(0\text{-}1.25)^2+(0\text{-}1.25)^2+(4\text{-}1.25)^2+(6\text{-}1.25)^2)/7 = 39.5/7 = 5.64$$

$$s = \sqrt{s^2} = \sqrt{5.643} = 2.38$$

   The standard deviation of 2.38 indicates a typical number of sick days taken is 2.38 days from the mean of 1.25.
   c)   Redo (a) and (b)
      a.   The range is sixty; this is the distance from the smallest to the largest observation. In this case, there are sixty days separating the fewest and most sick days taken (60-0=60).
      b.   The standard deviation is the typical distance of an observation from the mean (which is 8).

$$s^2 = \frac{\sum(x-\bar{x})^2}{n-1} = ((0\text{-}8)^2+(0\text{-}8)^2+(0\text{-}8)^2+(0\text{-}8)^2+(0\text{-}8)^2+(0\text{-}8)^2+(4\text{-}8)^2+(60\text{-}8)^2)/7 = 3104/7 = 443.43$$

$$s = \sqrt{s^2} = \sqrt{443.43} = 21.06$$

   The standard deviation of 21.06 indicates a typical number of sick days taken is 21.06 days from the mean of 8.
   The range and mean both increase when an outlier is added.

2.48   Life expectancy
   a)   Upon examination of the data, the countries in Africa will have a larger standard deviation since the spread of the data is greater for this group than for the countries in Western Europe.
   b)   Western Europe:

$$\bar{x} = \frac{\sum x}{n} = \frac{\sum(77+77+\ldots+80)}{15} = 78.6667$$

$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} = \sqrt{\frac{(77-78.6667)^2+(77-78.6667)^2+\ldots+(80-78.6667)^2}{15-1}} = 1.047$$

   Africa:

$$\bar{x} = \frac{\sum x}{n} = \frac{\sum(37+37+\ldots+57)}{16} = 46.625$$

$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} = \sqrt{\frac{(37-46.625)^2+(37-46.625)^2+\ldots+(57-46.625)^2}{16-1}} = 7.136$$

   Note that the standard deviation for the Western Europe group, 1.0 (rounded), is much smaller than for the Africa group, 7.1.

2.49   Life expectancy including Russia
   We would expect the standard deviation to be larger since the value for Russia is significantly smaller than the rest of the group adding additional spread to the data. The standard deviation including Russia is, in fact, 3.56.

2.50 **Shape of home prices?**
The most plausible value is $60,000. -$15,000 is not possible because a standard deviation cannot be negative. $1,000 and $1,000,000 are unlikely because they are too small or too big, respectively, for a typical deviation. One would not expect the typical deviation to be that far from the median for home prices.

2.51 **Exam standard deviation**
The most realistic value is 12. There are problems with all the others.
-10: We can't have a negative standard deviation.
0: We know that there is spread because the scores ranged from 35 to 98, so the standard deviation is not 0.
3: This standard deviation seems very small for this range.
63: This standard deviation is too large for a typical deviation. In fact, <u>no</u> score differed from the mean by this much.

2.52 **Heights**
a) According to the Empirical Rule, 68% of men would be within one standard deviation of the mean – between 68 and 74 inches. 95% of men would be within two standard deviations of the mean – between 65 and 77 inches. All or nearly all men would be within three standard deviations of the mean – between 62 and 80 inches.
b) The mean for women is lower than the mean for men. This indicates that the mean for women is lower than the mean for all students, and the mean for men is higher than the mean for all students. Because women's heights would tend to be closer to the mean for women than to the overall mean, the standard deviation would be smaller when we compared them with the appropriate gender group than when we compared them to the overall group. Similarly, men's heights would tend be closer to the mean for men than to the overall mean.

2.53 **Histograms and standard deviation**
a) The sample on the right has the largest standard deviation, and the sample in the middle has the smallest.
b) The Empirical Rule is relevant only for the distribution on the left because the distribution is bell-shaped.

2.54 **Female strength**
According to the Empirical Rule, 68% of women would be able to lift within one standard deviation from the mean – between 66.6 and 93.2 pounds. 95% of women would be able to lift within two standard deviations from the mean – between 53.3 and 106.5 pounds. All or nearly all women would be able to lift within three standard deviations from the mean – between 40 and 119.8 pounds.

2.55 **Female body weight**
a) 95% of weights would fall within two standard deviations from the mean – between 99 and 167.
b) An athlete who is three standard deviations above the mean would weight 184 pounds. This would be an unusual observation because typically all or nearly all observations fall within three standard deviations from the mean. In a bell-shaped distribution, this would likely be about the highest score one would obtain.

2.56 **Shape of cigarettes taxes**
With a bell-shaped distribution, we expect scores to extend about three standard deviations from the mean in either direction. The lowest possible value, of 0, however, is only 1.52 standard deviations below the mean, and so the distribution likely is skewed to the right.

2.57   Empirical rule and skewed, highly discrete distribution
a) One standard deviation from the mean: between -0.21 and 0.53
    Two standard deviations from the mean: between -0.58 and 0.9
    Three standard deviations from the mean: between -0.95 and 1.27

	Observations	Predicted by Empirical Rule
One standard deviation from the mean	84.0%	68%
Two standard deviations from the mean	84.0%	95%
Three standard deviations from the mean	99.9%	About 100%

There are more observations within one standard deviation of the mean and fewer within two standard deviations than would be predicted by the Empirical Rule.
b) The Empirical Rule is only valid when used with data from a bell-shaped distribution. This is not a bell-shaped distribution; rather, it is highly skewed to the right. Most observations have a value of 0, and hardly any have the highest value of 2.
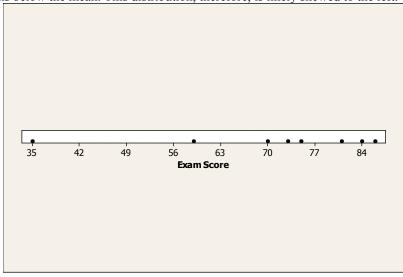
2.58   How much TV?
These statistics suggest that this distribution is highly skewed toward the right for two main reasons. The mean is larger than the median, and the standard deviation is almost as large as the mean. In fact, the lowest possible value of 0 is only 2.98/2.66 = 1.12 standard deviations below the mean.

2.59   How many friends?
a) The standard deviation is larger than the mean; in addition, the mean is higher than the median. In fact, the lowest possible value of 0 is only 7.4/11.0 = 0.67 standard deviations below the mean. These situations occur when the mean and standard deviation are affected by an outlier or outliers. It appears that this distribution is skewed to the right.
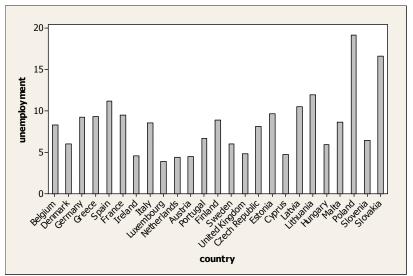b) The Empirical Rules does not apply to these data because they do not appear to be bell-shaped.

2.60   Judging skew using $\bar{x}$ and *s*
The largest observation is 86 is less than one standard deviation above the mean of 70.4. Specifically, 86 is only 15.6/16.7 = 0.93 standard deviations above the mean. The smallest observation is 35, 35.4/16.7 = 2.11 standard deviations below the mean. This distribution, therefore, is likely skewed to the left.

⌨2.61 EU data file
  a)

  b) Variable   StDev
      unemployment   3.695
      This is a typical deviation from the mean unemployment for EU nations.

⌨2.62 Create data with a standard deviation
  a) Many possible answers, such as 2, 2, 4, 4, 5, 5, 6, 6, 8, 8
  b) Many possible answers, such as 0, 0, 1, 2, 4, 6, 8, 9, 10, 10
  c) The largest standard deviation we can get is 5.27 when we have five 0's and five 10's.

SECTION 2.5: PRACTICING THE BASICS

2.63 Vacation days
  a) Median: Find the middle value of 13, 25, 26, 28, 34, 35, 37, 42
      The median is 31, the average of the two middle values, 28 and 34.
  b) The first quartile is the median of 13, 25, 26, and 28. The first quartile is 25.5, the average of the two middle values, 25 and 26.
  c) The third quartile is the median of 34, 35, 37, and 42. The third quartile is 36, the average of the two middle values.
  d) 25% of countries have residents who take fewer than 25.5 vacation days, half of countries have residents who take fewer than 31 vacation days, and 75% of countries have residents who take fewer than 36 vacation days per year. The middle 50% of countries have residents who take an average of between 25.5 and 36 vacation days annually.

⌨2.64 European unemployment
  a) The middle number is 6.7 (the 8th number out of 15). Thus, Italy has the median unemployment rate for these data. Half of the remaining countries have unemployment rates that are less than 6.7 and half have rates that are higher.
  b) The first quartile is the median of the observations below the median: 3.2, 3.6, 4.4, 4.5, 5, 5.4 and 6. It is 4.5 The third quartile is the median of the observations above the median: 7, 7.2, 7.7, 7.8, 8.4, 8.6 and 8.7. It is 7.8.
  c) Mean: $\bar{x} = \dfrac{\sum x}{n}$ =(3.2+3.6+4.4+4.5+5+5.4+6+6.7+7+7.2+7.7+7.8+8.4+8.6+8.7)/15=6.28; the typical score is 6.28.

2.65 **Female strength**
a) One fourth had maximum bench press less than 70 pounds, and one fourth had maximum bench press greater than 90 pounds.
b) The mean and median are the about the same, and the first and third quartiles are equidistant from the median. These are both indicators of a roughly symmetric distribution.

2.66 **Female body weight**
a) One quarter had weight below 119 and one quarter had weight above 144.
b) The mean and median are the about the same, and the first and third quartiles are approximately equidistant from the median. These are both indicators of a roughly symmetric distribution.

2.67 **Ways to measure variability**
a) The range is even more affected by an outlier than is the standard deviation. The standard deviation takes into account the values of all observations and not just the most extreme two.
b) With a very extreme outlier, the standard deviation will be affected both because the mean will be affected and because the deviation of the outlier (and its square) will be very large. The IQR would not be affected by such an outlier.
c) The standard deviation takes into account the values of all observations and not just the two marking 25% and 75% of observations.

2.68 **Variability of cigarette taxes**
a) (i) Q1, marking the lowest 25% of states, has a value of 36. Thus, 75% of states have cigarette taxes greater then 36 cents.
   (ii) Q3, marking the highest 25% of states, has a value of 100. Thus, 25% of states have cigarette taxes greater than $1.00.
b) The two values that demarcate the middle 50% are Q1 = 36 cents and Q3 = 100 cents (one dollar).
c) The interquartile range (IQR) is the difference between Q1 and Q3. IQR=Q3-Q1=100-36=64 cents. For the middle 50% of state cigarette taxes, $0.64 is the distance between the largest and smallest cigarette tax amount.
d) With a bell-shaped distribution, we expect Q1 and Q3 to be roughly equidistant from the median which is not the case here. The maximum value is also quite far from Q3. Thus, it appears that the distribution is skewed to the right.

2.69 **Sick leave**
a) The range is six; this is the distance from the smallest to the largest observation. In this case, there are six days separating the fewest and most sick days taken (6-0=6).
b) The interquartile range is the difference between Q3 and Q1. IQR=Q3-Q1=2.
c) Redo (a) and (b).
   a. The range is sixty; this is the distance from the smallest to the largest observation. In this case, there are sixty days separating the fewest and most sick days taken (60-0=60).
   b. Q1, the median of all scores below the median, is still 0. Q3, the median of all scores above the median, is still 2 (the average of 0 and 4). The interquartile range remains the same: IQR=Q3-Q1=2. The IQR is least affected by the outlier because it doesn't take the magnitudes of the two extreme scores into account at all, whereas the range and *s* do.

2.70 **Infant mortality Africa**
a) Q1 is the median of the lower half of the sorted data: 54, 63, 68, 76, 78, 79 and 80. It is 76. Q3 is the median of the upper half of the sorted data: 81, 84, 96, 101, 110, 121 and 154. It is 101.
b) IQR=Q3-Q1=25. For the middle half of the infant mortality rates, the distance between the largest and smallest rates is 25.

2.71 **Infant mortality Europe**
Q1 is the median of the lower half of the sorted data: 3, 3, 3, 4, 4, 4 and 4. It is 4. Q2, the median, is the average of the middle two data values (4+4)/2=4. Q3 is the median of the upper half of the sorted data: 4, 4, 4, 4, 5, 5 and 5. It is 4.

2.72   Computer use
a)   This five-number summary suggests that the distribution is skewed to the right. The distance between the minimum and the median is much smaller than the distance between the median and the maximum.
b)   In this case, outliers would be those values more than 1273.5 points from the first and third quartiles:
1.5(IQR)= 1.5 X 849=1273.5.
(IQR=Q3-Q1=1105-256=849)

The lower boundary: Q1-1.5(IQR)=256-1273.5= -1017.5
The upper boundary: Q3+1.5(IQR)=1105+1273.5=2378.5

In the current example, the lowest score is 4, so there are no scores below -1017.5. The highest score, on the other hand, is 320,000, much higher than 2378.5. Thus, there are potential outliers according to this criterion.
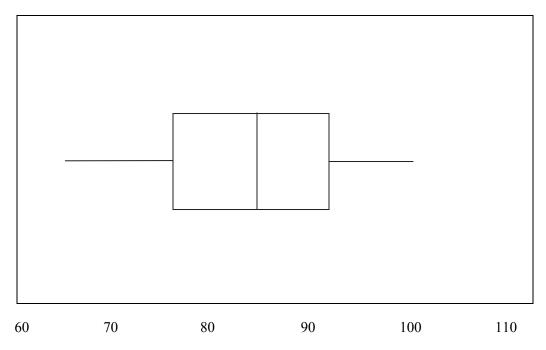
2.73   Central Park temperature distribution revisited
a)   We would expect it to be skewed to the left because the maximum is closer to the median than is the minimum.
b)   Numbers are approximate:
Minimum: 49.3
Q1: 52.8
Median: 53.85
Q3: 55.0
Maximum: 57.2
These approximations support the premise that the distribution is skewed to the left if it is skewed. The median is closer to the maximum and Q3 than it is to the minimum and Q1.
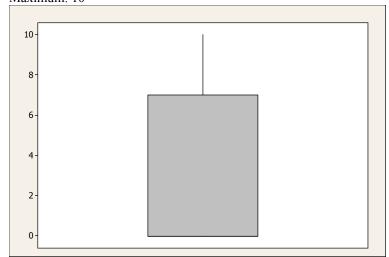
2.74   Box plot for easy exam

The minimum, Q1, median, Q3, and maximum are used in the box plot.

2.75   **Public transportation**
   a)   Minimum: 0
        Q1: 0
        Median: 0
        Q3: 4
        Maximum: 10

   b)   Q1 and the median share the same line in the box because so many employees have a score of zero that the middle score of the whole set of data is zero and the middle score of the lower half of the data also is zero.
   c)   There is no whisker because the minimum score also is zero. This situation resulted because there are so many people with the lowest score.

2.76   **Energy statistics**
   a)   Numbers are approximate:
        Minimum: 75
        Q1: 115
        Median: 160
        Q3: 180
        Maximum: 425
        One country was a potential outlier, the one around 425.
   b)   We can know how far Italy was from the mean in terms of standard deviations by calculating its z-score. It is 0.40 standard deviations below the mean.

$$z = \frac{x - \bar{x}}{s}, \; z = \frac{138.7 - 167.8}{72.8} = -0.40.$$

   c)   The U.S. is 2.28 standard deviations above the mean.

$$z = \frac{x - \bar{x}}{s}, \; z = \frac{334 - 167.8}{72.8} = 2.28.$$

2.77   **European Union unemployment rates**
   a)   In a box plot, Q1, 4.5, and Q3, 7.8, would be at the outer edges of the box. The whiskers would extend on the left to the minimum, 3.2, and on the right to the maximum, 8.7.
   b)   The score is 1.33 standard deviations above the mean, and thus, is not an outlier according to the three standard deviation criterion.

$$z = \frac{x - \bar{x}}{s}, z = \frac{8.7 - 6.3}{1.8} = 1.33.$$

   c)   A z-score of 0 indicates that the country's unemployment rate is zero standard deviations from the mean; hence, the unemployment rate is equal to the mean. In this case, a country with an unemployment rate of 6.3 would have a z-score of 0.

2.78   **Air pollution**
   a)   Canada is 2.28 standard deviations above the mean in terms of air pollution. In other words, their air pollution is comparatively high.

$$z = \frac{x - \bar{x}}{s}, \quad z = \frac{16.5 - 8.3}{3.6} = 2.28.$$

   b)   Sweden is -0.92 standard deviations above the mean in terms of air pollution. In other words, their air pollution is comparatively low.

$$z = \frac{x - \bar{x}}{s}, \quad z = \frac{5.0 - 8.3}{3.6} = -0.92.$$

2.79   **Female heights**
   a) $z = \dfrac{x - \bar{x}}{s}, \quad z = \dfrac{56 - 65.3}{3.0} = -3.1$.

   b)   The negative sign indicates that the height of 56 inches is below the mean.
   c)   Because the height of 56 inches is more than three standard deviations from the mean, it is a potential outlier.
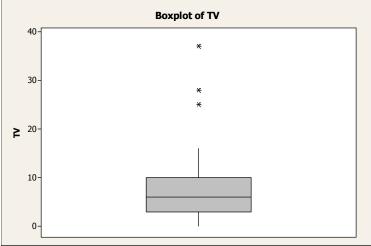
2.80   **Hamburger sales**
   The appropriate statistic here would be the *z*-score.

$$z = \frac{x - \bar{x}}{s} = \frac{2000 - 1165}{220} = 3.80.$$

   This z-score indicates that the sales for this day are more than three standard deviations above the mean, and thus, would be a potential outlier – in other words, an unusually good day.
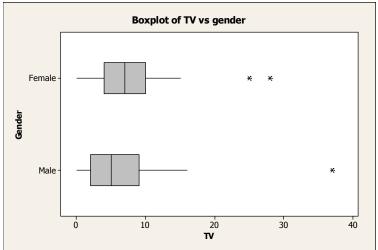
⌨2.81   **Florida students again**
   a)   The distribution depicted in the boxplot below is skewed to the right. The right whisker is longer, and there are a few outliers (shown by asterisks) representing large values.

   b)   The 1.5(IQR) criterion would indicate that all data should fall between about -7.5 and 20.5 (based on IQR = 10 – 3 = 7; 1.5(IQR)1.5 = 10.5). Because some data points fall beyond this range, it appears that there are potential outliers.

⌨**2.82** **Females or males watch more TV?**

Based on the Florida survey data, females tend to watch more TV. The median and Q1 and Q3 are higher for females than for males.

2.83 **CO2 comparison**
a) The outlier shown in the box plot for Europe is around 22.
b) Based on the box plot, we would predict the shape of the distribution of carbon dioxide emissions for South America to be skewed to the right since the distance between Q3 and Q2 is much larger than between Q1 and Q2 and the upper whisker is also much longer than the lower whisker.
c) Carbon dioxide emissions tend to be much higher in Europe than South America. The smallest level reported in Europe is close to the third quantile of the levels reported in South America. Thus, roughly 75% of the levels reported in South America are less than the smallest level reported in Europe.

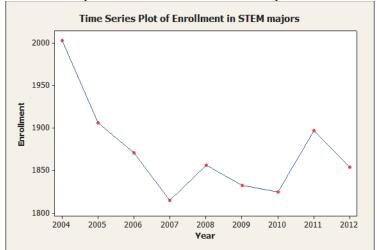# SECTION 2.6: PRACTICING THE BASICS

2.84 **Great pay (on the average)**
a) The mean is $43,700 and the median is $9300.
b) It is misleading because the mean is so heavily influenced by the outlier (her own salary) that it is not a typical value. The median would be a much more accurate summary of these salaries.
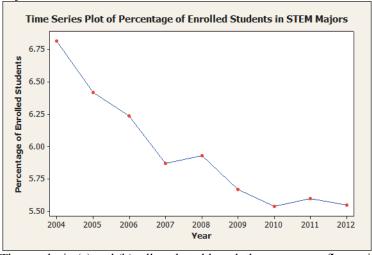
2.85 **Market share for food sales**
a) One problem with this chart is that the percentages do not add up to 100. Second, the Tesco slice seems too large for 27.2%. A third problem is that contiguous colors are very similar, particularly the two white sections. This increases the difficulty in easily reading this chart.
b) It would be easier to identify the mode with a bar graph because one would merely have to identify the highest bar.

⌨**2.86** **Enrollment trends**
a) This graph shows an overall decrease in enrollment in STEM majors at first, with what appears to be somewhat of a plateau toward the end of the time span.

b) This graph shows a gradual decrease over time in the percentage of students who are enrolled in STEM majors.

c) The graphs in (a) and (b) tell us that although there are some fluctuations in the numbers of students enrolling in STEM majors over the years, there is a steady decrease in the percentage of enrolling students who are enrolled in STEM majors over the years. We cannot learn this from Figures 2.18 and 2.19.

2.87 **Terrorism and war in Iraq**
a) This graph is misleading. Because the vertical axis does not start at 0, it appears that six times as many people are in the "no, not" column than in the "yes, safer" column, when really it's not even twice as many.
b) With a pie chart, the area of each slice represents the percentage who fall in that category. Therefore the relative sizes of the slices will always represent the relative percentages in each category.

2.88 **BBC license fee**
The 2013 projection is shown where the observation would be plotted for the year 2007, not 2013.

2.89 **Federal government spending**
The slices do not seem to have the correct sizes, for instance the slice with 16% seems larger than the slice with 19%.

2.90   Bad graph
The response to this exercise will be different for each student.

CHAPTER PROBLEMS: PRACTICING THE BASICS

2.91   Categorical or quantitative?
a)   Number of children in family: quantitative
b)   Amount of time in football game before first points scored: quantitative
c)   Choice of major (English, history, chemistry, …): categorical
d)   Preference for type of music (rock, jazz, classical, folk, other): categorical

2.92   Continuous or discrete?
a)   Age of mother: continuous
b)   Number of children in a family: discrete
c)   Cooking time for preparing dinner: continuous
d)   Latitude and longitude of a city: continuous
e)   Population size of a city: discrete

2.93   Immigration into U.S.
a)   "Place of birth" is categorical.
Percentages:

Europe	13.7%
Caribbean	9.6%
Central America	37.6%
South America	6.1%
Asia	25.4%
Other	7.6%
Total	100%

b)   These categories are not ordered. Thus, the mode would be the most sensible measure. In this case, the modal category is Central America.
c)   In a Pareto chart, the place of birth categories would be organized from the one with the highest percentage (i.e., Central America) to the one with the lowest percentage (i.e., South America). The advantage of a Pareto chart is that it's much easier to make comparisons and to identify easily the most common outcomes (i.e., to identify which places of birth have higher percentages of current foreign-born Americans).

2.94   Cool in China
a)   The variable being measured is personality trait that defines "cool."
b)   This is a categorical variable.
c)   Because the data are categorical with unordered categories, we could use only the bar chart and the modal category.
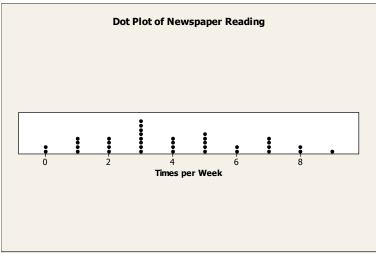
2.95   Chad voting problems
a)   We first locate the dot directly above 11.6% on the horizontal x axis. We then look at the vertical y axis across from this point to determine the label for that dot: optical, two column. This tells us that the over-vote was highest among those using optical scanning with a two-column ballot.
b)   We first locate the dots above the lowest percentages on the x axis. We then determine the labels across from these dots on the y axis to determine the lowest two combinations: optical, one column, and votomatic, one column. Thus, the lowest over-voting occurred when voters had a ballot with only one column that was registered either using optical scanning or votomatic (manual punching of chads).
c)   We could summarize these data further by using a bar for each combination: optical, one column; optical, two column; votomatic, one column, etc. For each bar, we could then plot the average over-vote of all counties in that category. To do this, we would need the exact percentages of each county in each category.

2.96   Number of children
   a)   The most appropriate graph would be a histogram because there are so many data points. Dot plots and stem-and-leaf plots are unwieldy with many data points.
   b)   A histogram would show that the data are skewed to the right. There are many data points at the lower numbers, but only a few as we get to higher numbers of children.

2.97   Newspaper reading
   a)

   b)   Stem-and-leaf of Times per Week   N = 36
        Leaf Unit = 0.10
         0 00
         1 0000
         2 0000
         3 00000000
         4 0000
         5 00000
         6 00
         7 0000
         8 00
         9 0
        The leaf unit is identified above. The stems are the whole numbers, 0 through 9.
   c)   The median is the middle number. There are 36 numbers, so the median is between 18 and 19 which have the values 3 and 4, respectively. Thus, the median is 3.5.
   d)   The distribution is slightly skewed to the right.

2.98   Match the histogram
   (a) Symmetric and bimodal
   (b) Skewed to the right
   (c) Skewed to the left
   (d) Symmetric and unimodal

2.99 **Sandwiches and protein**
a)

```
0 | 8
1 |
1 | 7889
2 | 113
2 | 666
```

b) A stem-and-leaf plot allows one to see the individual amounts.
c) The protein amounts are mostly between 17 and 21 grams with a few sandwiches having a higher protein value of 26 grams. There appears to be one outlier having only 8 grams of protein.

2.100 **Sandwiches and cost**
a) (the data values have been truncated)

```
2 | 4
2 | 999
3 | 1444
3 | 688
```

b) A stem-and-leaf plot allows one to see the individual prices.
c) Most of the sandwiches cost between $2.90 and $3.89. The prices are skewed to the left with one sandwich costing only $2.49.

2.101 **What shape do you expect?**
a) Number of times arrested in past year – skewed to the right because most values are at 0 but there are some large values.
b) Time needed to complete difficult exam (maximum time is 1 hour) – skewed to the left because most values are at 1 hour or slightly less, but some could be quite a bit less.
c) Assessed value of home – skewed to the right because there are some extremely large values.
d) Age at death – skewed to the left because most values are high, but some very young people die.

2.102 **Sketch plots** NOTE: plots will look different for each student, but should have the following characteristics.
a) It would be skewed to the right, and the mean would be greater than the median because of a few mansions that sell for millions.
b) It would be skewed to the right, and the mean would be higher than the median. Most women do not give birth over age 40. Thus, the median would be zero. The mean, however, would be positive, because some women do give birth over the age of 40.
c) It would be skewed to the left, and the mean would be lower than the median. The mean would be pulled down by the outlier of 50. The standard deviation is only 10, so there probably aren't lots of low scores. Moreover, the highest possible score of 100 is only $12/10 = 1.2$ standard deviations above the mean.
d) It would be skewed to the left, and the mean would likely be lower than the median. Most people with cars drive them every month, but a few drive them less, and some hardly or not at all. These outliers would pull the mean, but not the median, lower. The median and mode probably would be 12.

2.103 **Median versus mean sales price of new homes**
We would expect the mean sales price to have been higher due to the distribution being skewed to the right. A few very expensive homes will greatly affect the mean, but not the median sales price.

2.104 **Household net worth**
a) The distribution of these families' net worth is likely to be skewed to the right because relatively few families would have very high net worth so that we expect the mean to be greater than the median.
b) When assets such as homes and retirement savings decline due to a recession, it is typical for the highest valued assets to be affected the most. Thus, we would expect the mean net worth to drop more than the median net worth.

2.105 **Golfers' gains**
a) The data for the 90 players would be skewed to the right with the majority of the golfers earning between $1 and $3 million and a few earning over $3 million.

b) Since the data is skewed to the right, the mean would be the higher value of $2,090,012 and the median the lesser value of $1,646,853.

2.106 Hiking
The classification into easy, medium or hard is categorical and the length classification is quantitative.

2.107 Lengths of hikes
a) One example is 1, 2, 4, 6, 7. Both the mean and median are 4.
b) One example is 2, 5, 2, 6 and 3.

2.108 Central Park monthly temperatures
a) Both distributions are fairly symmetric with January having greater variability than July.
b) The mean temperature for January is around 32º and the mean temperature for July is around 76º. The average monthly temperature in January is approximately 44º less than the average monthly temperature in July.
c) The average monthly temperature in January is more variable than in July. The range of average temperatures for January is approximately 22º to 43º and the standard deviation is approximately 5º. The range of average temperatures for July is approximately 71º to 81º and the standard deviation is approximately 2º.   It may be a bit surprising to see how much more variable are the average monthly temperatures in January than in July.

2.109 What does *s* equal?
a) Given the mean and range, the most realistic value is 12. -10 is not realistic because standard deviation must be 0 or positive. Given that there is a large range, it is not realistic that there would be almost no spread; hence, the standard deviation of 1 is unrealistic. 60 is unrealistically large; the whole range is hardly any more than 60.
b) -20 is impossible because standard deviations must be nonnegative.

2.110 Female heights
a) According to the Empirical Rule, 95% of scores in a bell-shaped distribution fall within two standard deviations of the mean.
$\bar{x} - 2s$ =65-2(3.5)=58
$\bar{x} + 2s$ =65+2(3.5)=72
Thus, 95% of heights likely fall between 58 and 72 inches.
b) The height for a woman who is three standard deviations below the mean is 54.5.
$\bar{x} - 3s$ =65-3(3.5)=54.5
This is on the cusp of what would be considered an outlier according to the z-score criterion. Scores that are beyond three standard deviations from the mean are considered to be potential outliers. So, yes, this height is bordering on unusual.

2.111 Energy and water consumption
a) The distribution is likely skewed to the right because the maximum is much farther from the mean than the minimum is, and also because the standard lowest possible value of 0 is only 780/506 = 1.54 standard deviations below the mean.
b) The distribution is likely skewed to the right because the standard deviation is almost as large as the mean, and the smallest possible value is zero, only 1.15 standard deviation below the mean.
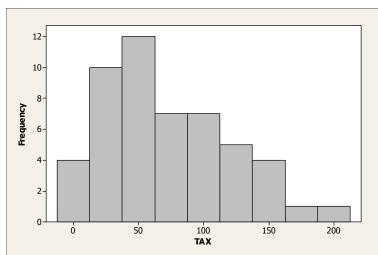
2.112 Mean versus Median and Income
For the same reason that income is skewed to the right (a small percentage of the population earn very high salaries), each percentile class is also skewed to the right. The skewness becomes greater as the income percentile increases; there is less difference between the incomes of the 45th and 49th percentiles, for example than, between the 95th and 99th percentiles.

⌨**2.113 Student heights**
    a) For a bell-shaped distribution, such as the heights of all men, the Empirical Rule states that all or nearly all scores will fall within three standard deviations of the mean ($\bar{x} \pm 3s$). In this case, that means that nearly all scores would fall between 62.2 and 79.6. In this example, almost all men's scores do fall between these values.
    b) The center for women is about five inches less than the center for men. The variability, however, is very similar. These distributions are likely very similar in shape; they are just centered around different values.
    c) The lowest score for men is 62. This would have a *z*-score of $z = \dfrac{x - \bar{x}}{s} = \dfrac{62 - 70.9}{2.9} = -3.07$. Thus, it falls 3.07 standard deviations below the mean.

⌨**2.114 Cigarette tax**
    a)

    The histogram shows a unimodal distribution that is skewed to the right. If there are any outliers, they would be the most extreme scores, such as the one around 200.
    b) The mean is 72.85 and the median is 60. The mean is inflated relative to the median as one would expect from the distribution depicted in the histogram that is skewed to the right. The few high scores would pull the mean higher, but not the median.
    c) The standard deviation is 48.00. This indicates that the typical score falls about 48.0 from the mean.

2.115 Cereal sugar values
    a) Numbers are approximate:
    Minimum: 0
    Q1: 4
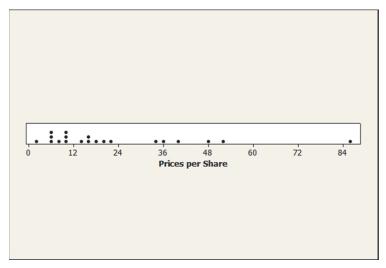   Median: 9.5
    Q3: 13.5
    Maximum: 18
    b) Because the median is closer to Q3 and the maximum than it is to Q1 or the minimum, it appears that this distribution is slightly skewed to the left.
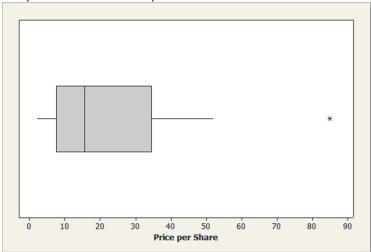   c) This sugar value falls 1.64 standard deviations below the mean.
    $z = \dfrac{x - \bar{x}}{s}$, $z = \dfrac{0 - 8.75}{5.32} = -1.64.$

2.116 Stock prices positions
    a) Note: Due to the large spread of the data, plotting is only done at even prices so that a dot above a price of 2n represents a price per share of either 2n-1 or 2n.
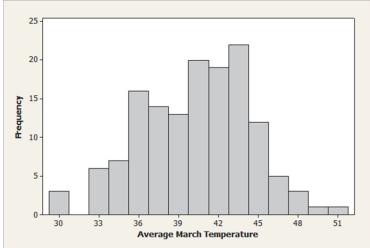
b)   The median is the average of the two middle numbers, 15 and 16. Thus, the median is 15.5. The first quartile is the median of all the numbers below the median: 2, 5, 5, 5, 7, 9, 9, 10, 13, and 15. Thus, the first quartile is 8. The third quartile is the median of all numbers above the median: 16, 17, 19, 21, 33, 35, 39, 48, 52 and 85. Thus, the third quartile is 34.

c)   The box plot does not show the gaps in the observations above 21. Also, the individual data values cannot be reproduced from the box plot.
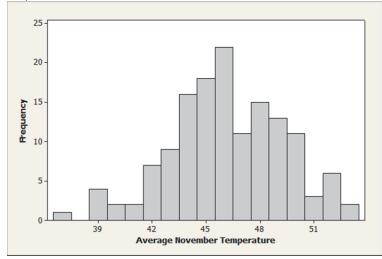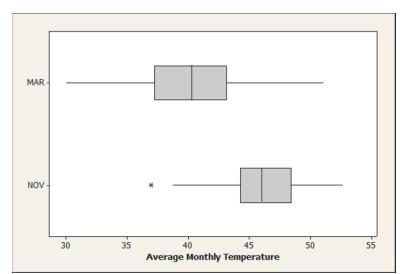
⌨2.117 Temperatures in Central Park
   a) The distribution appears to be fairly symmetric. The mean and median are almost the same, bolstering the contention that the data are normal. The spread is not large in comparison to the mean.

   b) Mean: 40.27; StDev: 4.169
   c) The mean average temperature is higher in November than in March and the spread of average temperatures is less in November than in March.

d)

As indicated by the histograms, the average monthly temperature is greater in November than in March and the standard deviation is less. The side-by-side box plot makes it easier to see the relative spreads of the data as well as the difference in the means.

2.118   Teachers' salaries
a)   The range is the maximum minus the minimum. Range=61,800-33,100=28,700.
The interquartile range (IQR) equals Q3-Q1. IQR=48,850-39,250=9,600.
These statistics indicate that the salaries range across a $28,700 span, and that the middle 50% of salaries range across an $9,600 span.
b)   (i)   The values at the ends of the boxes would be 39,250 (Q1) and 48,850 (Q3).
(ii) The line in the middle of the box would be the median, 42,700.
(iii) The lower end of the left whisker would be the minimum, 33,100.
(iv) The upper end of the right whisker would be the maximum, 61,800.
c)   The minimum and Q1 are closer to the median than are Q3 and the maximum. This indicates that the data are likely skewed to the right.
d)   The most realistic standard deviation would be 6,000. 100 and 1,000 are too small for typical deviations given a range of 28,700. 25,000 is too big given that the whole range extends only slightly farther; the typical score could never be this far from the mean. 6,000 is the only realistic score.

2.119   Health insurance
a)   The distribution is probably skewed to the right because the distance of Q3 from the median and from Q3 to the maximum is longer than the distance of Q1 from the median and from Q1 to the minimum.
b)   The most plausible value for the standard deviation of this distribution is 4. The middle 50% of scores fall within a range of 5.4%., making it plausible that the typical score would deviate four percentage points from the mean. We cannot have a negative percentage point, so -16 is not plausible. We know that there is variation, so 0 is not plausible. The whole range is not much more than 15; thus, 15 and 25 are implausibly large for the standard deviation of this distribution.

2.120   What box plot do you expect?
Box plots will differ for each student, but should have the following characteristics.
a)   The center of these data is closer to the maximum than the minimum. Although the mean is likely to be pulled by an outlier more than the median, this still indicates that the data might be skewed to the left, and that the box plot might have more distance between the median and both Q1 and the minimum than between the median and both Q3 and the maximum.
b)   IQ scores are designed to be symmetric, and these data support that. The box plot, thus, would appear symmetric.

c) The mean is higher than is the median, indicating that the data are skewed to the right. Thus, the box plot would have more distance between the median and both Q3 and the maximum than between the median and both Q1 and the minimum.

2.121 High school graduation rates
a) The range is the difference between the lowest and highest scores: 92.3-78.3=14.
The interquartile range (IQR) is the difference between scores at the 25th and 75th percentiles: IQR=Q3-Q1=88.8-83.6=5.2.
b) 1.5(IQR) = 7.8 from Q1 or Q3; this criterion suggests that potential outliers would be those scores less than 75.8 and greater than 96.6. There are no scores beyond these values, and so it would not indicate any potential outliers.

2.122 SAT scores revisited
a) Because the right whisker extends further than does the left whisker, the box plot suggests that the distribution is somewhat skewed to the right.
b) Numbers are approximate:
Minimum: 1390
Q1: 1480
Median: 1560
Q3: 1695
Maximum: 1800
The lowest score is approximately 1390 and the highest is 1800. The score below which the lowest 25% fall is approximately 1560, and the score above which the highest 25% fall is approximately 1695. The middle score, that below which 50% of the scores fall, is 1560.
c) If only viewing the box plot, we would not see that the distribution may not be unimodal.

2.123 Blood pressure
a) $z = \dfrac{x - \bar{x}}{s} = \dfrac{140 - 121}{16} = 1.19$

A z-score of 1.19 indicates that a person with a blood pressure of 140, the cutoff for having high blood pressure, falls 1.19 standard deviations above the mean.
b) About 95% of all values in a bell-shaped distribution fall within two standard deviations of the mean – in this case, 32. Subtracting two times the standard deviation from the mean, and adding two times the standard deviation to the mean tells us that about 95% of systolic blood pressures fall between 89 and 153.

2.124 No cereal sodium
If a cereal has a sodium value of 0, it falls -2.16 standard deviations below the mean.
$z = \dfrac{x - \bar{x}}{s} = \dfrac{0 - 167}{77.3} = -2.16$

2.125 Who was Roger Maris?
a) $z = \dfrac{x - \bar{x}}{s} = \dfrac{5 - 22.92}{15.98} = -1.12$
$z = \dfrac{x - \bar{x}}{s} = \dfrac{61 - 22.92}{15.98} = 2.38$
Neither the minimum nor the maximum score reaches the criterion for a potential outlier of being more than three standard deviations from the mean (i.e., having a z-score less than -3 or greater than 3). Thus, there are no potential outliers according to three standard deviation criterion.
b) The maximum is much farther from the mean and median than is the minimum, an indicator that the distribution might not be bell-shaped. Moreover, the lowest possible value of 0 is only 22.92/15.98 = 1.43 standard deviations below the mean.

c) Based on the criteria noted above, this is not unusual. It does not even come close to meeting the three standard deviation criterion for a potential outlier and therefore is not an unusual number of homeruns for Roger Maris.

$$z = \frac{x - \bar{x}}{s} = \frac{13 - 22.92}{15.98} = -0.62$$

CHAPTER PROBLEMS: CONCEPTS AND INVESTIGATIONS
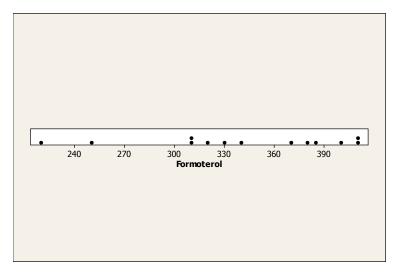
⌨2.126 Baseball's great homerun hitters

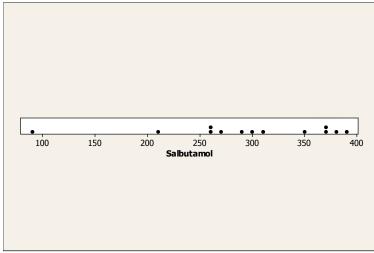The responses will be different for each student depending on the methods used.

⌨2.127 How much spent on haircuts?

The responses will be different for each student depending on the methods used.

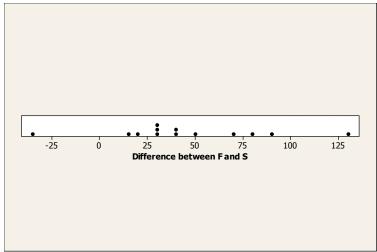**⌨2.128 Controlling asthma**

a) The distribution of children on Formoterol (F) is more normal than is the slightly skewed distribution of children on Salbutamol (S), although it also is slightly skewed to the left. There is a data point that qualifies as an outlier, indicated by a dot to the far left, in the S distribution. Moreover, children on F seem to be doing better, on average, than do children on S.

b) Here are the difference scores for each child. A positive difference indicates a higher score for Formoterol than for Salbutamol.

Child	Formoterol	Salbutamol	Difference
1	310	270	40
2	385	370	15
3	400	310	90
4	310	260	50
5	410	380	30
6	370	300	70
7	410	390	20
8	320	290	30
9	330	365	-35
10	250	210	40
11	380	350	30
12	340	260	80
13	220	90	130

Difference between F and S

If there is, on average, no difference between PEF levels for the two brands, the distribution of differences would be centered around 0, a score indicating no difference.

The current difference scores appear to have a roughly normal distribution, perhaps skewed slightly to the right. The difference scores indicate a positive difference, on average. The center is well into the positive side, and the data points are quite spread out. Thus, children on F have higher scores, on average, than when they are on S, although there is a quite a bit of variation in amount of improvement. Moreover, there appears to be one outlier, a child who responds more poorly on F than on S.

⌨2.129 Google trend
The response to this exercise will be different for each student.

2.130 Google again
The response to this exercise will be different for each student.

2.131 **Back-to-back stem-and-leaf plot**
a) The two distributions are different. The median for the adult cereals is 6000 milligrams, with a shape skewed to the right. The median for the child cereals is 10,500 grams with a shape skewed to the left. In addition, the sugar values are more variable for the adult cereals.
b) Leaf Unit = 10

```
     Adult    Child
        0| 0 |
       75| 0 |
        0| 1 |344
      885| 1 |689
       20| 2 |011
         | 2 |9
        4| 3 |
```

The distributions are different. The medians are similar, but the variability is much greater for the adult cereals than for the children's. The distribution of sodium levels is fairly symmetric for the children's cereals but is skewed to the left for the adult cereals with one possible outlier having a sodium content of 340 milligrams.

2.132 **You give examples**
Note: Responses for this problem will be different for each student. These are just examples.
a) Approximately symmetric – number of letters that can be remembered in a memory task, or IQ
b) Skewed to the right – number of alcoholic beverages consumed in a week (this would be skewed by a few extreme binge-drinkers) or distance traveled to work (skewed by a few with incredibly long commutes)
c) Skewed to the left – happiness levels on one's wedding day (most would be very happy, but there'd likely be a few who were sad) or score on an easy exam (skewed by a few who did poorly anyway)
d) Bimodal – writing ability in a university writing center (some would come because they need help, the rest would be skilled tutors, and there would be fewer in the middle) or income for a sample that includes people from the U.S. and people from a third world country (some would center on a very low income, an some around a much higher income)
e) Skewed to the right, with a mode and median of 0 but a positive mean – number of times students have eaten snake in their lives (most would never have eaten it, but a few would have tried it once, perhaps on Fear Factor, and an even smaller number would have had it several times) or number of times married for a sample of high school students (most would not have been married at all, but a few would have been married once, and an even smaller number would have been married more than once)

2.133 **Political conservatism and liberalism**
a) As seen in Example 12, one need not add up every separate number when calculating a mean. This would be unwieldy with the political conservatism and liberalism data. We would have to add up 69 ones, 240 twos, etc. (all the way up to 68 sevens), then divide by the 1,933 people in the study. There's a far easier way. We can find the sum of all values in the study ($\sum x$) by multiplying each possible value (1-7 in this case) by its frequency.

$$\bar{x} = \frac{\sum x}{n} = \frac{69(1) + 240(2) + 221(3) + 740(4) + 268(5) + 327(6) + 68(7)}{1933} = \frac{7950}{1933} = 4.11 .$$

b) The mode, the most common score, is four.
c) The median would be the 967th scores. In this case, that category is four.

2.134 **Mode but not median and mean**
We use the mode when we're interested in the category with the highest frequency, as opposed to merely finding the "center" of the data. To find a mean or median, we must have observations that measure a quantity. With unordered categories, observations do not do this. But, we can still find the most common outcome, so the mode is appropriate.

2.135 **Multipe choice - GRE scores**
The best answer is (a).

2.136   Multiple choice - Facts about *s*
The best answer is (b). (s can be zero if all observations hold the same value.)

2.137   Multiple choice - Relative GPA
The best answer is (a). (The standard deviation would allow her to calculate her *z*-score.)

2.138   True or false
a)   False; consider the following data set: 3, 3, 3, 3, 3. Note that the mean=median=mode=3.
b)   False; consider the following data set: 1 2 3 4. The mean is 2.5 which is not one of the data points.
c)   True; when *n* is odd, the median is the data point in the $\left(\dfrac{n+1}{2}\right)$ position of the sorted data.
d)   True; by definition, the median is the second quartile which is also known as the 50th percentile.

2.139   Bad statistic
The standard deviation was incorrectly recorded. The standard deviation represents a typical scores distance from the mean. For grades ranging between 26 and 100, a standard deviation of 76 is way too large.

2.140   Soccer true or false
False. The mean would be inflated by the salaries of the few players who earn exorbitant salaries, but the magnitudes of these salaries would not affect the median. Thus, the mean would be larger than would the median.

♦♦2.141 Mean for grouped data
In Example 12 and exercise 2.133, the mean could be expressed as a sum. Above, the mean was calculated by multiplying each score by its frequency, then summing these and dividing by the total number of subjects. Alternatively, we could first divide each frequency by the number of subjects, rather than dividing the sum by the number of subjects. Dividing the frequency for a given category by the total number of subjects would give us the proportion. We are just changing the order in which we perform the necessary operations to calculate the mean.

♦♦2.142 Male heights
a)   The median falls at the 50th percentile. In this case, the 50th percentile falls in the group that is 70 inches or less (i.e., 54% of all subjects), but above 69 inches. Thus, the median is in the category "70 inches or less," but above 69 inches.
b)   If the distribution is bell-shaped, the mean would fall in the middle, and be about 70 inches. Further, the Empirical Rule would apply, and nearly all scores would fall within three standard deviations of the mean. If nearly all scores fall within 10 inches from the mean (60 is 10 inches below the mean of 70, and 80 is ten inches above it), the standard deviation would be about 10 divided by 3, or about 3.3.

♦♦2.143 Range and standard deviation approximation
Based on the work of statisticians (the Empirical Rule), we know that most, if not all, data points fall within three standard deviations of the mean if we have a bell-shaped distribution. The formula for this is $\bar{x} = \pm 3s$. If the region from three standard deviations below the mean to three standard deviations above the mean encompasses just about everyone in the data set, we could add the section below the mean (3*s*) to the section above the mean (3*s*) to get everyone in the data set. 3*s*+3*s*=6*s*. Because the range is defined as everyone in the dataset, we can say that the range is equal, approximately, to 6*s*.

♦♦2.144 Range the least resistant
There are only two observations that are taken into account by the range, the minimum and maximum scores. The range is the difference between these two, so the range increases exactly the same amount as one of these scores increases (or in the case of the minimum, decreases). The mean and standard deviation, however, take the magnitude of all observations into account. Although an extreme score would pull the mean in its direction, and would increase the standard deviation, this "pull" would be offset, at least to some degree, by the values of the other observations.

♦♦2.145 Using MAD to measure variability

a) When calculating *MAD*, the value of a given observation is subtracted from the mean, the absolute value is taken, and the resulting number is added to the rest of the deviations before dividing by *n* (we're taking the mean of these absolute deviations).With greater variability, numbers tend to be further from the mean. Thus, the absolute values of their deviations from the mean would be larger. When we take the average of all these values, the overall *MAD* is larger than with distributions with less spread.

b) We would expect *MAD* to be a little less resistant than the standard deviation. Because we square the deviations, a large deviation has a greater effect.

♦♦2.146 Rescale the data

a) The constant *c*, in this case, is 20, and we are using the rule that involves adding. The mean will be $\bar{x} + c$ = 57+20 = 77.
The standard deviation will not change, and will still be 20.

b) The constant *c,* in this case, is 0.5, and we are using the rule that involves multiplying.
The mean will be $c\bar{x}$ =0.5(39,000)=19,500 British pounds.
The standard deviation will be $|c|s$=0.5(15,000)=7,500 British pounds.

c) Adding a constant does not change the shape of a distribution, but merely moves it to the right or left. Multiplying by a constant makes the distribution either more or less spread out, but it does not change the shape; e.g., if the original distribution is skewed right, so is the new one.

CHAPTER PROBLEMS: STUDENT ACTIVITIES

The responses for exercises 2.147 and 2.148 will vary for each class depending on the data set used.

▣2.149 GSS

Frequency table:

TVHOURS	Count
0	87
1	273
2	354
3	215
4	184
5	63
6	59
7	10
8	35
9	3
10	18
11	1
12	9
14	2
15	2
16	2
17	1
18	1
20	1
24	4

Histogram: