1: Describing Data with Graphs

1.1 **a** The experimental unit, the individual or object on which a variable is measured, is the student.
b The experimental unit on which the number of errors is measured is the exam.
c The experimental unit is the patient.
d The experimental unit is the azalea plant.
e The experimental unit is the car.

1.2 **a** "Time to assemble" is a *quantitative* variable because a numerical quantity (1 hour, 1.5 hours, etc.) is measured.
b "Number of students" is a *quantitative* variable because a numerical quantity (1, 2, etc.) is measured.
c "Rating of a politician" is a *qualitative* variable since a quality (excellent, good, fair, poor) is measured.
d "State of residence" is a *qualitative* variable since a quality (CA, MT, AL, etc.) is measured.

1.3 **a** "Population" is a *discrete* variable because it can take on only integer values.
b "Weight" is a *continuous* variable, taking on any values associated with an interval on the real line.
c "Time" is a *continuous* variable.
d "Number of consumers" is integer-valued and hence *discrete*.

1.4 **a** "Number of boating accidents" is integer-valued and hence *discrete*.
b "Time" is a *continuous* variable.
c "Cost of a head of lettuce" is a *discrete* variable since money can be measured only in dollars and cents.
d "Yield in kilograms" is a *continuous* variable, taking on any values associated with an interval on the real line.

1.5 **a** The experimental unit, the item or object on which variables are measured, is the vehicle.
b Type (qualitative); make (qualitative); carpool or not? (qualitative); one-way commute distance (quantitative continuous); age of vehicle (quantitative continuous)
c Since five variables have been measured, this is *multivariate data*.

1.6 **a** The set of ages at death represents a population, because there have only been 38 different presidents in the United States history.
b The variable being measured is the continuous variable "age".
c "Age" is a quantitative variable.

1.7 The population of interest consists of voter opinions (for or against the candidate) <u>at the time of the election</u> for all persons voting in the election. Note that when a sample is taken (at some time prior or the election), we are not actually sampling from the population of interest. As time passes, voter opinions change. Hence, the population of voter opinions changes with time, and the sample may not be representative of the population of interest.

1.8 **a-b** The variable "survival time" is a quantitative continuous variable.
c The population of interest is the population of survival times for all patients having a particular type of cancer and having undergone a particular type of radiotherapy.
d-e Note that there is a problem with sampling in this situation. If we sample from all patients having cancer and radiotherapy, some may still be living and their survival time will not be measurable. Hence, we cannot sample directly from the population of interest, but must arrive at some reasonable alternate population from which to sample.

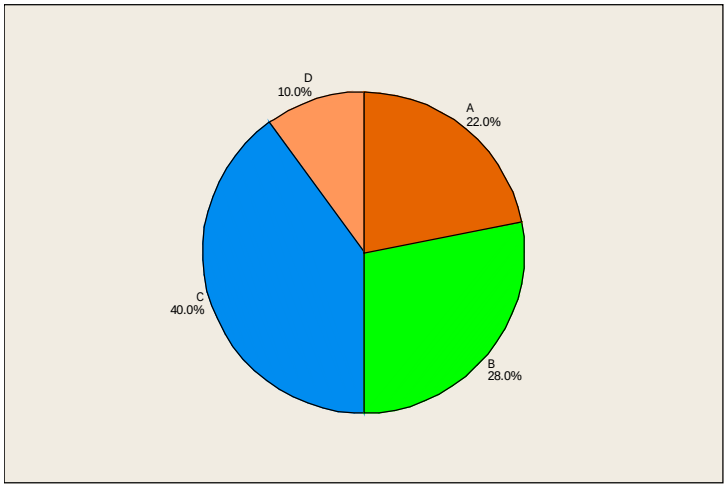
1.9 **a** The variable "reading score" is a quantitative variable, which is probably integer-valued and hence discrete.
b The individual on which the variable is measured is the student.

c The population is hypothetical – it does not exist in fact – but consists of the reading scores for all students who could possibly be taught by this method.

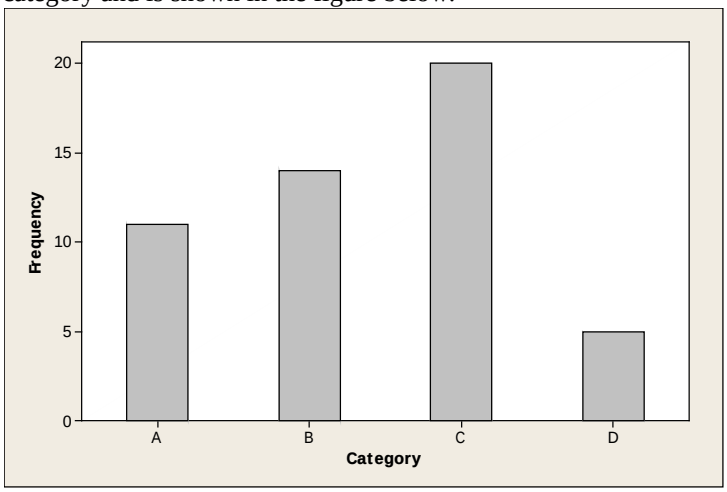
1.10 **a-b** The variable "category" is a qualitative variable measured for each of fifty people who constitute the experimental units.

c The pie chart is constructed by partitioning the circle into four parts, according to the total contributed by each part. Since the total number of people is 50, the total number in category A represents $11/50 = 0.22$ or 22% of the total. Thus, this category will be represented by a sector angle of $0.22(360) = 79.2^{o}$. The other sector angles are shown below. The pie chart is shown in the figure below.

Category	Frequency	Fraction of Total	Sector Angle
A	11	.22	79.2
B	14	.28	100.8
C	20	.40	144.0
D	5	.10	36.0



d The bar chart represents each category as a bar with height equal to the frequency of occurrence of that category and is shown in the figure below.



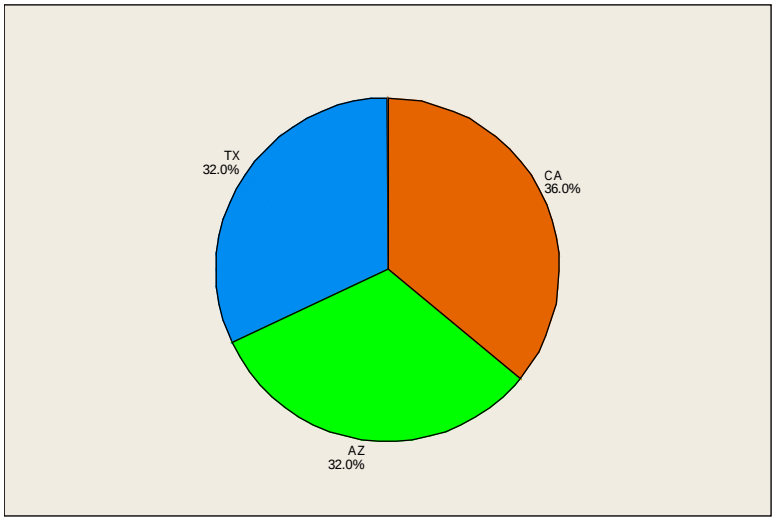
e Yes, the shape will change depending on the order of presentation. The order is unimportant.

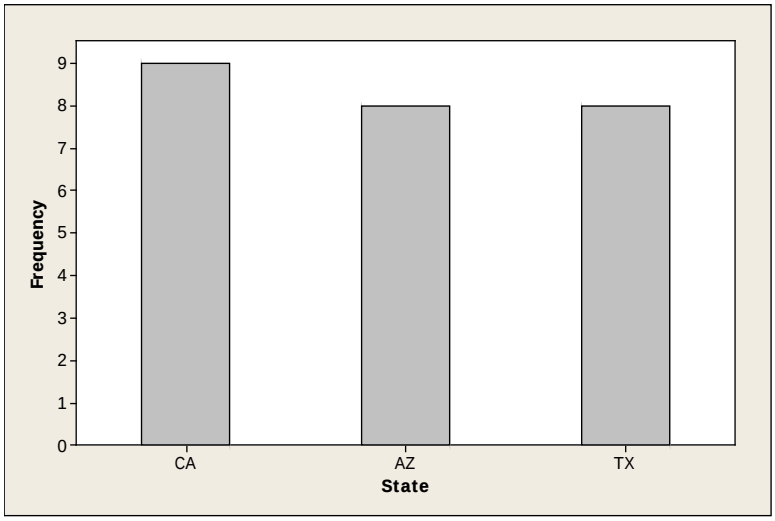
f The proportion of people in categories B, C, or D is found by summing the frequencies in those three categories, and dividing by n = 50. That is, $(14+20+5)/50 = 0.78$.

g Since there are 14 people in category B, there are $50-14=36$ who are not, and the percentage is calculated as $(36/50)100=72\%$.

1.11 **a-b** The experimental unit is the pair of jeans, on which the qualitative variable "state" is measured.
c-d Construct a statistical table to summarize the data. The pie and bar charts are shown in the figures below.

State	Frequency	Fraction of Total	Sector Angle
CA	9	.36	129.6
AZ	8	.32	115.2
TX	8	.32	115.2





e From the table or the chart, Texas produced $8/25=0.32$ of the jeans.

f The highest bar represents California, which produced the most pairs of jeans.

g Since the bars and the sectors are almost equal in size, the three states produced roughly the same number of pairs of jeans.

1.12 **a** The population of interest consists of voter opinions (for or against the candidate) at the time of the election for all persons voting in the 2012 election.

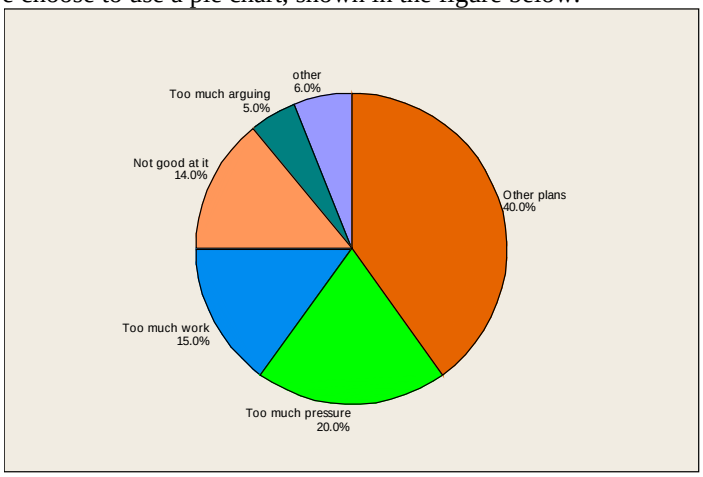
b The population from which the pollsters have sampled is the population of voter preferences on April 9-11, 2010 for all voters registered voters nationwide.

c Registered voters are not necessarily those voters who will actually vote in the election, while likely voters are those who have indicated that they are "likely" to vote. The second group is a subset of the first group.

d Not necessarily. The registered voters surveyed on April 9-11 may fail to actually vote in the election, and/or they may change their minds before the election actually occurs. Moreover, once the actual Democratic and Republican candidates are chosen, the preference proportions for these two candidates may change dramatically.

1.13 **a** The percentages given in the exercise only add to 94%. We should add another category called "Other", which will account for the other 6% of the responses.

b Either type of chart is appropriate. Since the data is already presented as percentages of the whole group, we choose to use a pie chart, shown in the figure below.



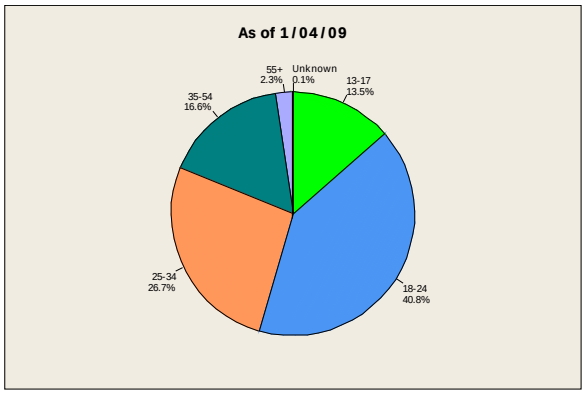
c-d Answers will vary.

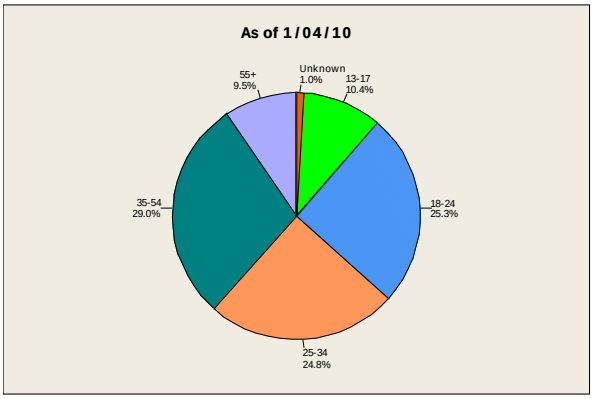
1.14 **a-b** The underlying variable being measured is a quantitative variable, which would be described as "age of Facebook users". However, it is being recorded in age group categories, a qualitative variable.

c The numbers represent the number (in thousands) of Facebook users who fall in each of the six categories.

d-e The percentages falling in each of the six categories are shown below, and the pie charts for January 4, 2009 and January 4, 2010 follow.

Age	As of 1/04/09	As of 1/04/10
13-17	13.5%	10.4%
18-24	40.8%	25.3%
25-34	26.7%	24.8%
35-54	16.6%	29.0%
55+	2.3%	9.5%
Unknown	0.1%	1.0%
Total	**100%**	**100%**



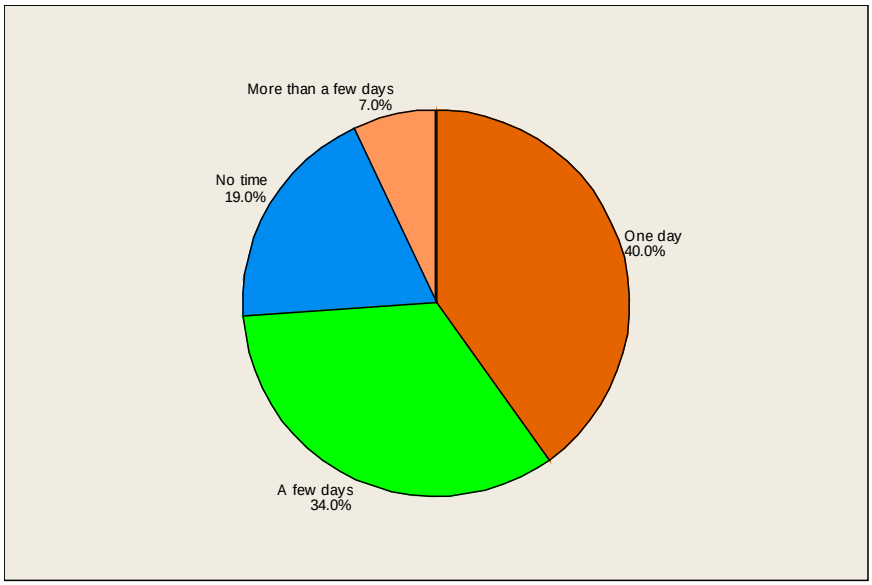


f The user base appears to have shifted towards the older age categories.

1.15 **a** The total percentage of responses given in the table is only $(40+34+19)\% = 93\%$. Hence there are 7% of the opinions not recorded, which should go into a category called "Other" or "More than a few days".

b Yes. The bars are very close to the correct proportions.

c Similar to previous exercises. The pie chart is shown below. The bar chart is probably more interesting to look at.



1.16 The most obvious choice of a stem is to use the ones digit. The portion of the observation to the right of the ones digit constitutes the leaf. Observations are classified by row according to stem and also within each stem according to relative magnitude. The stem and leaf display is shown below.

```
1   6 8
2   1 2 5 5 5 7 8 8 9 9
3   1 1 4 5 5 6 6 6 7 7 7 7 8 9 9 9          leaf digit = 0.1
4   0 0 0 1 2 2 3 4 5 6 7 8 9 9 9          1 2  represents 1.2
5   1 1 6 6 7
6   1 2
```

a The stem and leaf display has a mound shaped distribution.

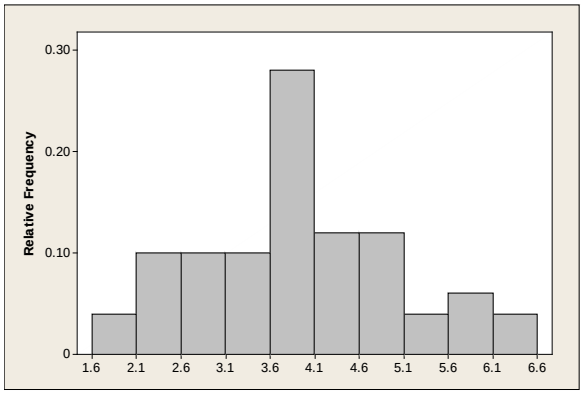
b From the stem and leaf display, the smallest observation is 1.6 (1 6).

c The eight and ninth largest observations are both 4.9 (4 9).

1.17 **a** For $n = 5$, use between 8 and 10 classes.

b

Class i	Class Boundaries	Tally	f_i	Relative frequency, f_i/n
1	1.6 to < 2.1	11	2	.04
2	2.1 to < 2.6	11111	5	.10
3	2.6 to < 3.1	11111	5	.10
4	3.1 to < 3.6	11111	5	.10
5	3.6 to < 4.1	11111 11111 1111	14	.28
6	4.1 to < 4.6	11111 11	7	.14
7	4.6 to < 5.1	11111	5	.10
8	5.1 to < 5.6	11	2	.04
9	5.6 to < 6.1	111	3	.06
10	6.1 to < 6.6	11	2	.04



c From **b**, the fraction less than 5.1 is that fraction lying in classes 1-7, or

$$(2+5+\cdots+7+5)/50 = 43/50 = 0.86$$

d From **b**, the fraction larger than 3.6 lies in classes 5-10, or.

$$(14+7+\cdots+3+2)/50 = 33/50 = 0.66$$

e The stem and leaf display has a more peaked mound-shaped distribution than the relative frequency histogram because of the smaller number of groups.

1.18 **a** As in Exercise 1.16, the stem is chosen as the ones digit, and the portion of the observation to the right of the ones digit is the leaf.

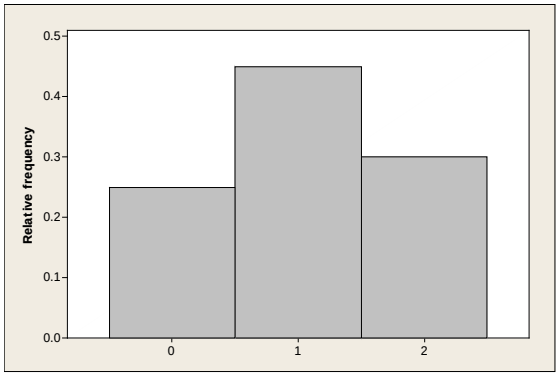
```
3  |  2 3 4 5 5 5 6 6 7 9 9 9 9
4  |  0 0 2 2 3 3 3 4 4 5 8        leaf digit = 0.1  1 2 represents 1.2
```

b The stems are split, with the leaf digits 0 to 4 belonging to the first part of the stem and the leaf digits 5 to 9 belonging to the second. The stem and leaf display shown below improves the presentation of the data.

```
3 | 2 3 4
3 | 5 5 5 6 6 7 9 9 9 9          leaf digit = 0.1  1 2 represents 1.2
3 | 0 0 2 2 3 3 3 4 4
4..| 5 8
```

1.19 **a** Since the variable of interest can only take the values 0, 1, or 2, the classes can be chosen as the integer values 0, 1, and 2. The table below shows the classes, their corresponding frequencies and their relative frequencies. The relative frequency histogram is shown below.

Value	Frequency	Relative Frequency
0	5	.25
1	9	.45
2	6	.30



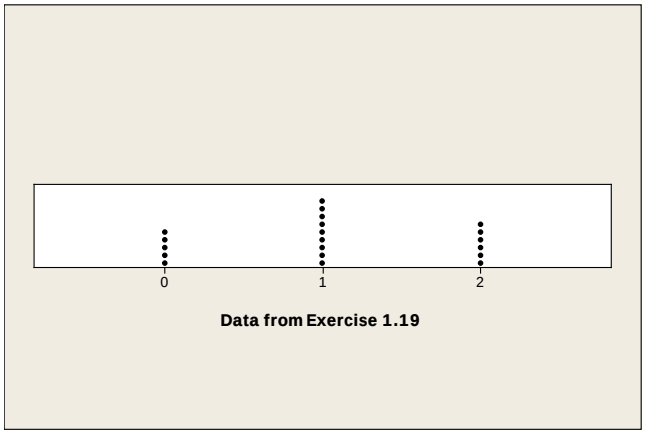
b Using the table in part **a**, the proportion of measurements greater then 1 is the same as the proportion of "2"s, or 0.30.

c The proportion of measurements less than 2 is the same as the proportion of "0"s and "1"s, or $0.25 + 0.45 = .70$.

d The probability of selecting a "2" in a random selection from these twenty measurements is $6/20 = .30$.

e There are no outliers in this relatively symmetric, mound-shaped distribution.

1.20 **a** The scale is drawn on the horizontal axis and the measurements are represented by dots.

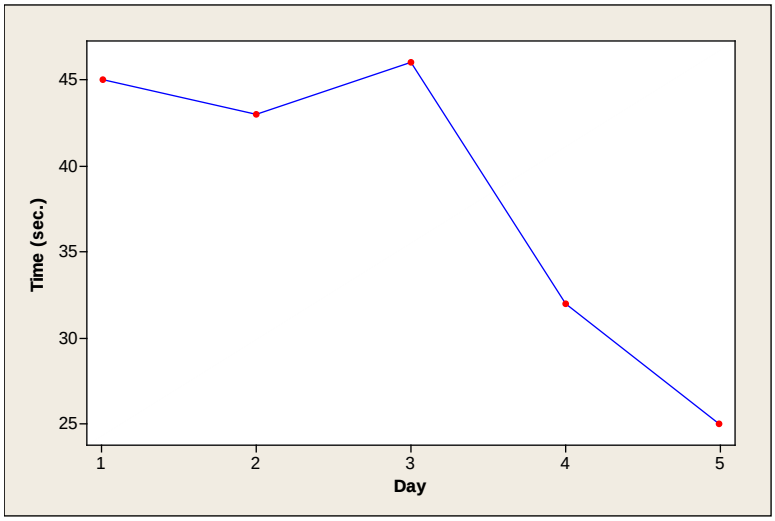


b Since there is only one digit in each measurement, the ones digit must be the stem, and the leaf will be a zero digit for each measurement.

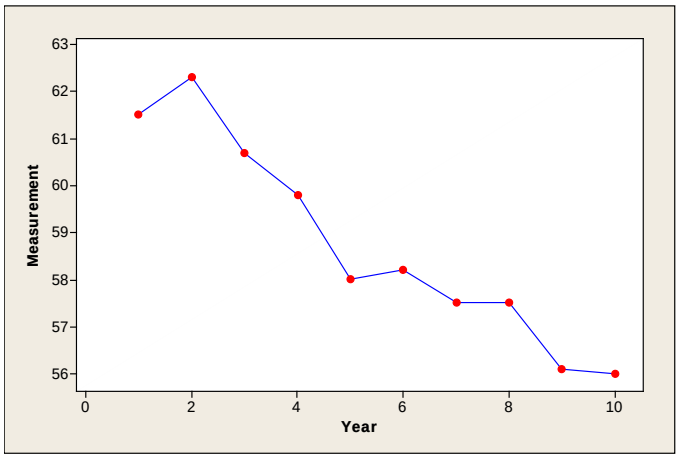
c
0 | 0 0 0 0 0
1 | 0 0 0 0 0 0 0 0 0
2 | 0 0 0 0 0 0

d The two plots convey the same information if the stem and leaf plot is turned 90^o and stretched to resemble the dotplot.

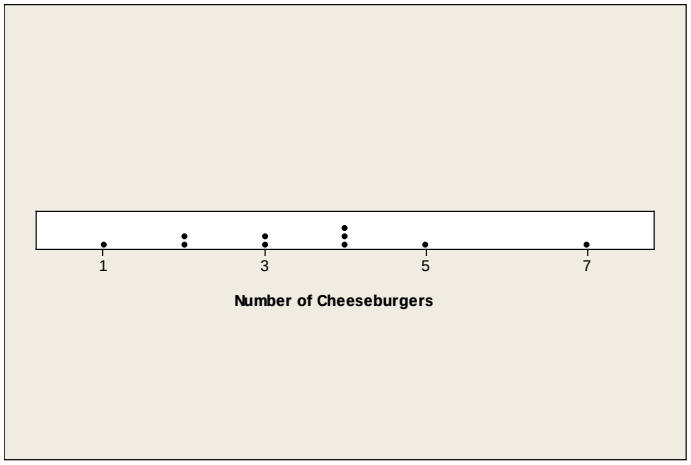
1.21 The line chart plots "day" on the horizontal axis and "time" on the vertical axis. The line chart shown below reveals that learning is taking place, since the time decreases each successive day.



1.22 **a-b** The line graph is shown below. Notice the change in y as x increases. The measurements are decreasing over time.



1.23 The dotplot is shown below.



a The distribution is somewhat mound-shaped (as much as a small set can be); there are no outliers.

b $2/10 = 0.2$

1.24 **a** The test scores are graphed using a stem and leaf plot generated by ***Minitab***.

Stem-and-Leaf Display: Scores

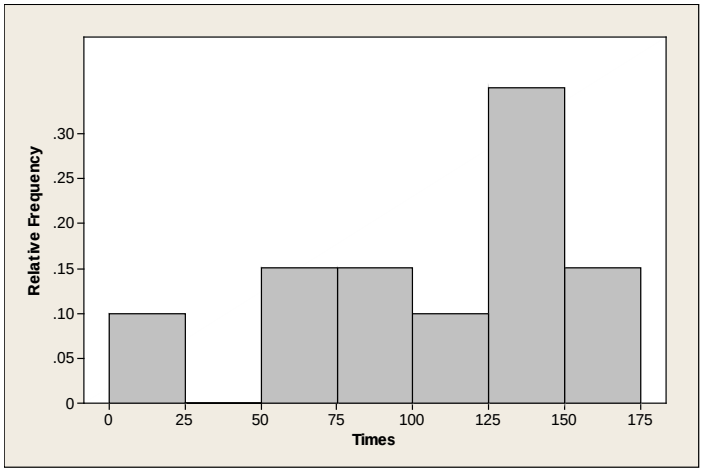
```
Stem-and-leaf of Scores  N  = 20
Leaf Unit = 1.0

 2   5  57
 5   6  123
 8   6  578
 9   7  2
(2)  7  56
 9   8  24
 7   8  6679
 3   9  134
```

b-c The distribution is not mound-shaped, but is rather bimodal with two peaks centered around the scores 65 and 85. This might indicate that the students are divided into two groups – those who understand the material and do well on exams, and those who do not have a thorough command of the material.

1.25 **a** There are a few extremely small numbers, indicating that the distribution is probably skewed to the left.

b The range of the data $165 - 8 = 157$. We choose to use seven class intervals of length 25, with subintervals 0 to < 25, 25 to < 50, 50 to < 75, and so on. The tally and relative frequency histogram are shown below.

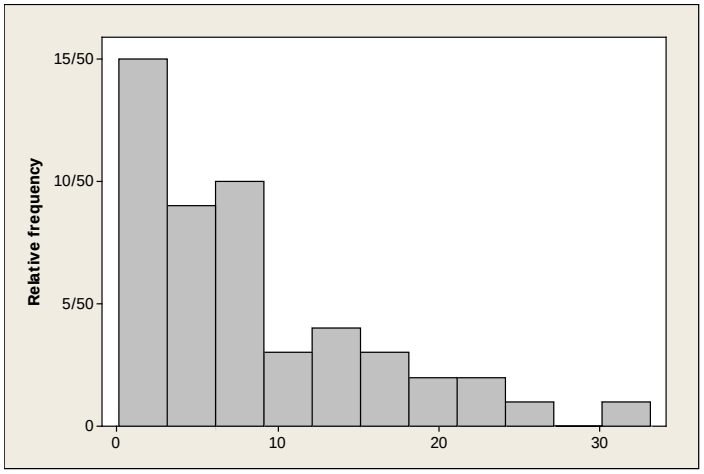
Class i	Class Boundaries	Tally	f_i	Relative frequency, f_i/n
1	0 to < 25	11	2	2/20
2	25 to < 50		0	0/20
3	50 to < 75	111	3	3/20
4	75 to < 100	111	3	3/20
5	100 to < 125	11	2	2/20
6	125 to < 150	11111 11	7	7/20
7	150 to < 175	111	3	3/20



c The distribution is indeed skewed left with two possible outliers – $x = 8$ and $x = 11$.

1.26 **a** The range of the data $32.3 - 0.2 = 32.1$. We choose to use eleven class intervals of length 3 ($32.1/11 = 2.9$, which when rounded to the next largest integer is 3). The subintervals 0.1 to < 3.1, 3.1 to < 6.1, 6.1 to < 9.1, and so on, are convenient and the tally and relative frequency histogram are shown on the next page.

Class i	Class Boundaries	Tally	f_i	Relative frequency, f_i/n
1	0.1 to < 3.1	11111 11111 11111	15	15/50
2	3.1 to < 6.1	11111 1111	9	9/50
3	6.1 to < 9.1	11111 11111	10	10/50
4	9.1 to < 12.1	111	3	3/50
5	12.1 to < 15.1	1111	4	4/50
6	15.1 to < 18.1	111	3	3/50
7	18.1 to < 21.1	11	2	2/50
8	21.1 to < 24.1	11	2	2/50
9	24.1 to < 37.1	1	1	1/50
10	27.1 to < 30.1		0	0/50
11	30.1 to < 33.1	1	1	1/50

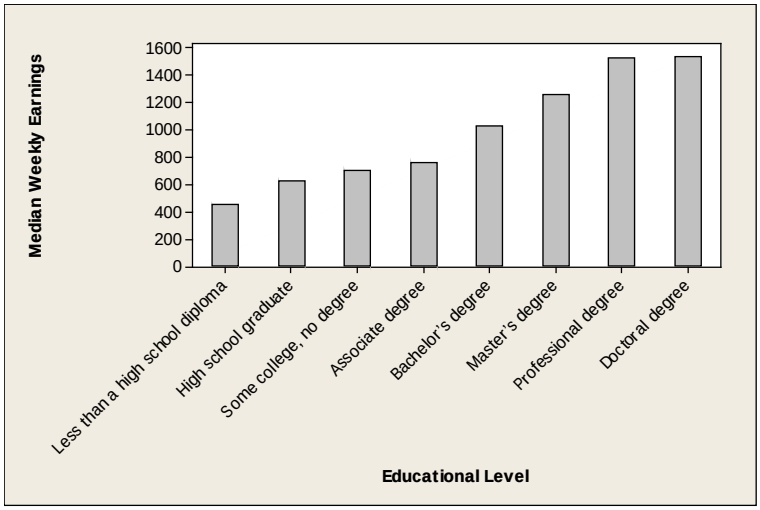


b The data is skewed to the right, with a few unusually large measurements.

c Looking at the data, we see that 36 patients had a disease recurrence within 10 months. Therefore, the fraction of recurrence times less than or equal to 10 is $36/50 = 0.72$.

1.27 **a** The data represent the median weekly earnings for six different levels of education. A bar chart would be the most appropriate graphical method.

b The bar chart is shown below.



c The median weekly earnings increases substantially as the person's educational level increases.

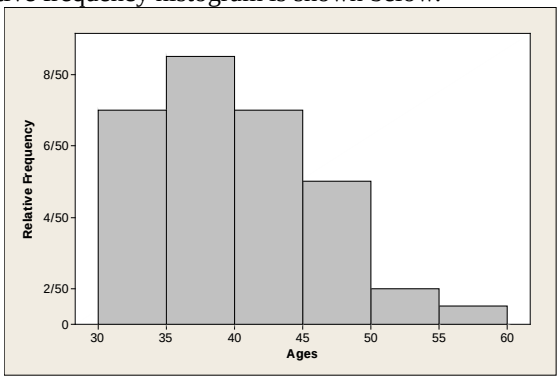
1.28 **a** Use the tens digit as the stem, and the ones digit as the leaf, dividing each stem into two parts.

```
3  | 0 0 0 1 1 2 2 2 3 3 4 4
3  | 5 5 5 6 6 6 6 6 7 7 8 8 9 9 9
4  | 0 0 0 0 1 1 1 1 2 2 3 3
4  | 5 5 6 6 6 7 8 8
5  | 0 0
5  | 5
```

b We use class intervals of length 5, beginning with the subinterval 30 to < 35. The tally is shown below .

Class i	Class Boundaries	Tally	f_i	Relative frequency, f_i/n
1	30 to < 35	11111 11111 11	12	12/50
2	35 to < 40	11111 11111 11111	15	15/50
3	40 to < 45	11111 11111 11	12	12/50
4	45 to < 50	11111 111	8	8/50
5	50 to < 55	11	2	2/50
6	55 to < 60	1	1	1/50

The relative frequency histogram is shown below.

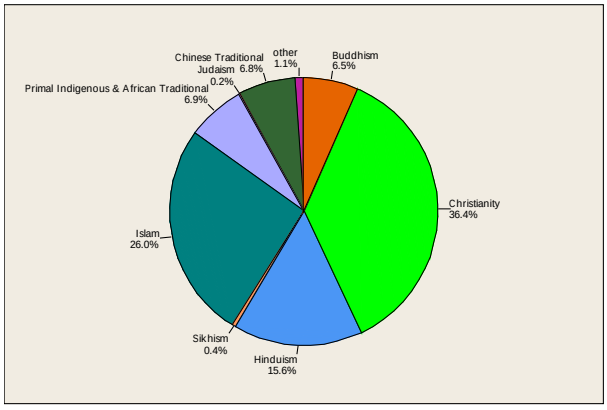


c The two graphs are very similar, with the relative frequency histogram a bit more visually appealing. If the student chose to create the stem and leaf plot without splitting the stems into two parts, the stem and leaf plot would not be very helpful in describing the data set.

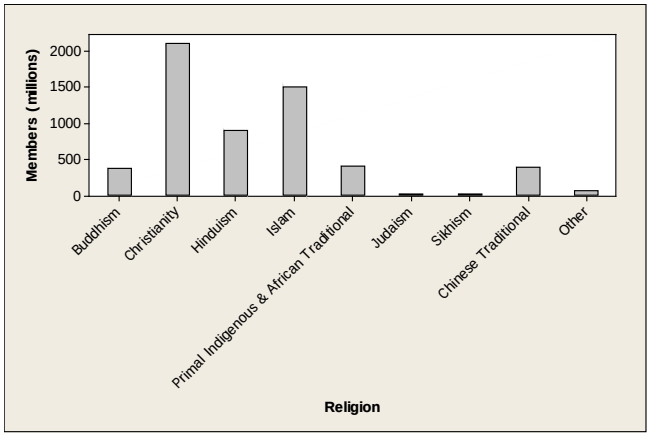
d Use either the stem and leaf plot, the table or the relative frequency histogram. The proportion of children in the interval 35 to < 45 is (15 + 12)/50 = .54.

e The proportion of children aged less than 50 months is (12 + 15 + 12 + 8)/50 = .94.

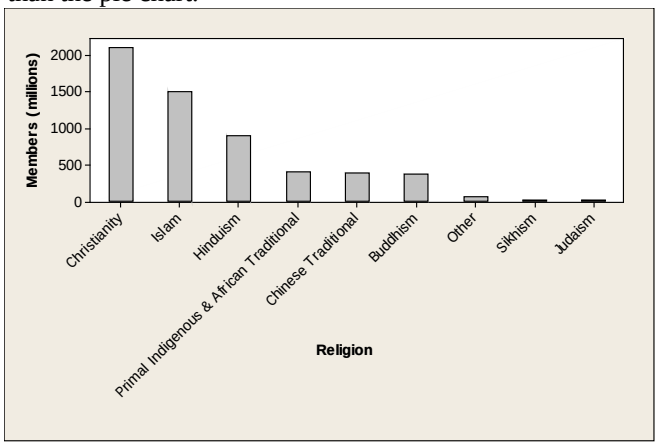
1.29 **a** Similar to previous exercises. The pie chart is shown below.



b The bar chart is shown below.



c The Pareto chart is a bar chart with the heights of the bars ordered from large to small. This display is more effective than the pie chart.



1.30 **a** Use the ones digit as the stem, and the portion to the right of the ones digit as the leaf, dividing each stem into two parts.

```
0  | 2 2 3 3 3 4 4 4
0  | 5 5 6 6 6 6 7 7 7 8 8 8 8 9 9
1  | 0 0 1 1 1 1 1 1 1 2 2 2 3 3 3 4 4
1  | 6 6 7 7 8 8 8 8 9 9
2  | 1 2 3
2  | 5 8                    leaf digit = 0.1
3  | 1 1                  1  2 represents 1.2
3  | 6
4  |
4  | 5
5  | 2
```

b Looking at the original data, we see that 25 customers waited one minute or less. Therefore, the fraction of service times less than or equal to one is $25/60 = 0.4167$.

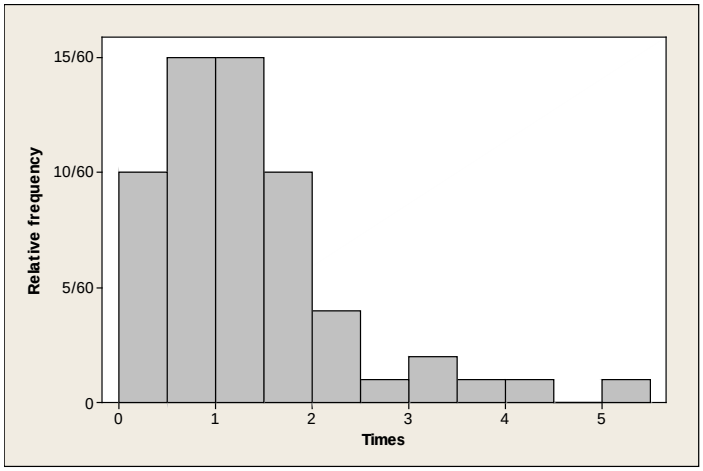
c The smallest measurement is 0 2 which is translated as 0.2.

1.31 **a** The data ranges from .2 to 5.2, or 5.0 units. Since the number of class intervals should be between five and twelve, we choose to use eleven class intervals, with each class interval having length 0.50 ($5.0/11 = .45$, which, rounded to the nearest convenient fraction, is .50). We must now select interval

boundaries such that no measurement can fall on a boundary point. The subintervals .1 to < .6, .6 to < 1.1, and so on, are convenient and a tally is constructed.

Class i	Class Boundaries	Tally	f_i	Relative frequency, f_i/n
1	0.1 to < 0.6	11111 11111	10	.167
2	0.6 to < 1.1	11111 11111 11111	15	.250
3	1.1 to < 1.6	11111 11111 11111	15	.250
4	1.6 to < 2.1	11111 11111	10	.167
5	2.1 to < 2.6	1111	4	.067
6	2.6 to < 3.1	1	1	.017
7	3.1 to < 3.6	11	2	.033
8	3.6 to < 4.1	1	1	.017
9	4.1 to < 4.6	1	1	.017
10	4.6 to < 5.1		0	.000
11	5.1 to < 5.6	1	1	.017

The relative frequency histogram is shown below.

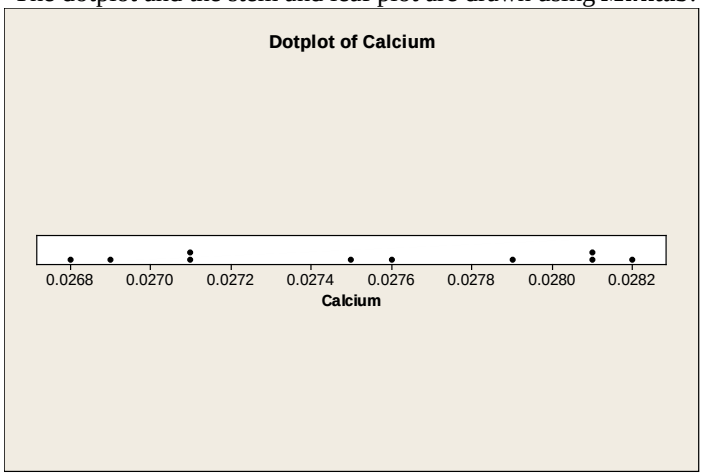


The distribution is skewed to the right, with several unusually large observations.

b For some reason, one person had to wait 5.2 minutes. Perhaps the supermarket was understaffed that day, or there may have been an unusually large number of customers in the store.

c The two graphs convey the same information. The stem and leaf plot allows us to actually recreate the actual data set, while the histogram does not.

1.32 **a-b** The dotplot and the stem and leaf plot are drawn using ***Minitab***.



Stem-and-Leaf Display: Calcium

```
Stem-and-leaf of Calcium  N  = 10
Leaf Unit = 0.00010

 2  26  89
 4  27  11
 4  27
 5  27  5
 5  27  6
 4  27  9
 3  28  11
 1  28  2
```

c The measurements all seem to be within the same range of variability. There do not appear to be any outliers.

1.33 **a** Answers will vary.

b The stem and leaf plot is constructed using the tens place as the stem and the ones place as the leaf. ***Minitab*** divides each stem into two parts to create a better descriptive picture. Notice that the distribution is roughly mound-shaped.

Stem-and-Leaf Display: Ages

```
Stem-and-leaf of Ages  N  = 38
Leaf Unit = 1.0

2   4  69
3   5  3
7   5  6678
13  6  003344
19  6  567778
19  7  011234
13  7  7889
9   8  013
6   8  58
4   9  0033
```

c Three of the five youngest presidents – Kennedy, Lincoln and Garfield – were assassinated while in office. This would explain the fact that their ages at death were in the lower tail of the distribution.

1.34 **a** We choose a stem and leaf plot, using the ones and tenths place as the stem, and a zero digit as the leaf. The ***Minitab*** printout is shown next.

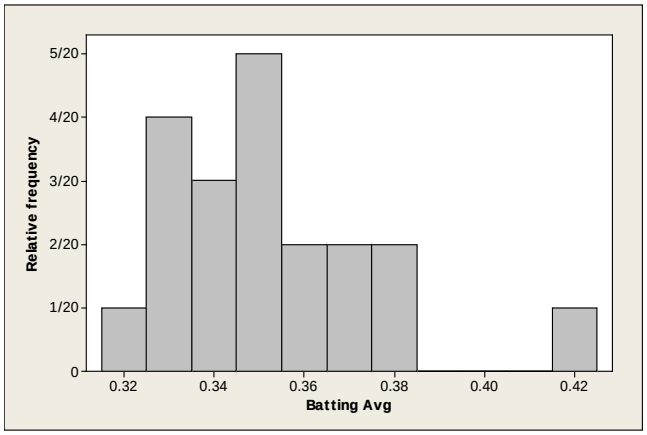
Stem-and-Leaf Display: Cells

```
Stem-and-leaf of Cells  N  = 15
Leaf Unit = 0.010
 1   49  0
 2   50  0
 3   51  0
(5)  52  00000
 7   53  000
 4   54  000
 1   55  0
```

b The data set is relatively mound-shaped, centered at 5.2.

c The value $x = 5.7$ does not fall within the range of the other cell counts, and would be considered somewhat unusual.

1.35 **a** Histograms will vary from student to student. A typical histogram, generated by ***Minitab*** is shown on the next page.



b Since 1 of the 20 players has an average above 0.400, the chance is 1 out of 20 or $1/20 = 0.05$.

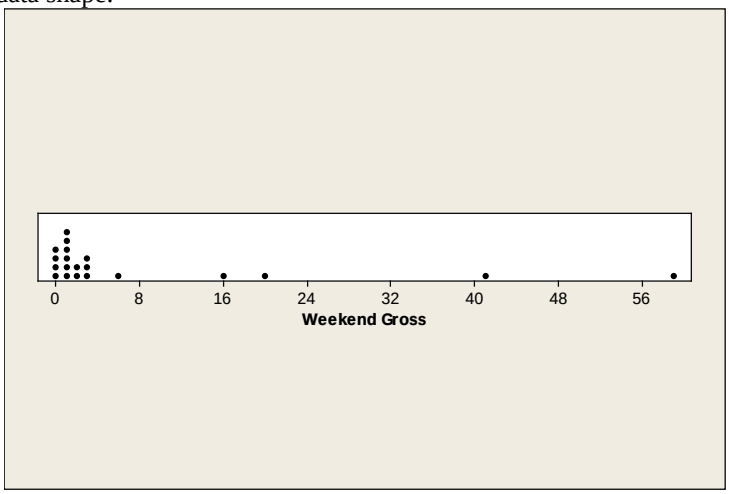
1.36 **a**

Stem-and-Leaf Display: Weekend Gross

```
Stem-and-leaf of Weekend Gross  N  = 20
Leaf Unit = 0.10

4   0  3444
7   0  556
10  1  024
10  1  69
8   2
8   2  8
7   3  11
5   3
5   4
5   4
5   5
5   5
5   6  2

HI 155, 201, 405, 593
```

The distribution is skewed to the right, with five outliers, four of which are marked by "HI" in the stem and leaf plot.

b The dotplot is more informative. Because it does not trim off the outliers, it gives a better display of the data shape.



1.37 **a** The variable being measured is a discrete variable – the number of hazardous waste sites in each of the 50 United States.

b The distribution is skewed to the right, with a several unusually large measurements. The five states marked as HI are California, Michigan, New Jersey, New York and Pennsylvania.
c Four of the five states are quite large in area, which might explain the large number of hazardous waste sites. However, the fifth state is relatively small, and other large states do not have unusually large number of waste sites. The pattern is not clear.

1.38 **a** "Ethnic origin" is a *qualitative variable* since a quality (ethnic origin) is measured.
b "Score" is a *quantitative variable* since a numerical quantity (0-100) is measured.
c "Type of establishment" is a *qualitative variable* since a category (Carl's Jr., McDonald's or Burger King) is measured.
d "Mercury concentration" is a *quantitative variable* since a numerical quantity is measured.

1.39 To determine whether a distribution is likely to be skewed, look for the likelihood of observing extremely large or extremely small values of the variable of interest.
a The distribution of non-secured loan sizes might be skewed (a few extremely large loans are possible).
b The distribution of secured loan sizes is not likely to contain unusually large or small values.
c Not likely to be skewed.
d Not likely to be skewed.
e If a package is dropped, it is likely that all the shells will be broken. Hence, a few large number of broken shells is possible. The distribution will be skewed.
f If an animal has one tick, he is likely to have more than one. There will be some "0"s with uninfected rabbits, and then a larger number of large values. The distribution will not be symmetric.

1.40 **a** The number o f homicides in Detroit during a 1-month period is a discrete random variable since it can take only the values 0, 1, 2…
b The length of time between arrivals at an outpatient clinic is a continuous random variable, since it can be any of the infinite number of positive real values.
c The number of typing errors is a discrete random variable, since it can take only the values 0, 1, 2, …
d Again, this is a discrete random variable since it can take only the values 0, 1, 2, 3, 4.
e The time required to finish an examination is a continuous random variable as was the random variable described in part **b**.

1.41 **a** Weight is continuous, taking any positive real value.
b Body temperature is continuous, taking any real value.
c Number of people is discrete, taking the values 0, 1, 2, …
d Number of properties is discrete.
e Number of claims is discrete.

1.42 **a** Number of people is discrete, taking the values 0, 1, 2, …
b Depth is continuous, taking any non-negative real value.
c Length of time is continuous, taking any non-negative real value.
d Number of aircraft is discrete.

1.43 Stem and leaf displays may vary from student to student. The most obvious choice is to use the tens digit as the stem and the ones digit as the leaf.

```
 7 | 8 9
 8 | 0 1 7
 9 | 0 1 2 4 4 5 6 6 6 8 8
10 | 1 7 9
11 | 2
```

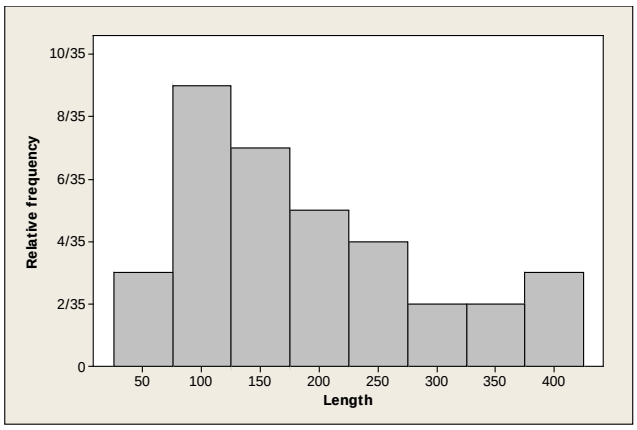
The display is fairly mound-shaped, with a large peak in the middle.

1.44 **a**

Stem-and-Leaf Display: Length

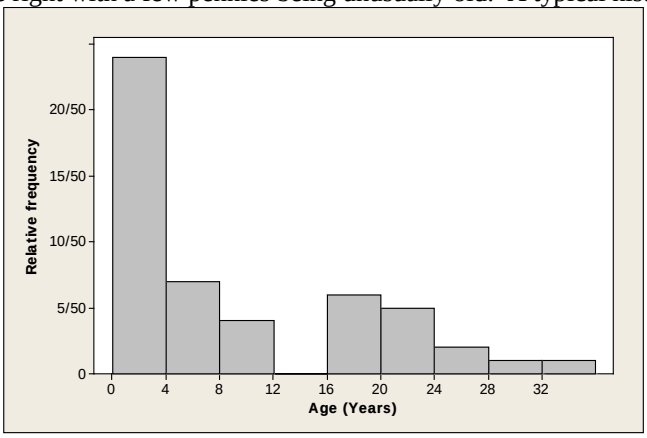
```
Stem-and-leaf of Length  N  = 35
Leaf Unit = 10

 7    0  6779999
(11)  1  00122334444
 17   1  5799
 13   2  004
 10   2  5669
 6    3  0
 5    3  5679
 1    4  2
```

b

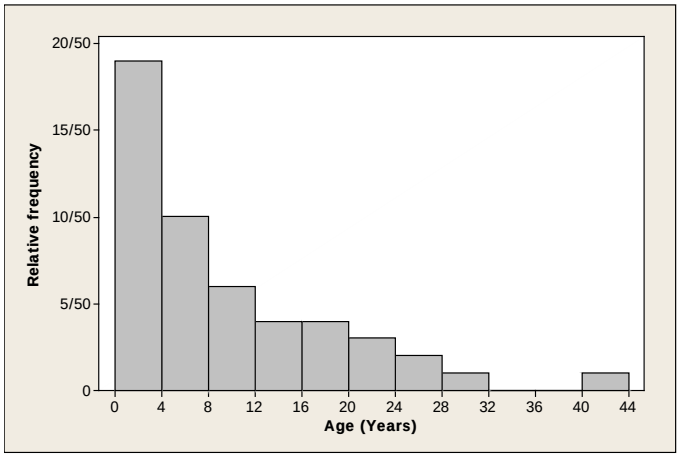


c These data are skewed right.

1.45 **a-b** Answers will vary from student to student. The students should notice that the distribution is skewed to the right with a few pennies being unusually old. A typical histogram is shown below.



1.46 **a** Answers will vary from student to student. A typical histogram is shown on the next page. It looks very similar to the histogram from Exercise 1.45.

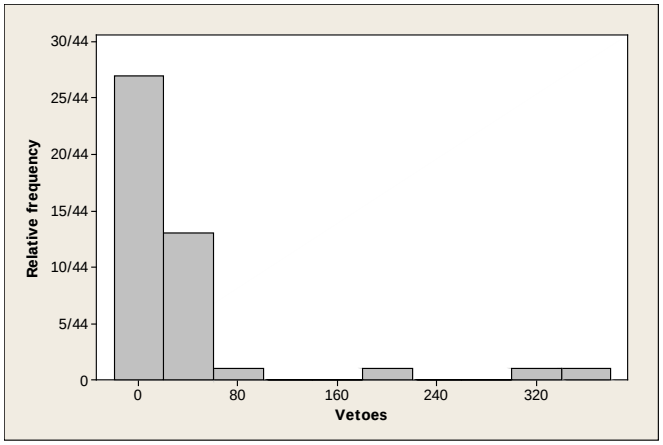


b The stem and leaf plot is drawn using ***Minitab***. There is one outlier, $x = 41$.

Stem-and-Leaf Display: Age (Years)

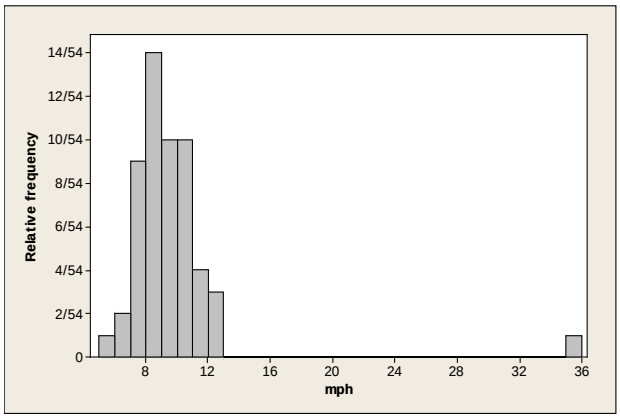
```
Stem-and-leaf of Age (Years)  N  = 50
Leaf Unit = 1.0
 9   0  000000011
 19  0  2223333333
(7)  0  4444555
 24  0  777
 21  0  88999
 16  1  0
 15  1  2
 14  1  444
 11  1  677
 8   1  9
 7   2  01
 5   2  3
 4   2  45
 2   2
 2   2  8

HI 41
```

1.47 Answers will vary from student to student. The students should notice that the distribution is skewed to the right with a few presidents (Truman, Cleveland, and F.D. Roosevelt) casting an unusually large number of vetoes.



1.48 **a** Answers will vary from student to student. The relative frequency histogram below was constructed using classes of length 1.0 starting at $x = 5$. The value $x = 35.1$ is not shown in the table, but appears on the graph on the next page.

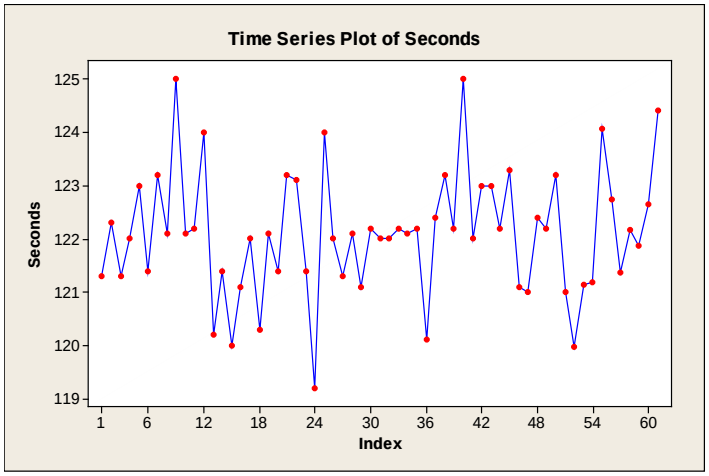
Class i	Class Boundaries	Tally	f_i	Relative frequency, f_i/n
1	5.0 to < 6.0	1	1	1/54
2	6.0 to < 7.0	11	2	2/54
3	7.0 to < 8.0	11111 1111	9	9/54
4	8.0 to < 9.0	11111 11111 1111	14	14/54
5	9.0 to < 10.0	11111 11111	10	10/54
6	10.0 to < 11.0	11111 11111	10	10/54
7	11.0 to < 12.0	1111	4	4/54
8	12.0 to < 13.0	111	3	3/54



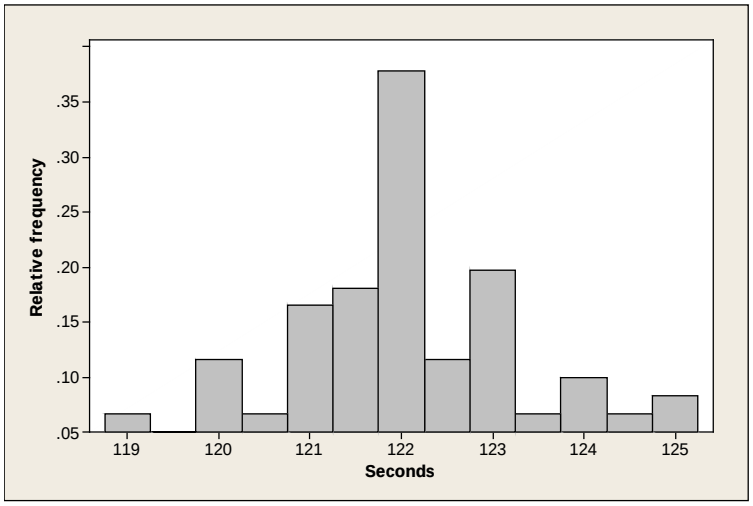
b Since Mt. Washington is a very mountainous area, it is not unusual that the average wind speed would be very high.

c The value $x = 10.3$ does not lie far from the center of the distribution (excluding $x = 35.1$). It would not be considered unusually high.

1.49 **a** The line chart is shown below. The year in which a horse raced does not appear to have an effect on his winning time.



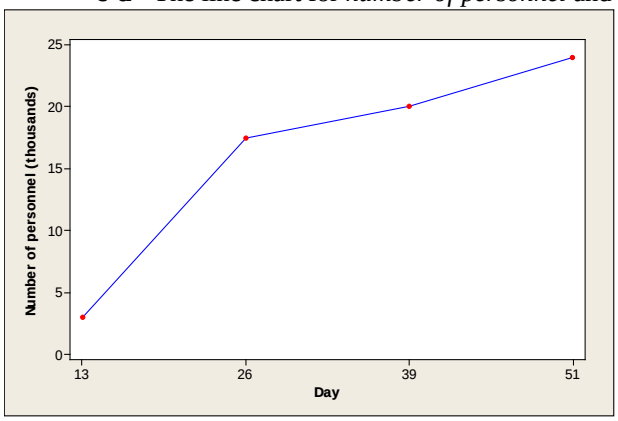
b Since the year of the race is not important in describing the data set, the distribution can be described using a relative frequency histogram. The distribution shown below is roughly mound-shaped with an unusually fast ($x = 119.2$) race times the year that *Secretariat* won the derby.

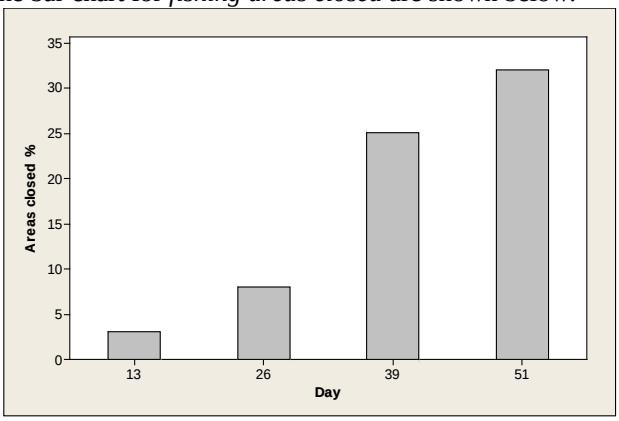


1.50 **a** The five quantitative variables are measured over time two months after the oil spill. Some sort of comparative bar charts (side-by-side or stacked) or a line chart should be used.

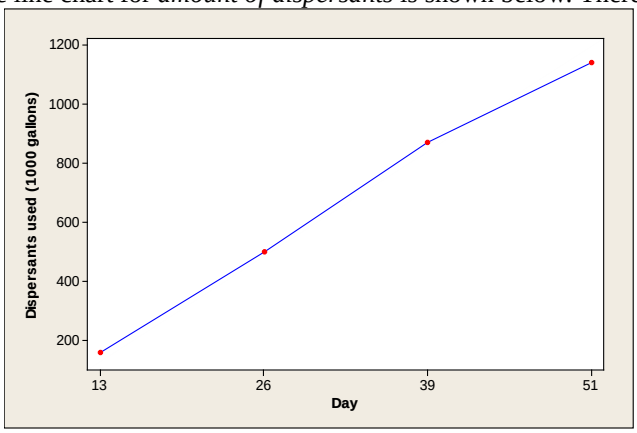
b As the time after the spill increases, the values of all five variables increase.

c-d The line chart for *number of personnel* and the bar chart for *fishing areas closed* are shown below.



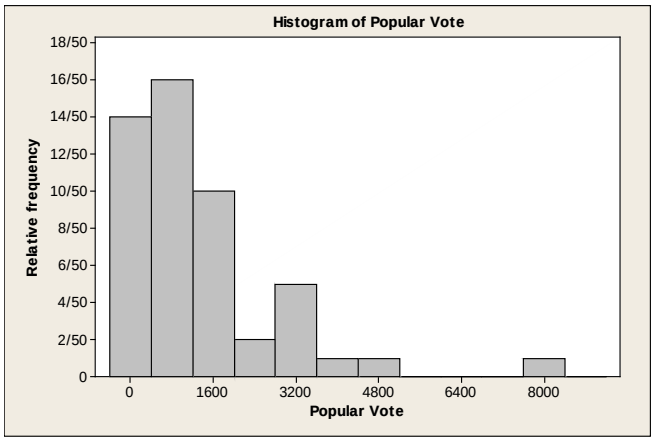


e The line chart for *amount of dispersants* is shown below. There appears to be a straight line trend.

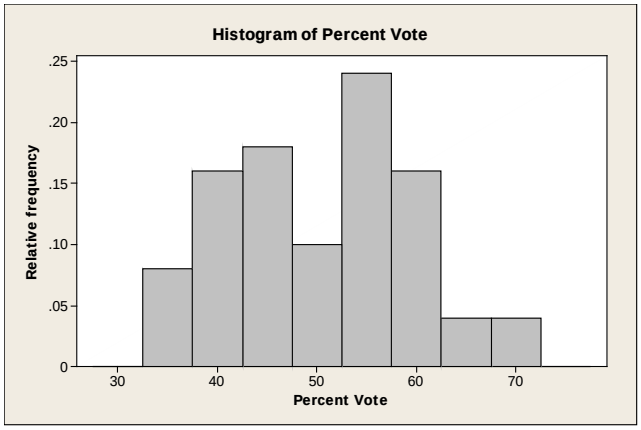


1.51 **a** The popular vote within each state should vary depending on the size of the state. Since there are several very large states (in population) in the United States, the distribution should be skewed to the right.

b-c Histograms will vary from student to student, but should resemble the histogram generated by ***Minitab*** in the figure on the next page. The distribution is indeed skewed to the right, with one "outlier" – California (and possibly Florida and New York).



1.52 **a-b** Once the size of the state is removed by calculating the percentage of the popular vote, the unusually large values in the Exercise 1.51 data set will disappear, and each state will be measured on an equal basis. Student histograms should resemble the histogram shown below. Notice the relatively mound-shape and the lack of any outliers.



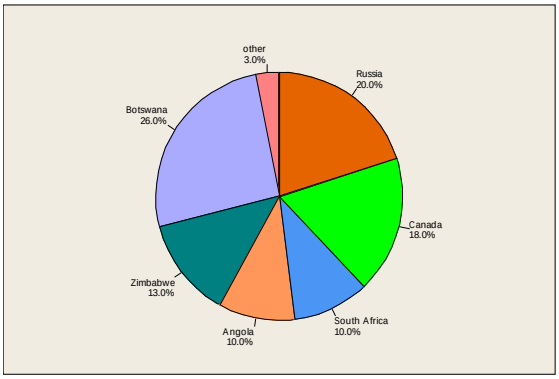
1.53 **a-b** Popular vote is skewed to the right while the percentage of popular vote is roughly mound-shaped. While the distribution of popular vote has outliers (California, Florida and New York), there are no outliers in the distribution of percentage of popular vote. When the stem and leaf plots are turned 90^{o}, the shapes are very similar to the histograms.

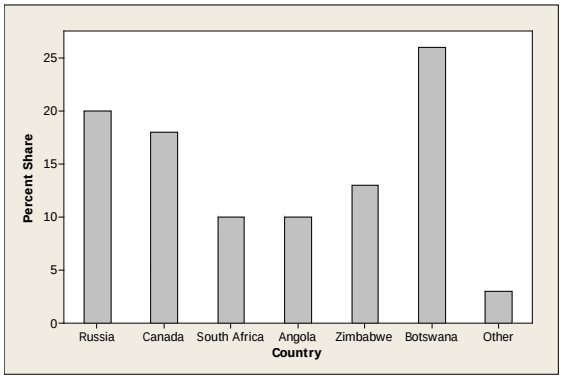
c Once the size of the state is removed by calculating the percentage of the popular vote, the unusually large values in the set of "popular votes" will disappear, and each state will be measured on an equal basis. The data then distribute themselves in a mound-shape around the average percentage of the popular vote.

1.54 **a-b** The data is somewhat mound-shaped, but it appears to have two local peaks – high points from which the frequencies drop off on either side.

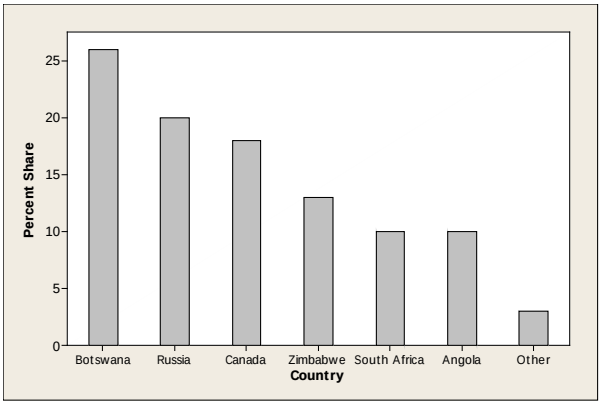
c Since these are student heights, the data can be divided into two groups – heights of males and heights of females. Both groups will have an approximate mound-shape, but the average female height will be lower than the average male height. When the two groups are combined into one data set, it causes a "mixture" of two mound-shaped distributions, and produces the bimodal distribution seen in the exercise.

1.55 **a-b** Answers will vary from student to student. Since the graph gives a range of values for Zimbabwe's share, we have chosen to use the 13% figure, and have used 3% in the "Other" category. The pie chart and bar charts are shown on the next page.





c-d The Pareto chart is shown below. Either the pie chart or the Pareto chart is more effective than the bar chart.



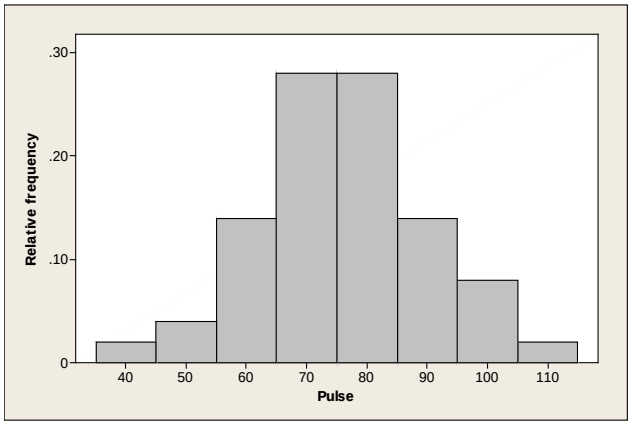
1.56 **a** The measurements are obtained by counting the number of beats for 30 seconds, and then multiplying by 2. Thus, the measurements should all be even numbers.

b The stem and leaf plot is shown below.

Stem-and-Leaf Display: Pulse

```
Stem-and-leaf of Pulse  N  = 50
Leaf Unit = 1.0
 1   4   2
 1   4
 3   5   24
 6   5   688
 10  6   0022
 15  6   66668
 24  7   000222224
 25  7   8
 25  8   0022444444444
 12  8   68888
 7   9   00
 5   9   66
 3   10  04
 1   10
 1   11  0
```

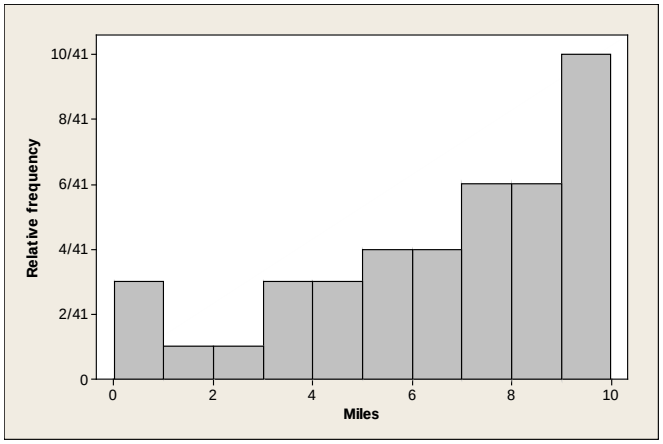
c Answers will vary. A typical histogram, generated by ***Minitab***, is shown on the next page.



d The distribution of pulse rates is mound-shaped and relatively symmetric around a central location of 75 beats per minute. There are no outliers.

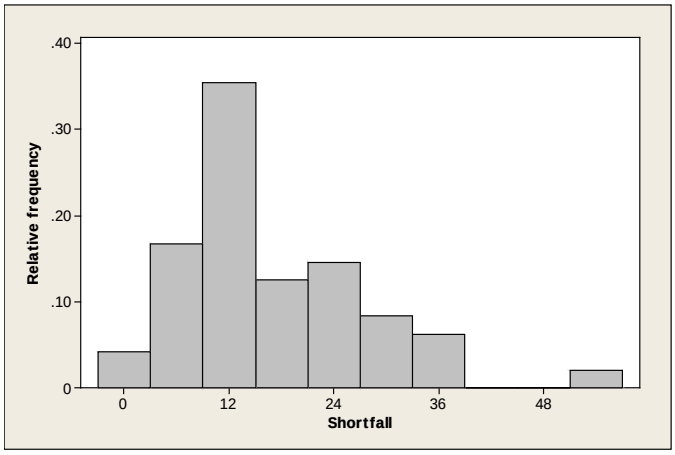
1.57 The relative frequency histogram below was constructed using classes of length 1.0 starting at $x = 0.0$.

Class i	Class Boundaries	Tally	f_i	Relative frequency, f_i/n
1	0.0 to < 1.0	111	3	3/41
2	1.0 to < 2.0	1	1	1/41
3	2.0 to < 3.0	1	1	1/41
4	3.0 to < 4.0	111	3	3/41
5	4.0 to < 5.0	111	3	3/41
6	5.0 to < 6.0	1111	4	4/41
7	6.0 to < 7.0	1111	4	4/41
8	7.0 to < 8.0	11111 1	6	6/41
9	8.0 to < 9.0	11111 1	6	6/41
10	9.0 to < 10.0	11111 11111	10	10/41



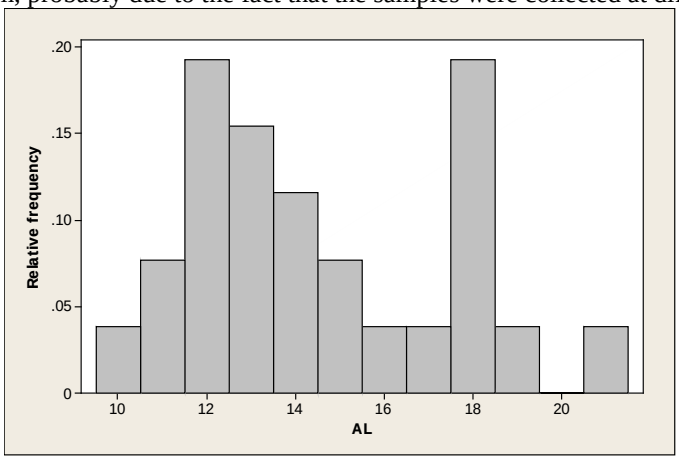
a The distribution is skewed to the left, with an unusual peak in the first class (within one mile of UCR).
b As the distance from UCR increases, each successive area increases in size, thus allowing for more Starbucks stores in that region.

1.58 **a-b** Answers will vary from student to student. The distribution is skewed to the right, with an extreme outlier (Nevada) in the upper part of the distribution. A typical histogram is shown on the next page.

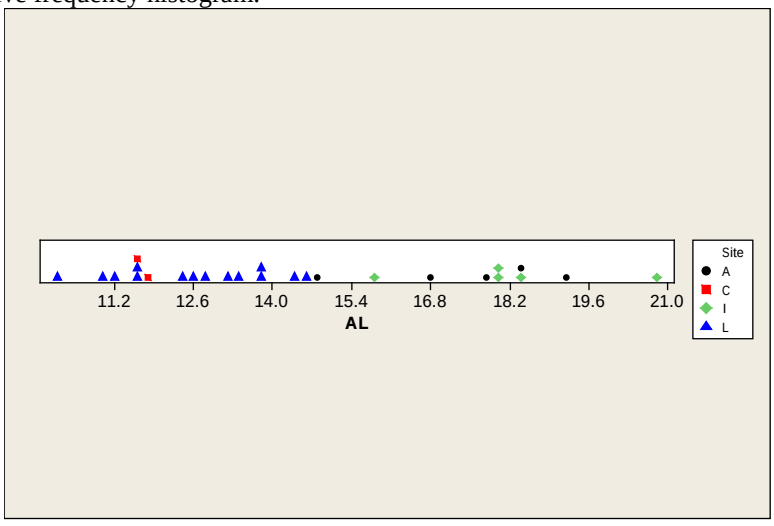


c Answers will vary. Perhaps the scarcity of the population in those three states means that there are fewer people who need to use the state's governmental services.

1.59 **a-b** Answers will vary. A typical histogram is shown below. Notice the gaps and the bimodal nature of the histogram, probably due to the fact that the samples were collected at different locations.

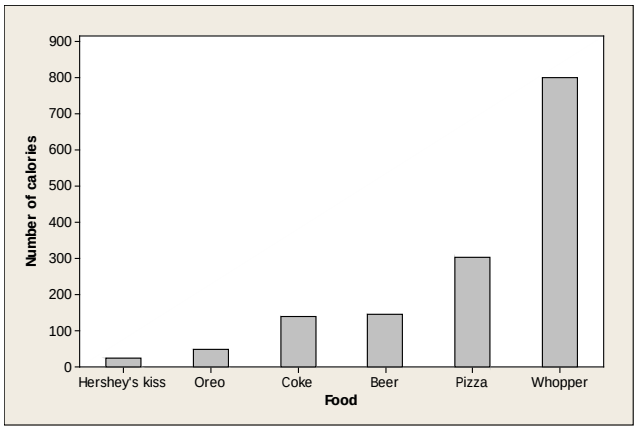


c The dotplot is shown below. The locations are indeed responsible for the unusual gaps and peaks in the relative frequency histogram.



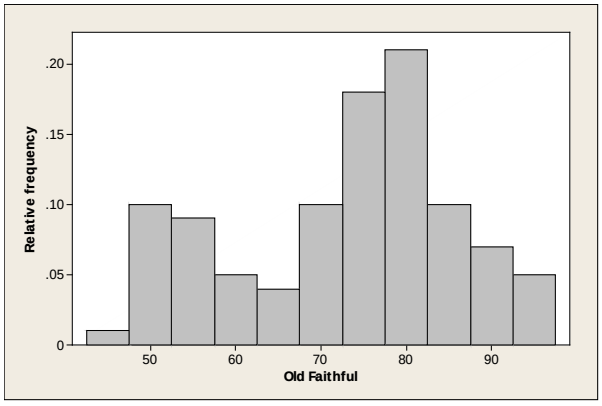
1.60 **a** The sizes and volumes of the food items do increase as the number of calories increase, but not in the correct proportion to the actual calories. The differences in calorie content are not accurately portrayed in the graph.

b The bar graph which accurately portrays the number of calories in the six food items is shown below.



1.61 Answers will vary from student to student. Students should notice that both distributions are skewed left. The higher peak with a low bar to its left in the laptop group may indicate that students who would generally receive average scores (65-75) are scoring higher than usual. This may or may not be *caused* by the fact that they used laptop computers.

1.62 Answers will vary. A typical relative frequency histogram is shown below. There is an unusual bimodal feature.



1.63 **a-b** The ***Minitab*** stem and leaf plot is shown below. The distribution is slightly skewed to the right.

Stem-and-Leaf Display: Tax

```
Stem-and-leaf of Tax  N  = 51
Leaf Unit = 1.0

 1    2  6
 3    3  22
 16   3  5557778888999
(15)  4  000011111223333
 20   4  566689
 14   5  00111234
 6    5  58
 4    6  133
 1    6  7
```

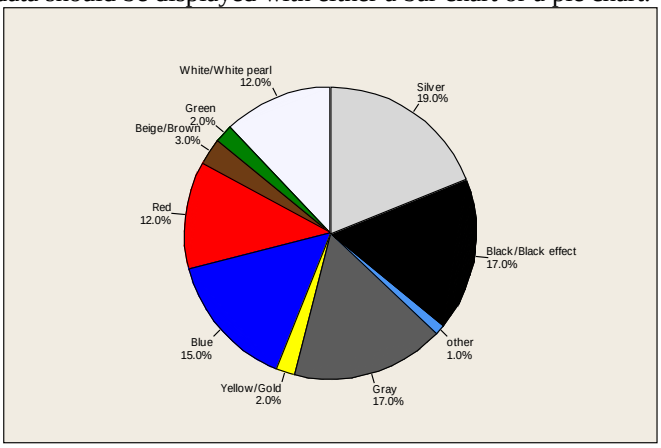
c Arkansas (26.4), Wyoming (32.4) and New Jersey (32.9) have gasoline taxes that are somewhat smaller than most, but they are not "outliers" in the sense that they lie far away from the rest of the measurements in the data set.

1.64 **a-b** Answers will vary. The ***Minitab*** stem and leaf plot is shown below. The distribution is skewed to the right.

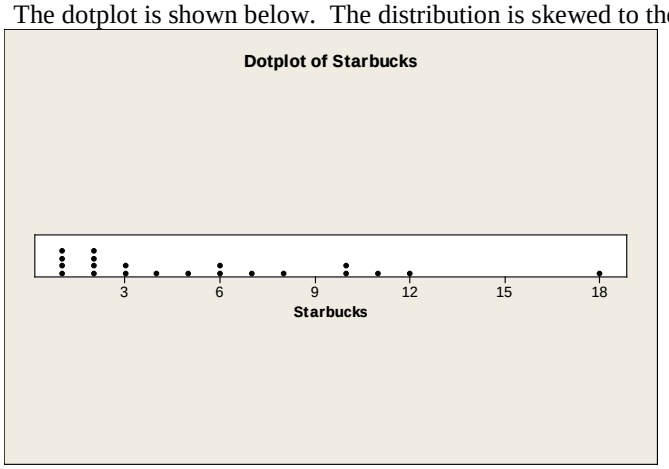
Stem-and-Leaf Display: Megawatts

```
Stem-and-leaf of Megawatts  N  = 20
Leaf Unit = 1000

 10  0  2222233333
 10  0  444
 7   0  666
 4   0  8
 3   1  0
 2   1  2
 1   1
 1   1
 1   1  8
```

1.65 The data should be displayed with either a bar chart or a pie chart. The pie chart is shown below.



1.66 **a-b** The dotplot is shown below. The distribution is skewed to the right.



c The Starbucks chain, which serves somewhat higher priced beverages, may be targeting clientele with higher median incomes than the typical American. Cities with higher median incomes or simply cities with larger populations may be more likely to have a larger number of Starbucks stores.

1.67 **a-b** The distribution is approximately mound-shaped, with one unusual measurement, in the class with midpoint at 100.8°. Perhaps the person whose temperature was 100.8 has some sort of illness coming on?
c The value 98.6° is slightly to the right of center.

CASE STUDY: How is Your Blood Pressure?

1. The following variables have been measured on the participants in this study: sex (qualitative); age in years (quantitative discrete); diastolic blood pressure (quantitative continuous, but measured to an integer value) and systolic blood pressure (quantitative continuous, but measured to an integer value). For each person, both systolic and diastolic readings are taken, making the data bivariate.

2. The important variables in this study are diastolic and systolic blood pressure, which can be described singly with histograms in various categories (male vs. female or by age categories). Further, the relationship between systolic and diastolic blood pressure can be displayed together using a scatterplot or a bivariate histogram.

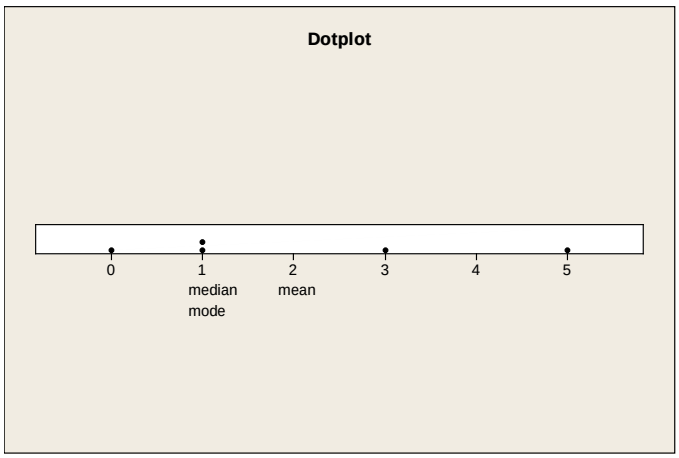
3. Answers will vary from student to student, depending on the choice of class boundaries or the software package which is used. The histograms should look fairly mound-shaped.

4. Answers will vary from student to student.

5. In determining how a student's blood pressure compares to those in a comparable sex and age group, female students (ages 15-20) must compare to the population of females, while male students (ages 15-20) must compare to the population of males. The student should use his or her blood pressure and compare it to the scatterplot generated in part 4.

2: Describing Data with Numerical Measures

2.1 **a** The dotplot shown below plots the five measurements along the horizontal axis. Since there are two "1"s, the corresponding dots are placed one above the other. The approximate center of the data appears to be around 1.



b The mean is the sum of the measurements divided by the number of measurements, or

$$\overline{x} = \frac{\sum x_i}{n} = \frac{0+5+1+1+3}{5} = \frac{10}{5} = 2$$

To calculate the median, the observations are first ranked from smallest to largest: 0, 1, 1, 3, 5. Then since $n = 5$, the position of the median is $0.5(n+1) = 3$, and the median is the 3rd ranked measurement, or $m = 1$. The mode is the measurement occurring most frequently, or mode = 1.

c The three measures in part **b** are located on the dotplot. Since the median and mode are to the left of the mean, we conclude that the measurements are skewed to the right.

2.2 **a** The mean is

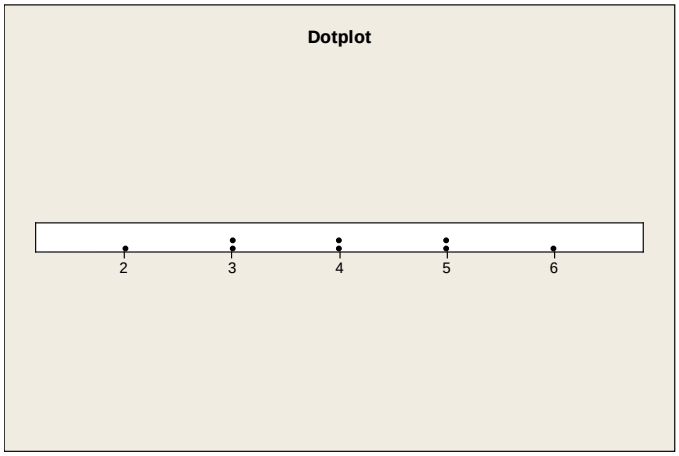
$$\overline{x} = \frac{\sum x_i}{n} = \frac{3+2+\cdots+5}{8} = \frac{32}{8} = 4$$

b To calculate the median, the observations are first ranked from smallest to largest:

2, 3, 3, 4, 4, 5, 5, 6

Since $n = 8$ is even, the position of the median is $0.5(n+1) = 4.5$, and the median is the average of the 4th and 5th measurements, or $m = (4+4)/2 = 4$.

c Since the mean and the median are equal, we conclude that the measurements are symmetric. The dotplot shown below confirms this conclusion.



2.3 **a** $\bar{x} = \dfrac{\sum x_i}{n} = \dfrac{58}{10} = 5.8$

b The ranked observations are: 2, 3, 4, 5, 5, 6, 6, 8, 9, 10. Since $n = 10$, the median is halfway between the 5th and 6th ordered observations, or $m = (5+6)/2 = 5.5$.

c There are two measurements, 5 and 6, which both occur twice. Since this is the highest frequency of occurrence for the data set, we say that the set is *bimodal* with modes at 5 and 6.

2.4 **a** $\bar{x} = \dfrac{\sum x_i}{n} = \dfrac{9715}{4} = 2428.75$ **b** $\bar{x} = \dfrac{\sum x_i}{n} = \dfrac{9618}{4} = 2404.50$

c The average premium cost in several different cities is not as important to the consumer as the average cost for a variety of consumers in his or her geographical area.

2.5 **a** Although there may be a few households who own more than one DVR, the majority should own either 0 or 1. The distribution should be slightly skewed to the right.

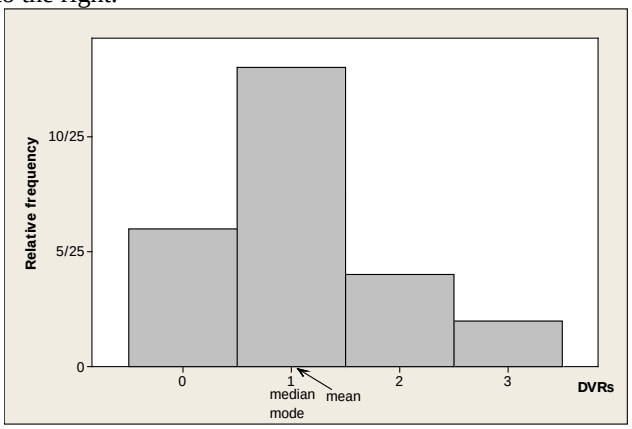
b Since most households will have only one DVR, we guess that the mode is 1.

c The mean is

$$\bar{x} = \frac{\sum x_i}{n} = \frac{1+0+\cdots+1}{25} = \frac{27}{25} = 1.08$$

To calculate the median, the observations are first ranked from smallest to largest: There are six 0s, thirteen 1s, four 2s, and two 3s. Then since $n = 25$, the position of the median is $0.5(n+1) = 13$, which is the 13th ranked measurement, or $m = 1$. The mode is the measurement occurring most frequently, or mode = 1.

d The relative frequency histogram is shown below, with the three measures superimposed. Notice that the mean falls slightly to the right of the median and mode, indicating that the measurements are slightly skewed to the right.



2.6 **a** The stem and leaf plot below was generated by ***Minitab***. It is skewed to the right.

Stem-and-Leaf Display: Revenues

```
Stem-and-leaf of Revenues  N  = 10
Leaf Unit = 10000

 2   0  33
(5)  0  66669
 3   1  0
 2   1  5
 1   2
 1   2  8
```

b The mean is

$$\bar{x} = \frac{\sum x_i}{n} = \frac{104589+95758+\cdots+284650}{10} = \frac{965180}{10} = 96,518$$

To calculate the median, rank the observations from smallest to largest.

31,515	36,537	61,867	65,357	66,176
68,281	95,758	104,589	150,450	284,650

Then since $n = 10$, the position of the median is $0.5(n+1) = 5.5$, the average of the 5th and 6th ranked measurements or $m = (66176 + 68281)/2 = 67,228.5$.

c Since the mean is strongly affected by outliers, the median would be a better measure of center for this data set.

2.7 It is obvious that any one family cannot have 2.5 children, since the number of children per family is a quantitative discrete variable. The researcher is referring to the average number of children per family calculated for all families in the United States during the 1930s. The average does not necessarily have to be integer-valued.

2.8 **a** Similar to previous exercises. The mean is

$$\bar{x} = \frac{\sum x_i}{n} = \frac{0.99 + 1.92 + \cdots + 0.66}{14} = \frac{12.55}{14} = 0.896$$

b To calculate the median, rank the observations from smallest to largest. The position of the median is $0.5(n+1) = 7.5$, and the median is the average of the 7th and 8th ranked measurement or $m = (0.67 + 0.69)/2 = 0.68$.

c Since the mean is slightly larger than the median, the distribution is slightly skewed to the right.

2.9 The distribution of sports salaries will be skewed to the right, because of the very high salaries of some sports figures. Hence, the median salary would be a better measure of center than the mean.

2.10 **a** Similar to previous exercises.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{2150}{10} = 215$$

b The ranked observations are shown below:

175	225
185	230
190	240
190	250
200	265

The position of the median is $0.5(n+1) = 5.5$ and the median is the average of the 5th and 6th observation or

$$\frac{200 + 225}{2} = 212.5$$

c Since there are no unusually large or small observations to affect the value of the mean, we would probably report the mean or average time on task.

2.11 **a** Similar to previous exercises.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{115}{21} = 5.476$$

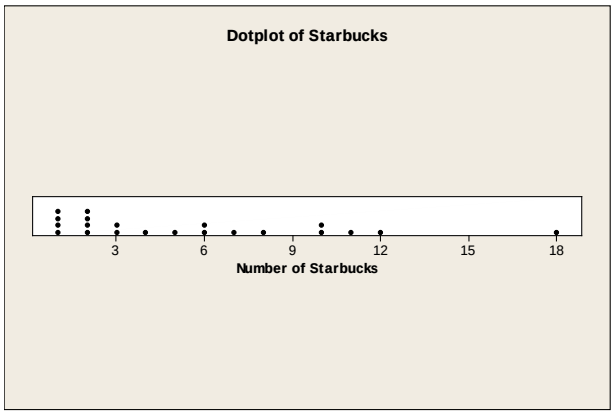
The ranked observations are:

1	2	3	6	10	18
1	2	3	6	10	
1	2	4	7	11	
1	2	5	8	12	

The position of the median is $0.5(n+1) = 11$, and the median is the 11th observation or $m = 4$. There are two observations, 1 and 2, both of which occur four times. Hence, the data set is *bimodal*—it has two modes, 1 and 2.

b Since the mean is larger than the median, the data are skewed to the right.

c The dotplot is shown below. The distribution is skewed to the right.



2.12 **a** $\bar{x} = \dfrac{\sum x_i}{n} = \dfrac{4513.75}{14} = 322.41$

b The ranked data are: 180.17, 184.86, 216.49, 222.84, 222.84, 231.04, 262.95, 279.90, 280.98, 289.97, 299.48, 384.99, 433.00, 1034.24 and the median is the average of the 7th and 8th observations or

$$m = \frac{262.95 + 279.90}{2} = 271.425$$

c Average cost would not be as important as minimum cost, but the consumer needs to consider many other variables, such as shipping costs, warranties, customer service for returns or damaged items and so on.

2.13 **a** $\bar{x} = \dfrac{\sum x_i}{n} = \dfrac{12}{5} = 2.4$

b Create a table of differences, $(x_i - \bar{x})$ and their squares, $(x_i - \bar{x})^2$.

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
2	–0.4	0.16
1	–1.4	1.96
1	–1.4	1.96
3	0.6	0.36
5	2.6	6.76
Total	0	11.20

Then

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{(2-2.4)^2 + \cdots + (5-2.4)^2}{4} = \frac{11.20}{4} = 2.8$$

c The sample standard deviation is the positive square root of the variance or

$$s = \sqrt{s^2} = \sqrt{2.8} = 1.673$$

d Calculate $\sum x_i^2 = 2^2 + 1^2 + \cdots + 5^2 = 40$. Then

$$s^2 = \frac{\sum x_i^2 - \dfrac{(\sum x_i)^2}{n}}{n-1} = \frac{40 - \dfrac{(12)^2}{5}}{4} = \frac{11.2}{4} = 2.8 \text{ and } s = \sqrt{s^2} = \sqrt{2.8} = 1.673.$$

The results of parts **a** and **b** are identical.

2.14 The results will vary from student to student, depending on their particular type of calculator. The results should agree with Exercise 2.13.

2.15 **a** The range is $R = 4 - 1 = 3$. **b** $\bar{x} = \dfrac{\sum x_i}{n} = \dfrac{17}{8} = 2.125$

c Calculate $\sum x_i^2 = 4^2 + 1^2 + \cdots + 2^2 = 45$. Then

$$s^2 = \frac{\sum x_i^2 - \dfrac{\left(\sum x_i\right)^2}{n}}{n-1} = \frac{45 - \dfrac{(17)^2}{8}}{7} = \frac{8.875}{7} = 1.2679 \text{ and } s = \sqrt{s^2} = \sqrt{1.2679} = 1.126.$$

2.16 **a** The range is $R = 6 - 1 = 5$. **b** $\bar{x} = \dfrac{\sum x_i}{n} = \dfrac{31}{8} = 3.875$

c Calculate $\sum x_i^2 = 3^2 + 1^2 + \cdots + 5^2 = 137$. Then

$$s^2 = \frac{\sum x_i^2 - \dfrac{\left(\sum x_i\right)^2}{n}}{n-1} = \frac{137 - \dfrac{(31)^2}{8}}{7} = \frac{16.875}{7} = 2.4107$$

and $s = \sqrt{s^2} = \sqrt{2.4107} = 1.55$.

d The range, $R = 5$, is $5/1.55 = 3.23$ standard deviations.

2.17 **a** The range is $R = 2.39 - 1.28 = 1.11$.

b Calculate $\sum x_i^2 = 1.28^2 + 2.39^2 + \cdots + 1.51^2 = 15.415$. Then

$$s^2 = \frac{\sum x_i^2 - \dfrac{\left(\sum x_i\right)^2}{n}}{n-1} = \frac{15.415 - \dfrac{(8.56)^2}{5}}{4} = \frac{.76028}{4} = .19007$$

and $s = \sqrt{s^2} = \sqrt{.19007} = .436$

c The range, $R = 1.11$, is $1.11/.436 = 2.5$ standard deviations.

2.18 **a** The range is $R = 370.23 - 216.85 = 153.38$. **b** $\bar{x} = \dfrac{\sum x_i}{n} = \dfrac{3426.64}{12} = 285.553$

c Calculate $\sum x_i^2 = 288.02^2 + 230.60^2 + \cdots + 298.12^2 = 1{,}005{,}678.51$. Then

$$s^2 = \frac{\sum x_i^2 - \dfrac{\left(\sum x_i\right)^2}{n}}{n-1} = \frac{1005678.51 - \dfrac{(3426.64)^2}{12}}{11} = 2471.821752$$

and $s = \sqrt{s^2} = \sqrt{2471.821752} = 49.7174$.

2.19 **a** The range of the data is $R = 6 - 1 = 5$ and the range approximation with $n = 10$ is

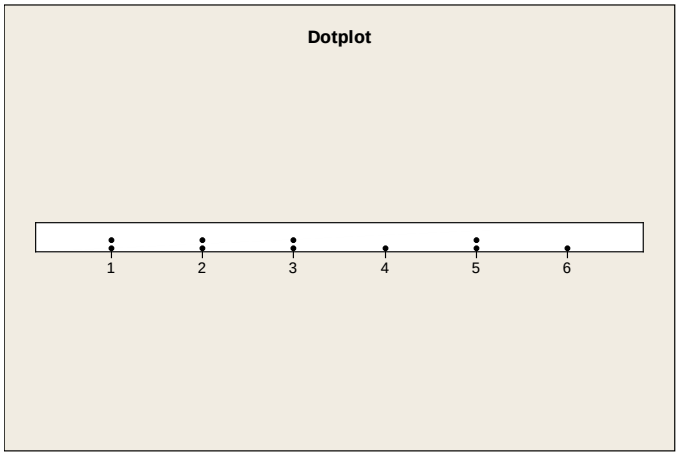
$$s \approx \frac{R}{3} = 1.67$$

b The standard deviation of the sample is

$$s = \sqrt{s^2} = \sqrt{\frac{\sum x_i^2 - \dfrac{\left(\sum x_i\right)^2}{n}}{n-1}} = \sqrt{\frac{130 - \dfrac{(32)^2}{10}}{9}} = \sqrt{3.0667} = 1.751$$

which is very close to the estimate for part **a**.

c-e From the dotplot on the next page, you can see that the data set is not mound-shaped. Hence you can use Tchebysheff's Theorem, but not the Empirical Rule to describe the data.



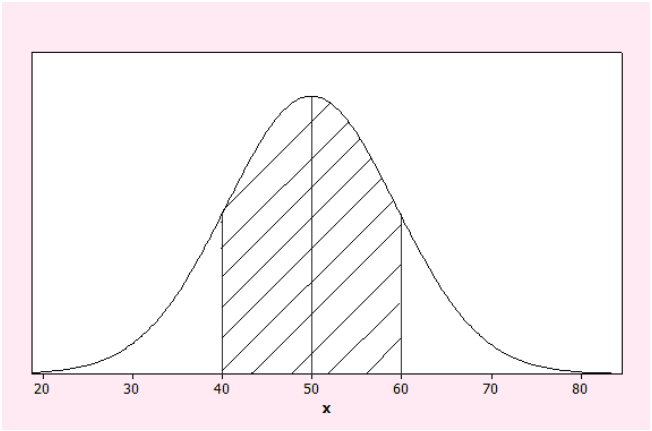
2.20 **a** First calculate the intervals:

$$\bar{x} \pm s = 36 \pm 3 \text{ or } 33 \text{ to } 39$$
$$\bar{x} \pm 2s = 36 \pm 6 \text{ or } 30 \text{ to } 42$$
$$\bar{x} \pm 3s = 36 \pm 9 \text{ or } 27 \text{ to } 45$$

According to the Empirical Rule, approximately 68% of the measurements will fall in the interval 33 to 39; approximately 95% of the measurements will fall between 30 and 42; approximately 99.7% of the measurements will fall between 27 and 45.

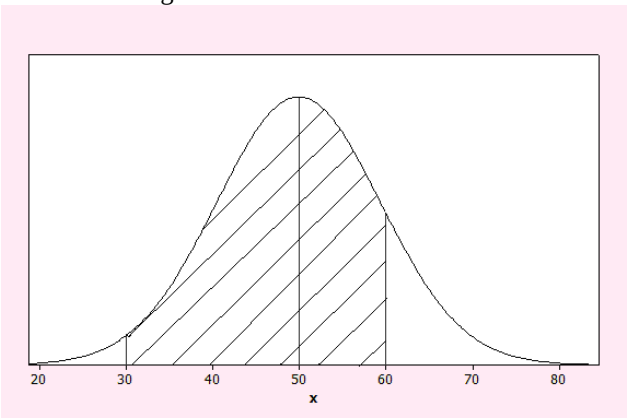
b If no prior information as to the shape of the distribution is available, we use Tchebysheff's Theorem. We would expect at least $\left(1-1/1^2\right) = 0$ of the measurements to fall in the interval 33 to 39; at least $\left(1-1/2^2\right) = 3/4$ of the measurements to fall in the interval 30 to 42; at least $\left(1-1/3^2\right) = 8/9$ of the measurements to fall in the interval 27 to 45.

2.21 **a** The interval from 40 to 60 represents $\mu \pm \sigma = 50 \pm 10$. Since the distribution is relatively mound-shaped, the proportion of measurements between 40 and 60 is .68 (or 68%) according to the Empirical Rule and is shown below.



b Again, using the Empirical Rule, the interval $\mu \pm 2\sigma = 50 \pm 2(10)$ or between 30 and 70 contains approximately .95 (or 95%) of the measurements.

c Refer to the figure below.



Since approximately 68% of the measurements are between 40 and 60, the symmetry of the distribution implies that 34% of the measurements are between 50 and 60. Similarly, since 95% of the measurements are between 30 and 70, approximately 47.5% are between 30 and 50. Thus, the proportion of measurements between 30 and 60 is

$$0.34 + 0.475 = 0.815$$

d From the figure in part **a**, the proportion of the measurements between 50 and 60 is 0.34 and the proportion of the measurements which are greater than 50 is 0.50. Therefore, the proportion that is greater than 60 must be

$$0.5 - 0.34 = 0.16$$

2.22 Since nothing is known about the shape of the data distribution, you must use Tchebysheff's Theorem to describe the data.

a The interval from 60 to 90 represents $\mu \pm 3\sigma$ which will contain at least 8/9 of the measurements.

b The interval from 65 to 85 represents $\mu \pm 2\sigma$ which will contain at least 3/4 of the measurements.

c The value $x = 65$ lies two standard deviations below the mean. Since at least 3/4 of the measurements are within two standard deviation range, *at most* 1/4 can lie outside this range, which means that at most 1/4 can be less than 65.

2.23 **a** The range of the data is $R = 1.1 - 0.5 = 0.6$ and the approximate value of s is

$$s \approx \frac{R}{3} = 0.2$$

b Calculate $\sum x_i = 7.6$ and $\sum x_i^2 = 6.02$, the sample mean is

$$\overline{x} = \frac{\sum x_i}{n} = \frac{7.6}{10} = .76$$

and the standard deviation of the sample is

$$s = \sqrt{s^2} = \sqrt{\frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1}} = \sqrt{\frac{6.02 - \frac{(7.6)^2}{10}}{9}} = \sqrt{\frac{0.244}{9}} = 0.165$$

which is very close to the estimate from part **a**.

2.24 **a** The stem and leaf plot (on the next page) generated by ***Minitab*** shows that the data is roughly mound-shaped. Note however the gap in the center of the distribution and the two measurements in the upper tail.

Stem-and-Leaf Display: Weight

```
Stem-and-leaf of Weight  N  = 27
Leaf Unit = 0.010

 1   7   5
 2   8   3
 6   8   7999
 8   9   23
 13  9   66789
 13  10
(3)  10  688
 11  11  2244
 7   11  788
 4   12  4
 3   12  8
 2   13
 2   13  8
 1   14  1
```

b Calculate $\sum x_i = 28.41$ and $\sum x_i^2 = 30.6071$, the sample mean is $\overline{x} = \dfrac{\sum x_i}{n} = \dfrac{28.41}{27} = 1.052$ and the standard deviation of the sample is

$$s = \sqrt{s^2} = \sqrt{\frac{\sum x_i^2 - \dfrac{\left(\sum x_i\right)^2}{n}}{n-1}} = \sqrt{\frac{30.6071 - \dfrac{(28.41)^2}{27}}{26}} = 0.166$$

c The following table gives the actual percentage of measurements falling in the intervals $\overline{x} \pm ks$ for $k = 1, 2, 3$.

k	$\overline{x} \pm ks$	Interval	Number in Interval	Percentage
1	1.052 ± 0.166	0.866 to 1.218	21	78%
2	1.052 ± 0.332	0.720 to 1.384	26	96%
3	1.052 ± 0.498	0.554 to 1.550	27	100%

d The percentages in part **c** do not agree too closely with those given by the Empirical Rule, especially in the one standard deviation range. This is caused by the lack of mounding (indicated by the gap) in the center of the distribution.

e The lack of any one-pound packages is probably a marketing technique intentionally used by the supermarket. People who buy slightly less than one-pound would be drawn by the slightly lower price, while those who need exactly one-pound of meat for their recipe might tend to opt for the larger package, increasing the store's profit.

2.25 According to the Empirical Rule, if a distribution of measurements is approximately mound-shaped,

a approximately 68% or 0.68 of the measurements fall in the interval $\mu \pm \sigma = 12 \pm 2.3$ or 9.7 to 14.3

b approximately 95% or 0.95 of the measurements fall in the interval $\mu \pm 2\sigma = 12 \pm 4.6$ or 7.4 to 16.6

c approximately 99.7% or 0.997 of the measurements fall in the interval $\mu \pm 3\sigma = 12 \pm 6.9$ or 5.1 to 18.9

Therefore, approximately 0.3% or 0.003 will fall outside this interval.

2.26 **a** The stem and leaf plots are shown below. The second set has a slightly higher location and spread.

Stem-and-Leaf Display: Method 1, Method 2

```
Stem-and-leaf of Method 1  N  = 10    Stem-and-leaf of Method 2  N  = 10
Leaf Unit = 0.00010                   Leaf Unit = 0.00010

 1   10  0
 3   11  00                           1  11  0
 4   12  0                            3  12  00
(4)  13  0000                         5  13  00
 2   14  0                            5  14  0
 1   15  0                            4  15  00
                                      2  16  0
                                      1  17  0
```

b *Method 1*: Calculate $\sum x_i = 0.125$ and $\sum x_i^2 = 0.001583$. Then $\overline{x} = \dfrac{\sum x_i}{n} = 0.0125$ and

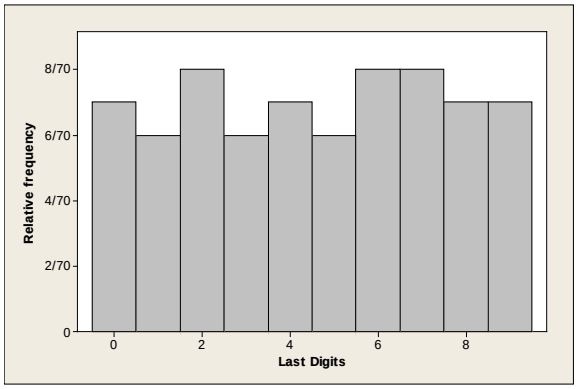
$$s = \sqrt{s^2} = \sqrt{\frac{\sum x_i^2 - \dfrac{\left(\sum x_i\right)^2}{n}}{n-1}} = \sqrt{\frac{0.001583 - \dfrac{(0.125)^2}{10}}{9}} = 0.00151$$

Method 2: Calculate $\sum x_i = 0.138$ and $\sum x_i^2 = 0.001938$. Then $\overline{x} = \dfrac{\sum x_i}{n} = 0.0138$ and

$$s = \sqrt{s^2} = \sqrt{\frac{\sum x_i^2 - \dfrac{\left(\sum x_i\right)^2}{n}}{n-1}} = \sqrt{\frac{0.001938 - \dfrac{(0.138)^2}{10}}{9}} = 0.00193$$

The results confirm the conclusions of part **a**.

2.27 **a** The relative frequency histogram is shown below.



The distribution is relatively "flat" or "uniformly distributed" between 0 and 9. Hence, the center of the distribution should be approximately halfway between 0 and 9 or $(0+9)/2 = 4.5$.

b The range of the data is $R = 9 - 0 = 9$. Using the range approximation, $s \approx R/4 = 9/4 = 2.25$.

c Using the data entry method the students should find $\overline{x} = 4.586$ and $s = 2.892$, which are fairly close to our approximations.

2.28 **a** Similar to previous exercises. The intervals, counts and percentages are shown in the table.

k	$\overline{x} \pm ks$	Interval	Number in Interval	Percentage
1	4.586 ± 2.892	1.694 to 7.478	43	61%
2	4.586 ± 5.784	–1.198 to 10.370	70	100%
3	4.586 ± 8.676	–4.090 to 13.262	70	100%

b The percentages in part **a** do not agree with those given by the Empirical Rule. This is because the shape of the distribution is not mound-shaped, but flat.

2.29 **a** Although most of the animals will die at around 32 days, there may be a few animals that survive a very long time, even with the infection. The distribution will probably be skewed right.

b Using Tchebysheff's Theorem, at least 3/4 of the measurements should be in the interval $\mu \pm \sigma \Rightarrow 32 \pm 72$ or 0 to 104 days.

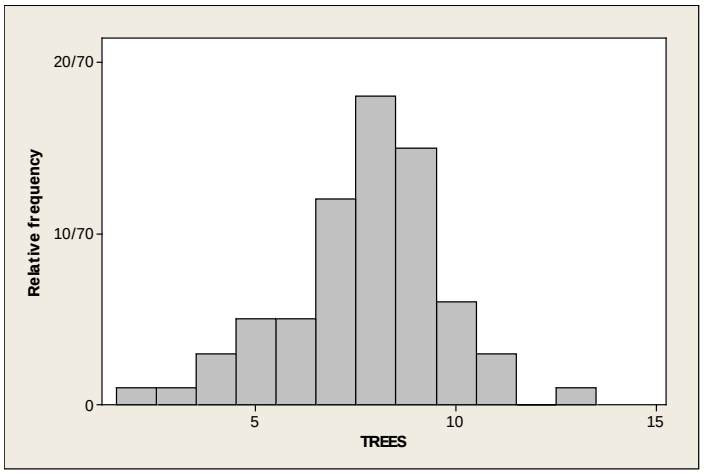
2.30 **a** The value of x is $\mu - \sigma = 32 - 36 = -4$.

b The interval $\mu \pm \sigma$ is 32 ± 36 should contain approximately $(100 - 68) = 34\%$ of the survival times, of which 17% will be longer than 68 days and 17% less than –4 days.

c The latter is clearly impossible. Therefore, the approximate values given by the Empirical Rule are not accurate, indicating that the distribution cannot be mound-shaped.

2.31 **a** We choose to use 12 classes of length 1.0. The tally and the relative frequency histogram follow.

Class i	Class Boundaries	Tally	f_i	Relative frequency, f_i/n
1	2 to < 3	1	1	1/70
2	3 to < 4	1	1	1/70
3	4 to < 5	111	3	3/70
4	5 to < 6	11111	5	5/70
5	6 to < 7	11111	5	5/70
6	7 to < 8	11111 11111 11	12	12/70
7	8 to < 9	11111 11111 11111 111	18	18/70
8	9 to < 10	11111 11111 11111	15	15/70
9	10 to < 11	11111 1	6	6/70
10	11 to < 12	111	3	3/70
11	12 to < 13		0	0
12	13 to < 14	1	1	1/70



b Calculate $n = 70,\ \sum x_i = 541$ and $\sum x_i^2 = 4453$. Then $\overline{x} = \dfrac{\sum x_i}{n} = \dfrac{541}{70} = 7.729$ is an estimate of μ.

c The sample standard deviation is

$$s = \sqrt{\frac{\sum x_i^2 - \dfrac{\left(\sum x_i\right)^2}{n}}{n-1}} = \sqrt{\frac{4453 - \dfrac{(541)^2}{70}}{69}} = \sqrt{3.9398} = 1.985$$

The three intervals, $\overline{x} \pm ks$ for $k = 1, 2, 3$ are calculated below. The table shows the actual percentage of measurements falling in a particular interval as well as the percentage predicted by Tchebysheff's Theorem and the Empirical Rule. Note that the Empirical Rule should be fairly accurate, as indicated by the mound-shape of the histogram in part **a**.

k	$\overline{x} \pm ks$	Interval	Fraction in Interval	Tchebysheff	Empirical Rule
1	7.729 ± 1.985	5.744 to 9.714	50/70 = 0.71	at least 0	≈ 0.68
2	7.729 ± 3.970	3.759 to 11.699	67/70 = 0.96	at least 0.75	≈ 0.95
3	7.729 ± 5.955	1.774 to 13.684	70/70 = 1.00	at least 0.89	≈ 0.997

2.32 **a** Calculate $R = 1.92 - 0.53 = 1.39$ so that $s \approx R/4 = 1.39/4 = 0.3475$.

b Calculate $n = 14,\ \sum x_i = 12.55$ and $\sum x_i^2 = 13.3253$.

Then

$$s^2 = \frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1} = \frac{13.3253 - \frac{(12.55)^2}{14}}{13} = 0.1596 \text{ and } s = \sqrt{0.15962} = 0.3995$$

which is fairly close to the approximate value of *s* from part **a**.

2.33 **a-b** Calculate $R = 93 - 51 = 42$ so that $s \approx R/4 = 42/4 = 10.5$.

c Calculate $n = 30$, $\sum x_i = 2145$ and $\sum x_i^2 = 158,345$. Then

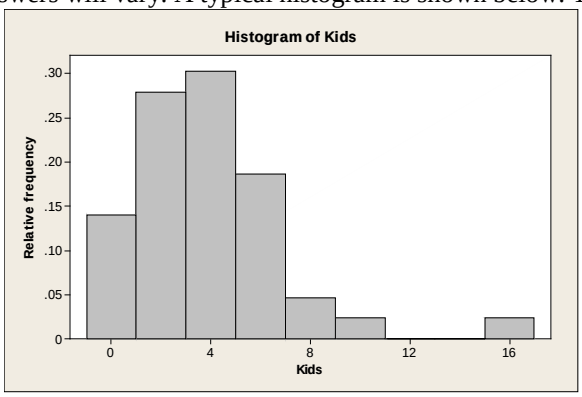
$$s^2 = \frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1} = \frac{158,345 - \frac{(2145)^2}{30}}{29} = 171.6379 \text{ and } s = \sqrt{171.6379} = 13.101$$

which is fairly close to the approximate value of *s* from part **b**.

d The two intervals are calculated below. The proportions agree with Tchebysheff's Theorem, but are not to close to the percentages given by the Empirical Rule. (This is because the distribution is not quite mound-shaped.)

k	$\bar{x} \pm ks$	Interval	Fraction in Interval	Tchebysheff	Empirical Rule
2	71.5 ± 26.20	45.3 to 97.7	30/30 = 1.00	at least 0.75	≈ 0.95
3	71.5 ± 39.30	32.2 to 110.80	30/30 = 1.00	at least 0.89	≈ 0.997

2.34 **a** Answers will vary. A typical histogram is shown below. The distribution is skewed to the right.



b Calculate $n = 43$, $\sum x_i = 153$ and $\sum x_i^2 = 901$. Then

$$\bar{x} = \frac{\sum x_i}{n} = \frac{153}{43} = 3.56,$$

$$s^2 = \frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1} = \frac{901 - \frac{(153)^2}{43}}{42} = 8.490587$$

and $s = \sqrt{8.490587} = 2.91$

c The three intervals, $\bar{x} \pm ks$ for $k = 1, 2, 3$ are calculated below. The table shows the actual percentage of measurements falling in a particular interval as well as the percentage predicted by Tchebysheff's Theorem and the Empirical Rule. Note that the Empirical Rule is not very accurate for the first interval, since the histogram in part **a** is skewed.

k	$\bar{x} \pm ks$	Interval	Fraction in Interval	Tchebysheff	Empirical Rule
1	3.56 ± 2.91	.65 to 6.47	33/43 = .77	at least 0	≈ 0.68
2	3.56 ± 5.82	–2.26 to 9.38	41/43 = .95	at least 0.75	≈ 0.95
3	3.56 ± 8.73	–5.17 to 12.29	42/43 = .977	at least 0.89	≈ 0.997

2.35 **a** Calculate $R = 2.39 - 1.28 = 1.11$ so that $s \approx R/2.5 = 1.11/2.5 = .444$.

b In Exercise 2.17, we calculated $\sum x_i = 8.56$ and $\sum x_i^2 = 1.28^2 + 2.39^2 + \cdots + 1.51^2 = 15.415$. Then

$$s^2 = \frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1} = \frac{15.451 - \frac{(8.56)^2}{5}}{4} = \frac{.76028}{4} = .19007$$

and $s = \sqrt{s^2} = \sqrt{.19007} = .436$, which is very close to our estimate in part **a**.

2.36 **a** Answers will vary. A typical stem and leaf plot is generated by ***Minitab.***

Stem-and-Leaf Display: Completed Passes

```
Stem-and-leaf of Completed Passes  N  = 15
Leaf Unit = 1.0

 1   0  7
 2   1  2
 7   1  58999
(3)  2  112
 5   2  5677
 1   3  4
```

b Calculate $n = 15$, $\sum x_i = 312$ and $\sum x_i^2 = 7106$. Then $\bar{x} = \frac{\sum x_i}{n} = \frac{312}{15} = 20.8$,

$$s^2 = \frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1} = \frac{7106 - \frac{(312)^2}{15}}{14} = 44.0285714$$

and $s = \sqrt{s^2} = \sqrt{44.0285714} = 6.635$.

c Calculate $\bar{x} \pm 2s \Rightarrow 20.8 \pm 13.27$ or 7.53 to 34.07. From the original data set, 14 of the 15 measurements, or about 93% fall in this interval.

2.37 **a** Calculate $n = 15$, $\sum x_i = 21$ and $\sum x_i^2 = 49$. Then $\bar{x} = \frac{\sum x_i}{n} = \frac{21}{15} = 1.4$ and

$$s^2 = \frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1} = \frac{49 - \frac{(21)^2}{15}}{14} = 1.4$$

b Using the frequency table and the grouped formulas, calculate

$$\sum x_i f_i = 0(4) + 1(5) + 2(2) + 3(4) = 21$$

$$\sum x_i^2 f_i = 0^2(4) + 1^2(5) + 2^2(2) + 3^2(4) = 49$$

Then, as in part **a**,

$$\bar{x} = \frac{\sum x_i f_i}{n} = \frac{21}{15} = 1.4$$

$$s^2 = \frac{\sum x_i^2 f_i - \frac{\left(\sum x_i f_i\right)^2}{n}}{n-1} = \frac{49 - \frac{(21)^2}{15}}{14} = 1.4$$

2.38 Use the formulas for grouped data given in Exercise 2.37. Calculate $n = 17$, $\sum x_i f_i = 79$, and $\sum x_i^2 f_i = 393$. Then,

$$\bar{x} = \frac{\sum x_i f_i}{n} = \frac{79}{17} = 4.65$$

$$s^2 = \frac{\sum x_i^2 f_i - \frac{\left(\sum x_i f_i\right)^2}{n}}{n-1} = \frac{393 - \frac{(79)^2}{17}}{16} = 1.6176 \text{ and } s = \sqrt{1.6176} = 1.27$$

2.39 **a** The data in this exercise have been arranged in a frequency table.

x_i	0	1	2	3	4	5	6	7	8	9	10
f_i	10	5	3	2	1	1	1	0	0	1	1

Using the frequency table and the grouped formulas, calculate

$$\sum x_i f_i = 0(10) + 1(5) + \cdots + 10(1) = 51$$

$$\sum x_i^2 f_i = 0^2(10) + 1^2(5) + \cdots + 10^2(1) = 293$$

Then

$$\overline{x} = \frac{\sum x_i f_i}{n} = \frac{51}{25} = 2.04$$

$$s^2 = \frac{\sum x_i^2 f_i - \frac{\left(\sum x_i f_i\right)^2}{n}}{n-1} = \frac{293 - \frac{(51)^2}{25}}{24} = 7.873 \text{ and } s = \sqrt{7.873} = 2.806.$$

b-c The three intervals $\overline{x} \pm ks$ for $k = 1, 2, 3$ are calculated in the table along with the actual proportion of measurements falling in the intervals. Tchebysheff's Theorem is satisfied and the approximation given by the Empirical Rule are fairly close for $k = 2$ and $k = 3$.

k	$\overline{x} \pm ks$	Interval	Fraction in Interval	Tchebysheff	Empirical Rule
1	2.04 ± 2.806	–0.766 to 4.846	21/25 = 0.84	at least 0	≈ 0.68
2	2.04 ± 5.612	–3.572 to 7.652	23/25 = 0.92	at least 0.75	≈ 0.95
3	2.04 ± 8.418	–6.378 to 10.458	25/25 = 1.00	at least 0.89	≈ 0.997

2.40 The ordered data are:

0, 1, 3, 4, 4, 5, 6, 6, 7, 7, 8

a With $n = 12$, the median is in position $0.5(n+1) = 6.5$, or halfway between the 6th and 7th observations. The lower quartile is in position $0.25(n+1) = 3.25$ (one-fourth of the way between the 3rd and 4th observations) and the upper quartile is in position $0.75(n+1) = 9.75$ (three-fourths of the way between the 9th and 10th observations). Hence, $m = (5+6)/2 = 5.5$, $Q_1 = 3 + 0.25(4-3) = 3.25$ and $Q_3 = 6 + 0.75(7-6) = 6.75$. Then the five-number summary is

Min	Q_1	Median	Q_3	Max
0	3.25	5.5	6.75	8

and

$$IQR = Q_3 - Q_1 = 6.75 - 3.25 = 3.50$$

b Calculate $n = 12$, $\sum x_i = 57$ and $\sum x_i^2 = 337$. Then $\overline{x} = \frac{\sum x_i}{n} = \frac{57}{12} = 4.75$ and the sample standard deviation is

$$s = \sqrt{\frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1}} = \sqrt{\frac{337 - \frac{(57)^2}{12}}{11}} = \sqrt{6.022727} = 2.454$$

c For the smaller observation, $x = 0$,

$$z\text{-score} = \frac{x - \overline{x}}{s} = \frac{0 - 4.75}{2.454} = -1.94$$

and for the largest observation, $x = 8$,

$$z\text{-score} = \frac{x - \overline{x}}{s} = \frac{8 - 4.75}{2.454} = 1.32$$

Since neither z-score exceeds 2 in absolute value, none of the observations are unusually small or large.

2.41 The ordered data are:

0, 1, 5, 6, 7, 8, 9, 10, 12, 12, 13, 14, 16, 19, 19

With $n = 15$, the median is in position $0.5(n+1) = 8$, so that $m = 10$. The lower quartile is in position $0.25(n+1) = 4$ so that $Q_1 = 6$ and the upper quartile is in position $0.75(n+1) = 12$ so that $Q_3 = 14$. Then the five-number summary is

Min	Q_1	Median	Q_3	Max
0	6	10	14	19

and $IQR = Q_3 - Q_1 = 14 - 6 = 8$.

2.42 The ordered data are:

1.0, 1.7, 2.0, 2.1, 2.3, 2.8, 2.9, 4.4, 5.1, 6.5, 8.8

a-b With $n = 11$, the lower quartile is in position $0.25(n+1) = 3$ (the 3rd observation) and the upper quartile is in position $0.75(n+1) = 9$ (the 9th observation). Hence, Q_1 = 2.0 and Q_3 = 5.1.

c Then $IQR = Q_3 - Q_1 = 5.1 - 2.0 = 3.1$.

2.43 Notice that the data is already ranked from smallest to largest.

a-b With $n = 8$, the lower quartile is in position $0.25(n+1) = 2.25$ (one-fourth of the way between the 2nd and 3rd observations) and the upper quartile is in position $0.75(n+1) = 6.75$ (three-fourths of the way between the 6th and 7th observations). Hence, $Q_1 = .30 + 0.25(.35 - .30) = .3125$ and $Q_3 = .58 + 0.75(.76 - .58) = .7150$. Then $IQR = Q_3 - Q_1 = .7150 - .3125 = .4025$.

c The upper and lower fences are then calculated as:

Lower fence = $Q_1 - 1.5IQR = .3125 - 1.5(.4025) = -.29125$

Upper fence = $Q_3 + 1.5IQR = .7150 + 1.5(.4025) = 1.31875$

There are no data points that lie outside these fences.

2.44 The ordered data are:

12, 18, 22, 23, 24, 25, 25, 26, 26, 27, 28

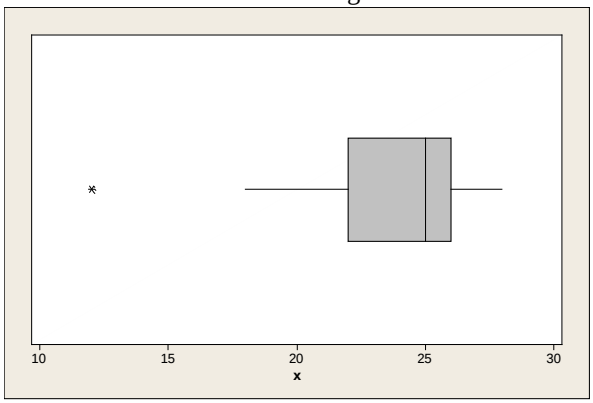
For $n = 11$, the position of the median is $0.5(n+1) = 0.5(11+1) = 6$ and $m = 25$. The positions of the quartiles are $0.25(n+1) = 3$ and $0.75(n+1) = 9$, so that $Q_1 = 22$, $Q_3 = 26$, and $IQR = 26 - 22 = 4$.

The *lower and upper fences* are:

$$Q_1 - 1.5IQR = 22 - 6 = 16$$

$$Q_3 + 1.5IQR = 26 + 6 = 32$$

The only observation falling outside the fences is $x = 12$ which is identified as an outlier. The box plot is shown below. The lower whisker connects the box to the smallest value that is not an outlier, $x = 18$. The upper whisker connects the box to the largest value that is not an outlier or $x = 28$.



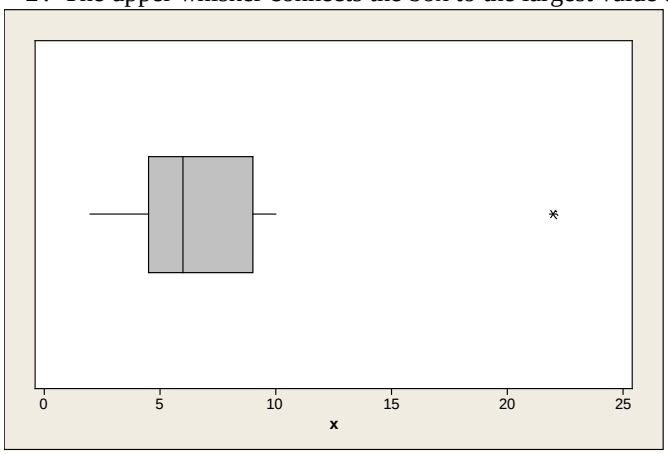
2.45 The ordered data are:

2, 3, 4, 5, 6, 6, 6, 7, 8, 9, 9, 10, 22

For $n = 13$, the position of the median is $0.5(n+1) = 0.5(13+1) = 7$ and $m = 6$. The positions of the quartiles are $0.25(n+1) = 3.5$ and $0.75(n+1) = 10.5$, so that $Q_1 = 4.5$, $Q_3 = 9$, and $IQR = 9 - 4.5 = 4.5$. The *lower and upper fences* are:

$$Q_1 - 1.5IQR = 4.5 - 6.75 = -2.25$$
$$Q_3 + 1.5IQR = 9 + 6.75 = 15.75$$

The value $x = 22$ lies outside the upper fence and is an outlier. The box plot is shown below. The lower whisker connects the box to the smallest value that is not an outlier, which happens to be the minimum value, $x = 2$. The upper whisker connects the box to the largest value that is not an outlier or $x = 10$.



2.46 From Section 2.6, the 69th percentile implies that 69% of all students scored below your score, and only 31% scored higher.

2.47 **a** The ordered data are shown below:

1.70	101.00	209.00	264.00	316.00	445.00
1.72	118.00	218.00	278.00	318.00	481.00
5.90	168.00	221.00	286.00	329.00	485.00
8.80	180.00	241.00	314.00	397.00	
85.40	183.00	252.00	315.00	406.00	

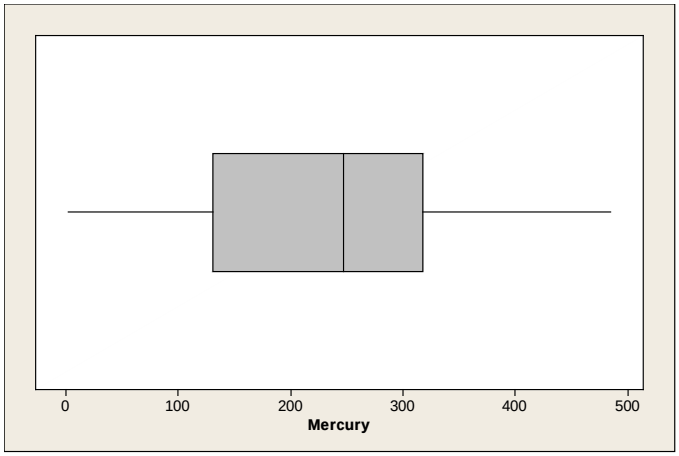
For $n = 28$, the position of the median is $0.5(n+1) = 14.5$ and the positions of the quartiles are $0.25(n+1) = 7.25$ and $0.75(n+1) = 21.75$. The lower quartile is ¼ the way between the 7th and 8th measurements or $Q_1 = 118 + 0.25(168 - 118) = 130.5$ and the upper quartile is ¾ the way between the 21st and 22nd measurements or $Q_3 = 316 + 0.75(318 - 316) = 317.5$. Then the five-number summary is

Min	Q_1	Median	Q_3	Max
1.70	130.5	246.5	317.5	485

b Calculate $IQR = Q_3 - Q_1 = 317.5 - 130.5 = 187$. Then the *lower and upper fences* are:

$$Q_1 - 1.5IQR = 130.5 - 280.5 = -150$$
$$Q_3 + 1.5IQR = 317.5 + 280.5 = 598$$

The box plot is shown below. Since there are no outliers, the whiskers connect the box to the minimum and maximum values in the ordered set.



c-d The boxplot does not identify any of the measurements as outliers, mainly because the large variation in the measurements cause the IQR to be large. However, the student should notice the extreme difference in the magnitude of the first four observations taken on young dolphins. These animals have not been alive long enough to accumulate a large amount of mercury in their bodies.

2.48 **a** See Exercise 2.24b.

b For $x = 1.38$,

$$z\text{-score} = \frac{x - \bar{x}}{s} = \frac{1.38 - 1.05}{0.17} = 1.94$$

while for $x = 1.41$,

$$z\text{-score} = \frac{x - \bar{x}}{s} = \frac{1.41 - 1.05}{0.17} = 2.12$$

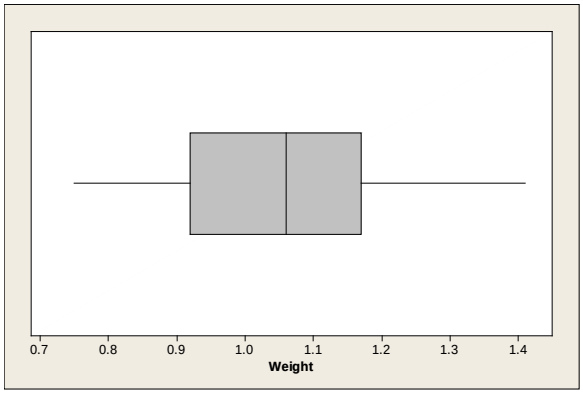
The value $x = 1.41$ would be considered somewhat unusual, since its *z*-score exceeds 2 in absolute value.

c For $n = 27$, the position of the median is $0.5(n+1) = 0.5(27+1) = 14$ and $m = 1.06$. The positions of the quartiles are $0.25(n+1) = 7$ and $0.75(n+1) = 21$, so that $Q_1 = 0.92$, $Q_3 = 1.17$, and $IQR = 1.17 - 0.92 = 0.25$.

The *lower and upper fences* are:

$$Q_1 - 1.5IQR = 0.92 - 0.375 = 0.545$$
$$Q_3 + 1.5IQR = 1.17 + 0.375 = 1.545$$

The box plot is shown below. Since there are no outliers, the whiskers connect the box to the minimum and maximum values in the ordered set.



Since the median line is almost in the center of the box, the whiskers are nearly the same lengths, the data set is relatively symmetric.

2.49 **a** For $n = 15$, the position of the median is $0.5(n+1) = 8$ and the positions of the quartiles are $0.25(n+1) = 4$ and $0.75(n+1) = 12$, while for $n = 16$, the position of the median is $0.5(n+1) = 8.5$ and the positions of the quartiles are $0.25(n+1) = 4.25$ and $0.75(n+1) = 12.75$. The sorted measurements are shown below.

Aaron Rodgers: 7, 12, 15, 18, 19, 19, 19, 21, 21, 22, 25, 26, 27, 27, 34
Drew Brees: 21, 22, 23, 24, 24, 25, 27, 27, 28, 29, 29, 30, 33, 34, 35, 37

For Aaron Rodgers,

$m = 21$, $Q_1 = 18$ and $Q_3 = 26$.

For Drew Brees,

$m = (27+28)/2 = 27.5$, $Q_1 = 24 + 0.25(24-24) = 24$ and $Q_3 = 30 + 0.75(33-30) = 32.25$.

Then the five-number summaries are

	Min	Q_1	Median	Q_3	Max
Rodgers	7	18	21	26	34
Brees	21	24	27.5	32.25	37

b For Aaron Rodgers, calculate $IQR = Q_3 - Q_1 = 26 - 18 = 8$. Then the *lower and upper fences* are:

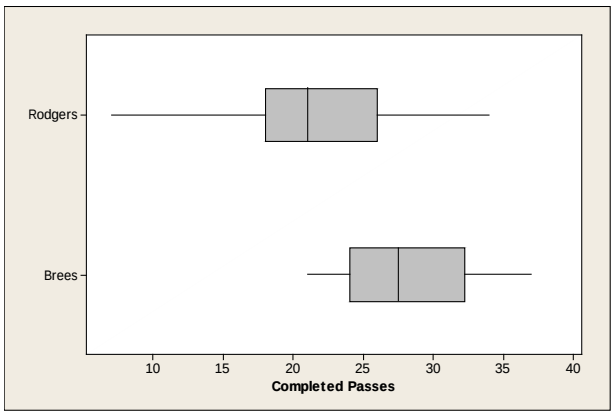
$$Q_1 - 1.5IQR = 18 - 12 = 6$$
$$Q_3 + 1.5IQR = 26 + 12 = 38$$

and there are no outliers.

For Drew Brees, calculate $IQR = Q_3 - Q_1 = 32.25 - 24 = 8.25$. Then the *lower and upper fences* are:

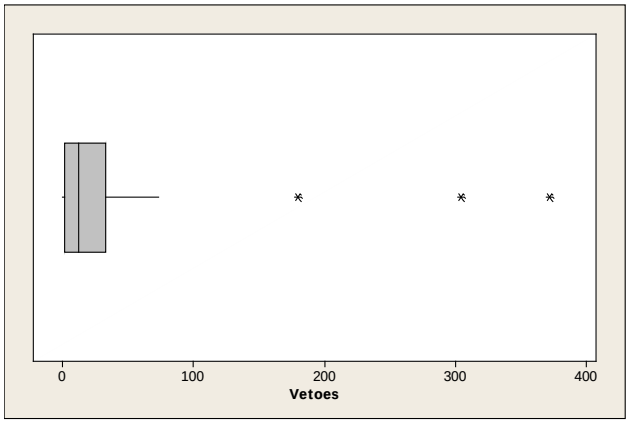
$$Q_1 - 1.5IQR = 24 - 12.375 = 11.625$$
$$Q_3 + 1.5IQR = 32.25 + 12.375 = 44.625$$

and there are no outliers. The box plots are shown below.



c Answers will vary. Both distributions are relatively symmetric and somewhat mound-shaped. The Rodgers distribution is slightly more variable; Brees has a higher median number of completed passes.

2.50 Answers will vary from student to student. The distribution is skewed to the right with three outliers (Truman, Cleveland and F. Roosevelt). The box plot is shown on the next page.



2.51 **a** Just by scanning through the 20 measurements, it seems that there are a few unusually small measurements, which would indicate a distribution that is skewed to the left.

b The position of the median is $0.5(n+1) = 0.5(25+1) = 10.5$ and $m = (120+127)/2 = 123.5$. The mean is

$$\overline{x} = \frac{\sum x_i}{n} = \frac{2163}{20} = 108.15$$

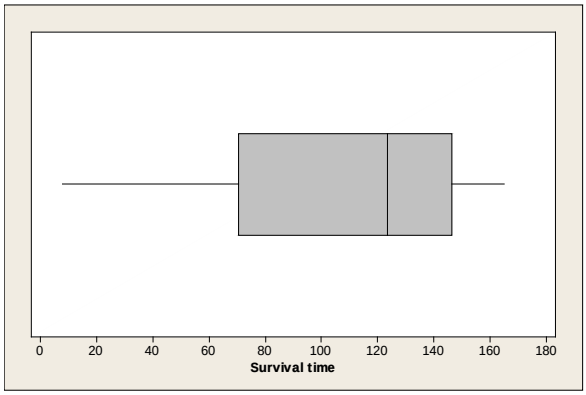
which is smaller than the median, indicate a distribution skewed to the left.

c The positions of the quartiles are $0.25(n+1) = 5.25$ and $0.75(n+1) = 15.75$, so that $Q_1 = 65 - .25(87-65) = 70.5$, $Q_3 = 144 + .75(147-144) = 146.25$, and $IQR = 146.25 - 70.5 = 75.75$.

The *lower and upper fences* are:

$$Q_1 - 1.5IQR = 70.5 - 113.625 = -43.125$$
$$Q_3 + 1.5IQR = 146.25 + 113.625 = 259.875$$

The box plot is shown below. There are no outliers. The long left whisker and the median line located to the right of the center of the box indicates that the distribution that is skewed to the left.



2.52 **a** The sorted data is:

216.85, 230.60, 236.96, 243.74, 271.99, 288.02
288.57, 298.12, 301.79, 311.20, 368.57, 370.23

The positions of the median and the quartiles are $0.5(n+1) = 6.5$, $0.25(n+1) = 3.25$ and $0.75(n+1) = 9.75$,

$$m = (288.02 + 288.57)/2 = 288.295$$

so that $Q_1 = 236.96 + .25(243.74 - 236.96) = 238.655$

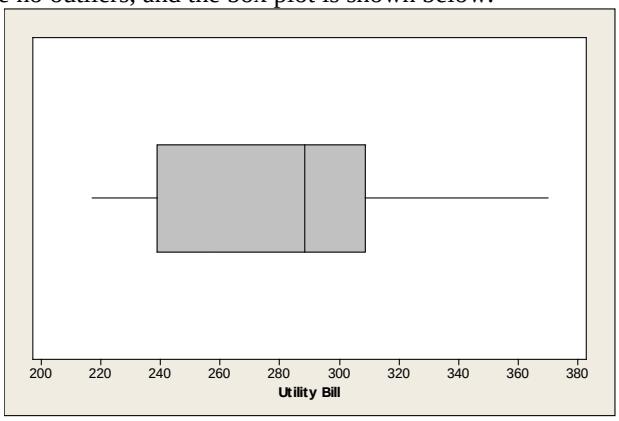
$$Q_3 = 301.79 + .75(311.20 - 301.79) = 308.8475$$

and $IQR = 308.8475 - 238.655 = 70.1925$.

The *lower and upper fences* are:

$$Q_1 - 1.5IQR = 238.655 - 105.28875 = 133.26125$$
$$Q_3 + 1.5IQR = 308.8475 + 105.28875 = 414.13625$$

There are no outliers, and the box plot is shown below.



b Because of the slightly longer right whisker, there are a few unusually large bills (probably in the summer due to air conditioning). A large portion of the bills are in the \$240 to \$285 range.

2.53 Answers will vary. The student should notice the outliers in the female group, and that the median female temperature is higher than the median male temperature.

2.54 **a** Calculate $n = 14,\ \sum x_i = 367$ and $\sum x_i^2 = 9641$. Then $\overline{x} = \dfrac{\sum x_i}{n} = \dfrac{367}{14} = 26.214$ and

$$s = \sqrt{\frac{\sum x_i^2 - \dfrac{\left(\sum x_i\right)^2}{n}}{n-1}} = \sqrt{\frac{9641 - \dfrac{(367)^2}{14}}{13}} = 1.251$$

b Calculate $n = 14,\ \sum x_i = 366$ and $\sum x_i^2 = 9644$. Then $\overline{x} = \dfrac{\sum x_i}{n} = \dfrac{366}{14} = 26.143$ and

$$s = \sqrt{\frac{\sum x_i^2 - \dfrac{\left(\sum x_i\right)^2}{n}}{n-1}} = \sqrt{\frac{9644 - \dfrac{(366)^2}{14}}{13}} = 2.413$$

c The centers are roughly the same; the Sunmaid raisins appear slightly more variable.

2.55 **a** The ordered sets are shown below:

Generic					Sunmaid				
24	25	25	25	26	22	24	24	24	24
26	26	26	26	27	25	25	27	28	28
27	28	28	28		28	28	29	30	

For $n = 14$, the position of the median is $0.5(n+1) = 0.5(14+1) = 7.5$ and the positions of the quartiles are $0.25(n+1) = 3.75$ and $0.75(n+1) = 11.25$, so that

Generic: $m = 26,\ Q_1 = 25,\ Q_3 = 27.25,$ and $IQR = 27.25 - 25 = 2.25$

Sunmaid: $m = 26,\ Q_1 = 24,\ Q_3 = 28,$ and $IQR = 28 - 24 = 4$

b **Generic:** *Lower and upper fences* are:

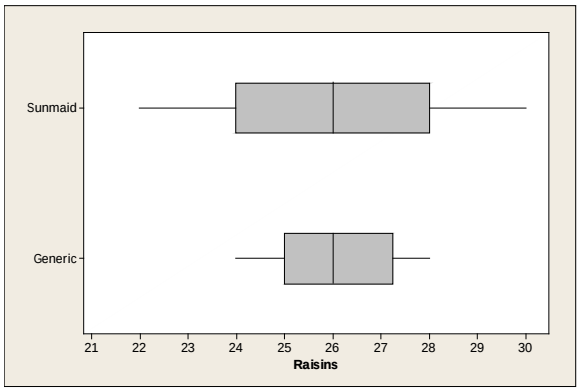
$$Q_1 - 1.5IQR = 25 - 3.375 = 21.625$$
$$Q_3 + 1.5IQR = 27.25 + 3.375 = 30.625$$

Sunmaid: *Lower and upper fences* are:

$$Q_1 - 1.5IQR = 24 - 6 = 18$$

$$Q_3 + 1.5IQR = 28 + 6 = 34$$

The box plots are shown below. There are no outliers.



d If the boxes are not being underfilled, the average size of the raisins is roughly the same for the two brands. However, since the number of raisins is more variable for the Sunmaid brand, it would appear that some of the Sunmaid raisins are large while others are small. The individual sizes of the generic raisins are not as variable.

2.56 **a** Calculate the range as $R = 15 - 1 = 14$. Using the range approximation, $s \approx R/4 = 14/4 = 3.5$.

b Calculate $n = 25$, $\sum x_i = 155.5$ and $\sum x_i^2 = 1260.75$. Then

$$\overline{x} = \frac{\sum x_i}{n} = \frac{155.5}{25} = 6.22 \text{ and}$$

$$s = \sqrt{\frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1}} = \sqrt{\frac{1260.75 - \frac{(155.5)^2}{25}}{24}} = 3.497$$

which is very close to the approximation found in part **a**.

c Calculate $\overline{x} \pm 2s = 6.22 \pm 6.994$ or –0.774 to 13.214. From the original data, 24 measurements or $(24/25)100 = 96\%$ of the measurements fall in this interval. This is close to the percentage given by the Empirical Rule.

2.57 **a** The largest observation found in the data from Exercise 1.26 is 32.3, while the smallest is 0.2. Therefore the range is $R = 32.3 - 0.2 = 32.1$.

b Using the range, the approximate value for *s* is: $s \approx R/4 = 32.1/4 = 8.025$.

c Calculate $n = 50$, $\sum x_i = 418.4$ and $\sum x_i^2 = 6384.34$. Then

$$s = \sqrt{\frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1}} = \sqrt{\frac{6384.34 - \frac{(418.4)^2}{50}}{49}} = 7.671$$

2.58 **a** Refer to Exercise 2.57. Since $\sum x_i = 418.4$, the sample mean is

$$\overline{x} = \frac{\sum x_i}{n} = \frac{418.4}{50} = 8.368$$

The three intervals of interest is shown in the table on the next page, along with the number of observations which fall in each interval.

k	$\bar{x} \pm ks$	Interval	Number in Interval	Percentage
1	8.368 ± 7.671	0.697 to 16.039	37	74%
2	8.368 ± 15.342	–6.974 to 23.710	47	94%
3	8.368 ± 23.013	–14.645 to 31.381	49	98%

b The percentages falling in the intervals do agree with Tchebysheff's Theorem. At least 0 fall in the first interval, at least $3/4 = 0.75$ fall in the second interval, and at least $8/9 = 0.89$ fall in the third. The percentages are not too close to the percentages described by the Empirical Rule (68%, 95%, and 99.7%).

c The Empirical Rule may be unsuitable for describing these data. The data distribution does not have a strong mound-shape (see the relative frequency histogram in the solution to Exercise 1.26), but is skewed to the right.

2.59 The ordered data are shown below.

0.2	2.0	4.3	8.2	14.7
0.2	2.1	4.4	8.3	16.7
0.3	2.4	5.6	8.7	18.0
0.4	2.4	5.8	9.0	18.0
1.0	2.7	6.1	9.6	18.4
1.2	3.3	6.6	9.9	19.2
1.3	3.5	6.9	11.4	23.1
1.4	3.7	7.4	12.6	24.0
1.6	3.9	7.4	13.5	26.7
1.6	4.1	8.2	14.1	32.3

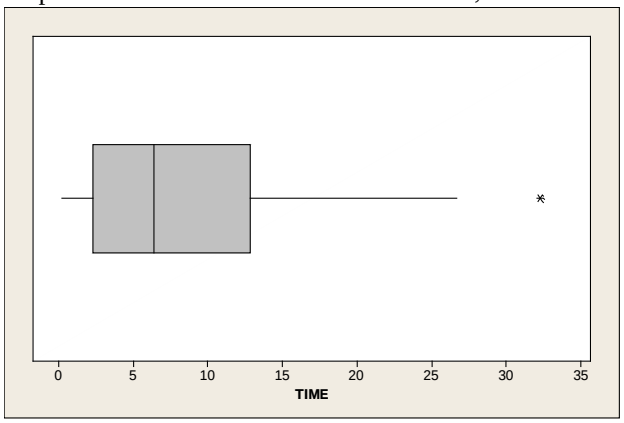
Since $n = 50$, the position of the median is $0.5(n+1) = 25.5$ and the positions of the lower and upper quartiles are $0.25(n+1) = 12.75$ and $0.75(n+1) = 38.25$.

Then $m = (6.1 + 6.6)/2 = 6.35$, $Q_1 = 2.1 + 0.75(2.4 - 2.1) = 2.325$ and $Q_3 = 12.6 + 0.25(13.5 - 12.6) = 12.825$. Then $IQR = 12.825 - 2.325 = 10.5$.

The *lower and upper fences* are:

$$Q_1 - 1.5IQR = 2.325 - 15.75 = -13.425$$
$$Q_3 + 1.5IQR = 12.825 + 15.75 = 28.575$$

and the box plot is shown below. There is one outlier, $x = 32.3$. The distribution is skewed to the right.



2.60 **a** For $n = 14$, the position of the median is $0.5(n+1) = 7.5$ and the positions of the quartiles are $0.25(n+1) = 3.75$ and $0.75(n+1) = 11.25$. The lower quartile is ¾ the way between the 3rd and 4th measurements or $Q_1 = 0.60 + 0.75(0.63 - 0.60) = 0.6225$ and the upper quartile is ¼ the way between the 11th and 12th measurements or $Q_3 = 1.12 + 0.25(1.23 - 1.12) = 1.1475$.

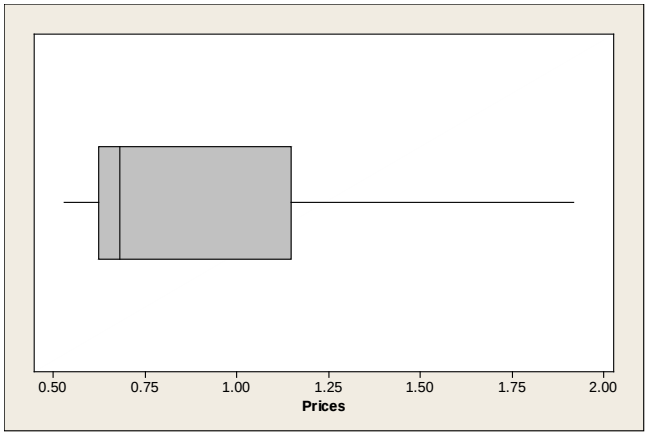
Then the five-number summary is

Min	Q_1	Median	Q_3	Max
0.53	0.6225	0.68	1.1475	1.92

b Calculate $IQR = Q_3 - Q_1 = 1.1475 - 0.6225 = 0.5250$. Then the *lower and upper fences* are:

$$Q_1 - 1.5IQR = 0.6225 - 0.7875 = -0.165$$
$$Q_3 + 1.5IQR = 1.1475 + 0.7875 = 1.935$$

The box plot is shown below. Since there are no outliers, the whiskers connect the box to the minimum and maximum values in the ordered set.



c Calculate $n = 14,\ \sum x_i = 12.55, \sum x_i^2 = 13.3253$. Then

$$\bar{x} = \frac{\sum x_i}{n} = \frac{12.55}{14} = 0.896 \text{ and}$$
$$s = \sqrt{\frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1}} = \sqrt{\frac{13.3253 - \frac{(12.55)^2}{14}}{13}} = 0.3995$$

The z-score for $x = 1.92$ is

$$z = \frac{x - \bar{x}}{s} = \frac{1.92 - 0.896}{0.3995} = 2.56$$

which is somewhat unlikely. This observation does not appear as an outlier in the box plot.

2.61 First calculate the intervals:

$\bar{x} \pm s = 0.17 \pm 0.01$ or 0.16 to 0.18
$\bar{x} \pm 2s = 0.17 \pm 0.02$ or 0.15 to 0.19
$\bar{x} \pm 3s = 0.17 \pm 0.03$ or 0.14 to 0.20

a If no prior information as to the shape of the distribution is available, we use Tchebysheff's Theorem. We would expect at least $\left(1 - 1/1^2\right) = 0$ of the measurements to fall in the interval 0.16 to 0.18; at least $\left(1 - 1/2^2\right) = 3/4$ of the measurements to fall in the interval 0.15 to 0.19; at least $\left(1 - 1/3^2\right) = 8/9$ of the measurements to fall in the interval 0.14 to 0.20.

b According to the Empirical Rule, approximately 68% of the measurements will fall in the interval 0.16 to 0.18; approximately 95% of the measurements will fall between 0.15 to 0.19; approximately 99.7% of the measurements will fall between 0.14 and 0.20. Since mound-shaped distributions are so frequent, if we do have a sample size of 30 or greater, we expect the sample distribution to be mound-shaped. Therefore, in this exercise, we would expect the Empirical Rule to be suitable for describing the set of data.