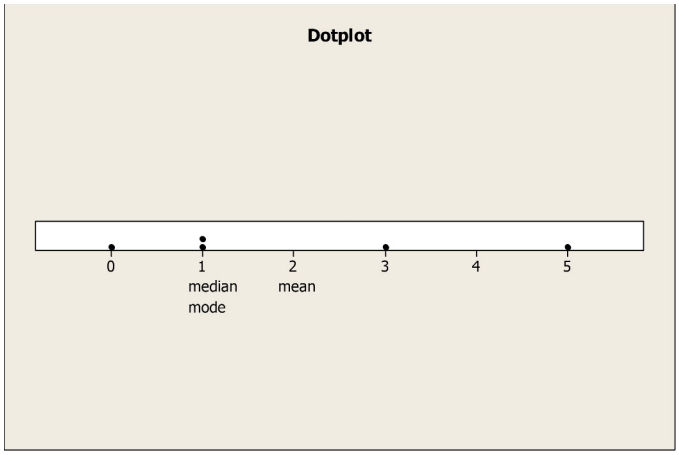
Chapter 2: Describing Data with Numerical Measures

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 centre of the data appears to be around 1.



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

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

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 third ranked measurement, or $m = 1$. The mode is the measurement occurring most frequently, or mode = 1.

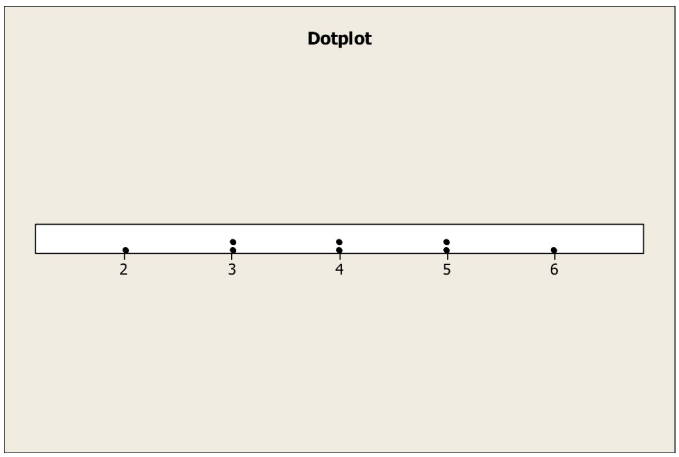
c The three measures in part **b** are located on the dotplot. Since the median and mode are to the left of the mean, we conclude that the measurements are skewed to the right.

2.2 **a** The mean is

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

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 fourth and fifth 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** $$\bar{x} = \frac{\sum x_i}{n} = \frac{58}{10} = 5.8$$

b The ranked observations are 2, 3, 4, 5, 5, 6, 6, 8, 9, 10. Since $n = 10,$ the median is halfway between the fifth and sixth 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** $$\bar{x} = \frac{\sum x_i}{n} = \frac{10823}{10} = 1082.3$$

b $$\bar{x} = \frac{\sum x_i}{n} = \frac{11025}{10} = 1102.5$$

c The average premium cost in different provinces 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 DVD player, the majority should own either 0 or 1. The distribution should be slightly skewed to the right.

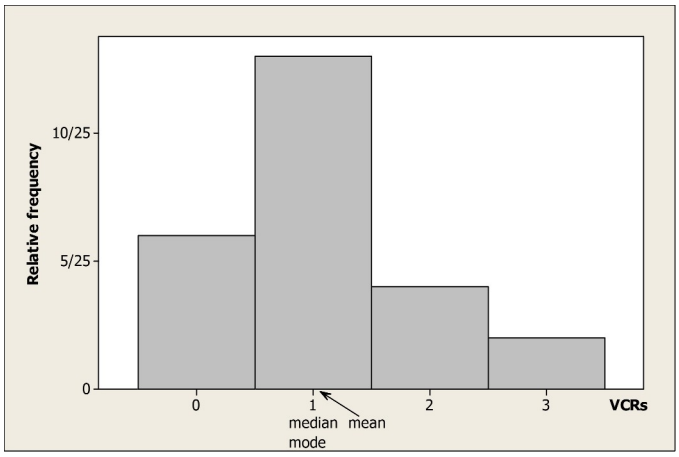
b Since most households will have only one DVD player, we guess that the mode is 1.

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 thirteenth 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 Plot: Wealth

```
Stem and leaf of Wealth  N  = 20
Leaf Unit = 1.0

 5   1  78888
(8)  2  00001234
 7   2  66
 5   3  23
 3   3
 3   4
 3   4  9
 2   5  2
 1   5  6
```

b The mean is

$$\bar{x} = \frac{\sum x_i}{n} = \frac{536.4}{20} = 26.82$$

To calculate the median, notice that the observations are already ranked from smallest to largest. Then since $n = 20$, the position of the median is $0.5(n + 1) = 10.5$, the average of the tenth and eleventh ranked measurements or $m = (21.5 + 22)/2 = 21.75$.

c Since the mean is strongly affected by outliers, the median would be a better measure of centre 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** This is 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 centre than the mean.

2.10 **a** This is 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 fifth and sixth 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** This is similar to previous exercises.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{417}{18} = 23.17$$

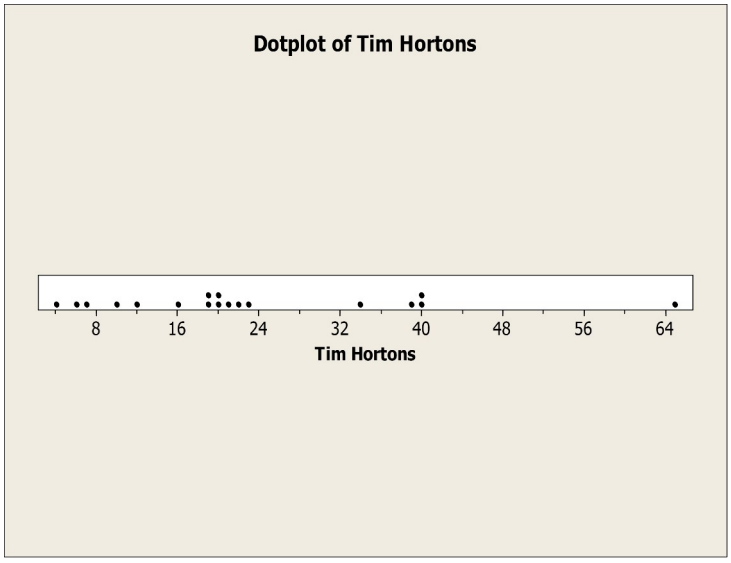
The ranked observations are shown below:

4	6	7	10	12	16	19	19	20
20	21	22	23	34	39	40	40	65

The median is the average of the 9th and 10th observations or $m = (20 + 20)/2 = 20$ and the mode is the most frequently occurring observation—mode = 19, 20, 40.

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

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



2.12 **a** $\bar{x} = \dfrac{\sum x_i}{n} = \dfrac{8120}{10} = 812$

b The ranked data are 1200, 1200, 1050, 800, 750, 700, 670, 650,600, 500 and the median is the average of the fifth and sixth observations or

$$m = \frac{700 + 750}{2} = 725$$

c Average cost would not be as important as many other variables, such as picture quality, sound quality, size, lowest cost for the best quality, and many other considerations.

2.13 **a** The sample mean is

$$\bar{x} = \frac{8 + 6 + 3 + \ldots + 8 + 10}{17} = 6.412$$

b We arrange the data in increasing order: {0,1,2,2,3,4,5,5,6,6,7,8,8,8,10,11,23}. The median is the number in the $0.5(n + 1) = 9$th position, which is 6. The mode is the number that occurs most often, which in this case is 8.

c Since the mean and median are only slightly off from each other, there is only slight skewness.

d The use of the median is probably better than the mean in this case, as the point with value 23 is likely an outlier (and the median is less sensitive to outliers). The mode is rarely employed in practice.

2.14 **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 **b** and **c** are identical.

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

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

2.16 **a** The range is $R = 6 - 1 = 5.$

b $\bar{x} = \frac{\sum x_i}{n} = \frac{31}{8} = 3.875$

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

$$s^2 = \frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1} = \frac{137 - \frac{(31)^2}{8}}{7} = \frac{16.875}{7} = 2.4107 \quad \text{and} \quad s = \sqrt{s^2} = \sqrt{2.4107} = 1.55$$

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

$$s^2 = \frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1} = \frac{15.451 - \frac{(8.56)^2}{5}}{4} = \frac{0.76028}{4} = 0.19007 \quad \text{and} \quad s = \sqrt{s^2} = \sqrt{0.19007} = 0.436$$

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

2.18 **a** The range is $R = 312.40 - 165.12 = 147.28.$

b $\bar{x} = \frac{\sum x_i}{n} = \frac{2726.60}{12} = 227.217$

c Calculate $\sum x_i^2 = 204.94^2 + 180.00^2 + \cdots + 222.23^2 = 647{,}847.084.$ Then

$$s^2 = \frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1} = \frac{647{,}847.084 - \frac{(2{,}726.60)^2}{12}}{11} = 2{,}574.37457$$

and $s = \sqrt{s^2} = \sqrt{2574.37457} = 50.738$

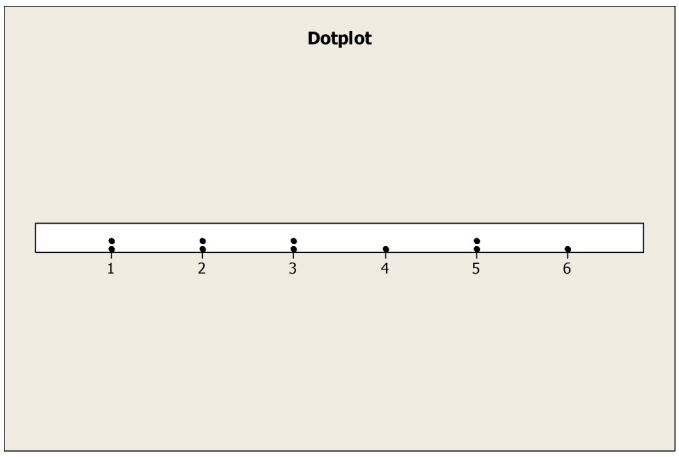
2.19 **a** The range of the data is $R = 6 - 1 = 5$ and the range approximation with $n = 10$ is $s \approx \frac{R}{3} = 1.67.$

b The standard deviation of the sample is

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

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

c–e From the dotplot below, 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:

$\bar{x} \pm s = 36 \pm 3$ or 33 to 39

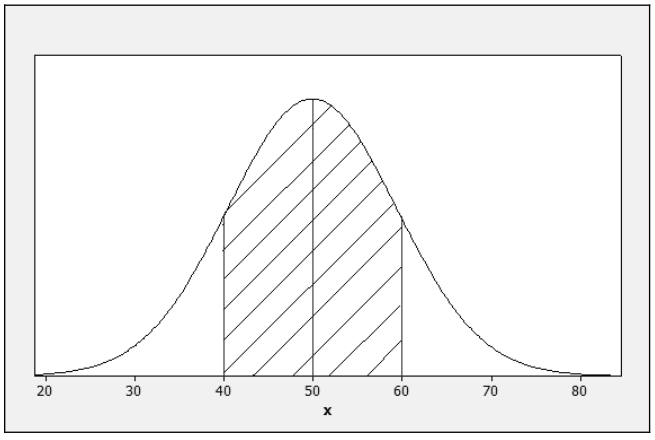
$\bar{x} \pm 2s = 36 \pm 6$ or 30 to 42

$\bar{x} \pm 3s = 36 \pm 9$ or 27 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.

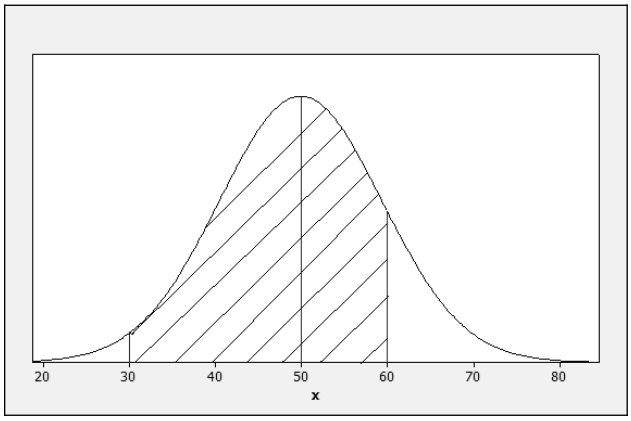
b If no prior information as to the shape of the distribution is available, we use Tchebysheff's Theorem. We would expect at least $\left(1 - 1/1^2\right) = 0$ of the measurements to fall in the interval 33 to 39; at least $\left(1 - 1/2^2\right) = 3/4$ of the measurements to fall in the interval 30 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% 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% 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 are 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 $\overline{x} = \frac{\sum x_i}{n} = \frac{7.6}{10} = 0.76$ 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 generated by *MINITAB* shows that the data is roughly mound-shaped. Note, however, the gap in the centre of the distribution and the two measurements in the upper tail.

Stem and Leaf Plot: 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

$$\bar{x} = \frac{\sum x_i}{n} = \frac{28.41}{27} = 1.052$$

and the standard deviation of the sample is

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

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

k	$\bar{x} \pm ks$	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 1 standard deviation range. This is caused by the lack of mounding (indicated by the gap) in the centre of the distribution.

e The lack of any 1-kilogram packages is probably a marketing technique intentionally used by the supermarket. People who buy slightly less than 1-kilogram would be drawn by the slightly lower price, while those who need exactly 1-kilogram 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 Plot: Method 1, Method 2

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

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

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

1   11  0
3   12  00
5   13  00
5   14  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 $\bar{x} = \dfrac{\sum x_i}{n} = 0.0125$ 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 $\bar{x} = \dfrac{\sum x_i}{n} = 0.0138$ 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 centre 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 $\bar{x} = 4.586$ 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	$\bar{x} \pm ks$	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.

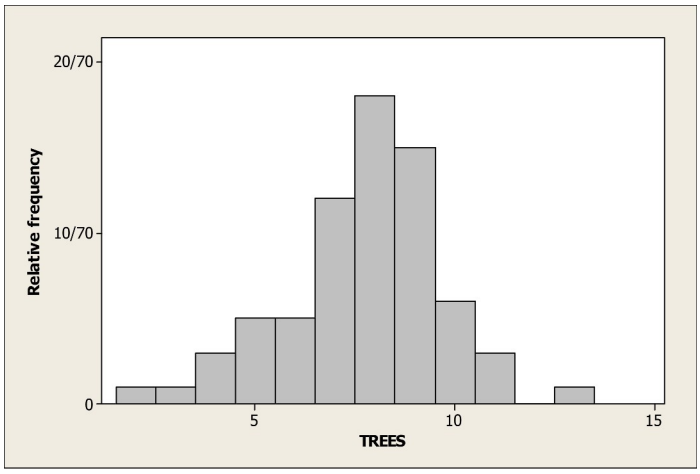
2.30 **a** The value of *x* is $\mu - \sigma = 32 - 36 = -4.$

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–d 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 $\bar{x} = \dfrac{\sum x_i}{n} = \dfrac{541}{70} = 7.729$ 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, $\bar{x} \pm ks$ 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 should be fairly accurate, as indicated by the mound-shape of the histogram in part **a**.

k	$\bar{x} \pm ks$	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

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

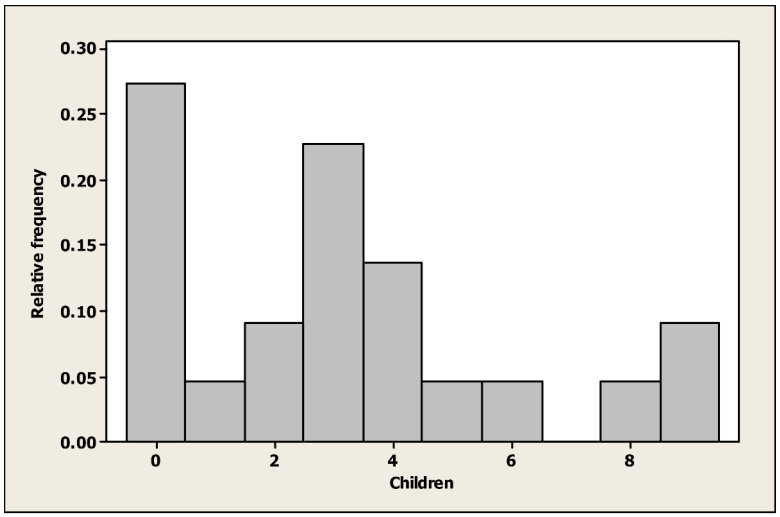
$$s^2 = \frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1} = \frac{158{,}345 - \frac{(2145)^2}{30}}{29} = 171.6379$$ and $s = \sqrt{171.6379} = 13.101$

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 $n = 22,\ \sum x_i = 69,$ and $\sum x_i^2 = 389.$ Then

$$\bar{x} = \frac{\sum x_i}{n} = \frac{69}{22} = 3.136,$$

$$s^2 = \frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1} = \frac{389 - \frac{(69)^2}{22}}{21} = 8.219$$

and $s = \sqrt{8.219} = 2.867$

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	3.136 ± 2.867	.269 to 6.003	13/22 = 0.59	at least 0	≈ 0.68
2	3.136 ± 5.734	–2.598 to 8.870	20/22 = 0.91	at least 0.75	≈ 0.95
3	3.136 ± 8.601	–5.465 to 11.737	22/22 = 1.00	at least 0.89	≈ 0.997

2.35 **a** Answers will vary. A typical stem and leaf plot is generated by *MINITAB***.**

Stem and Leaf Plot: Goals

```
Stem and leaf of Goals  N  = 21
Leaf Unit = 1.0

 1   0  9
 3   1  16
 6   2  335
 8   3  18
(4)  4  0016
 9   5  1245
 5   6  2
 4   7  13
 2   8  7
 1   9  2
```

b Calculate $n = 21,\ \sum x_i = 940,$ and $\sum x_i^2 = 53{,}036.$ Then $\bar{x} = \dfrac{\sum x_i}{n} = \dfrac{940}{21} = 44.76,$

$$s^2 = \frac{\sum x_i^2 - \dfrac{\left(\sum x_i\right)^2}{n}}{n-1} = \frac{53036 - \dfrac{(940)^2}{21}}{20} = 547.99 \quad \text{and} \quad s = \sqrt{s^2} = \sqrt{547.99} = 23.41$$

c Calculate $\bar{x} \pm 2s \Rightarrow 44.76 \pm 46.82$ or -2.06 to $91.58.$ From the original data set, 20 of the measurements, or 95.24% fall in this interval.

2.36 **a** The sample standard deviation is

$$s = \sqrt{\frac{\sum x_i^2 - \dfrac{\left(\sum x_i\right)^2}{n}}{n-1}} \Rightarrow s = \sqrt{\frac{128{,}122.7948 - \dfrac{(1{,}071.091)^2}{9}}{8}} = 9.0287$$

b The range is the largest minus the smallest value: 128.923 – 101.108 = 27.815. The range is about three times as big as the standard deviation. Both the range and the standard deviation measure the amount of "spread" in the data.

c The approximation for *s* based on *R* (as given in Section 2.5 of the text) is $s \approx \dfrac{R}{4} = \dfrac{27.815}{4} = 6.954$ in this case. The approximation is not very accurate.

d After subtracting 0.5, the sample variance is re-computed and turns out to be

$$s^2 = \frac{\sum x_i^2 - \dfrac{\left(\sum x_i\right)^2}{n}}{n-1} \Rightarrow s^2 = \frac{127{,}053.9538 - \dfrac{(1{,}066.591)^2}{9}}{8} = 81.51698$$

Sample variance from part **a** is 81.51698 (by squaring ***s*** computed in part **a**). Thus, subtracting (or adding) 0.5 (or any constant) does not affect the sample variance.

e For this example, three standard deviations from the mean is 119.0101 ± 3(9.0287) = (91.92, 146.10). As can be seen, 100% of the data lies in this interval, which is consistent with Tchebysheff's Theorem (which predicts that at least $1-\left(\frac{1}{3^2}\right)=88.89\%$ of the data/points are within 3 standard deviations from the position of the mean).

f For one standard deviation from the mean, the interval is 119.0101 ± 1(9.0287) = (109.98, 128.04), which contains roughly 6 out of the 9 points, or 66.67%. For 2 standard deviations from the mean, the interval is 119.0101 ± 2(9.0287) = (100.95, 137.07), which contains 9 out of the 9 points, or 100%. The Empirical Rule states that roughly 68% of the data is within 1 standard deviation, and 95% is within 2 standard deviations of the mean. Our results are fairly close (for such a small data set), indicating that our data might be mound-shaped.

g Cannot tell without more information.

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,

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

$$s^2=\frac{\sum x_i^2 f_i-\frac{\left(\sum x_i f_i\right)^2}{n}}{n-1}=\frac{393-\frac{(79)^2}{17}}{16}=1.6176$$ 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	0	1	0	0	1	1

Using the frequency table and the grouped formulas, calculate

$$\sum x_i f_i=0(10)+1(5)+\cdots+10(1)=46$$

$$\sum x_i^2 f_i=0^2(10)+1^2(5)+\cdots+10^2(1)=268$$

Then

$$\bar{x}=\frac{\sum x_i f_i}{n}=\frac{46}{24}=1.917$$

$$s^2 = \frac{\sum x_i^2 f_i - \frac{\left(\sum x_i f_i\right)^2}{n}}{n-1} = \frac{268 - \frac{(46)^2}{24}}{23} = 7.819 \quad \text{and} \quad s = \sqrt{7.819} = 2.796$$

b–c The three intervals $\bar{x} \pm ks$ 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	$\bar{x} \pm ks$	Interval	Fraction in Interval	Tchebysheff	Empirical Rule
1	1.917 ± 2.796	–0.879 to 4.713	21/24 = 0.875	at least 0	≈ 0.68
2	1.917 ± 5.592	–3.675 to 7.509	22/24 = 0.917	at least 0.75	≈ 0.95
3	1.917 ± 8.388	–6.471 to 10.305	24/24 = 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 sixth and seventh observations. The lower quartile is in position $0.25(n+1) = 3.25$ (one-fourth of the way between the third and fourth observations) and the upper quartile is in position $0.75(n+1) = 9.75$ (three-fourths of the way between the ninth and tenth 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 $\bar{x} = \frac{\sum x_i}{n} = \frac{57}{12} = 4.75$ 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,

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

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

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

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

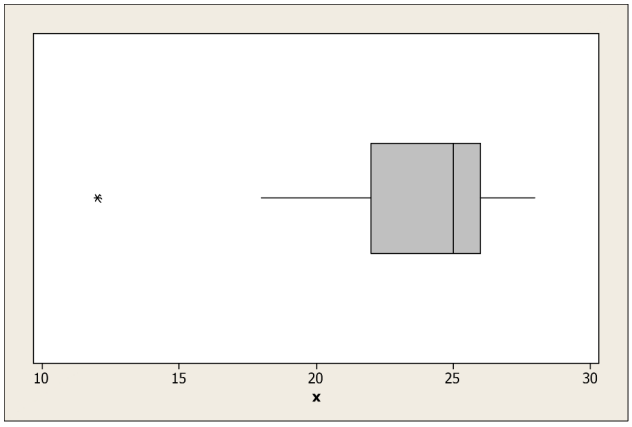
For $n = 11$, the position of the median is $0.5(n+1) = 0.5(11+1) = 6$ and $m = 25$. The positions of the quartiles are $0.25(n+1) = 3$ and $0.75(n+1) = 9,$ so that $Q_1 = 22,\ Q_3 = 26,$ and $IQR = 26 - 22 = 4.$

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 $x = 28$.



2.43 The ordered data are 2, 3, 4, 5, 6, 6, 6, 7, 8, 9, 9, 10, 22.

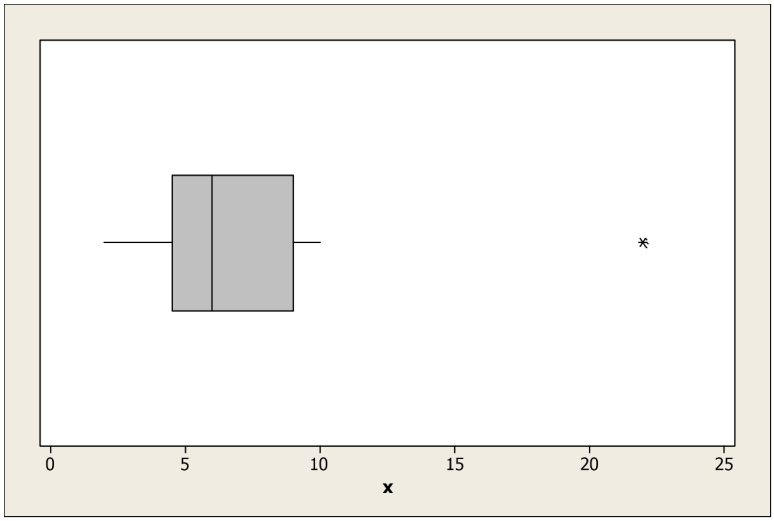
For $n = 13$, the position of the median is $0.5(n+1) = 0.5(13+1) = 7$ and $m = 6$. The positions of the quartiles are $0.25(n+1) = 3.5$ and $0.75(n+1) = 10.5,$ so that $Q_1 = 4.5,\ Q_3 = 9,$ and $IQR = 9 - 4.5 = 4.5.$

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 $x = 10$.



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

2.45 **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 one-fourth the way between the seventh and eighth measurements or $Q_1 = 118 + 0.25(168 - 118) = 130.5$ and the upper quartile is three-fourths the way between the 21st and 22nd measurements or $Q_3 = 316 + 0.75(318 - 316) = 317.5.$ Then the five-number summary is

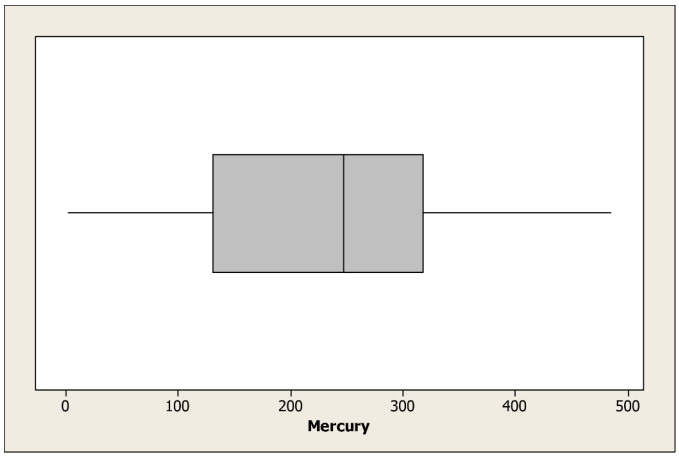
Min	Q_1	Median	Q_3	Max
1.70	130.5	246.5	317.5	485

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 box plot does not identify any of the measurements as outliers, mainly because the large variation in the measurements cause the IQR to be large. However, students 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.

2.46 **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.

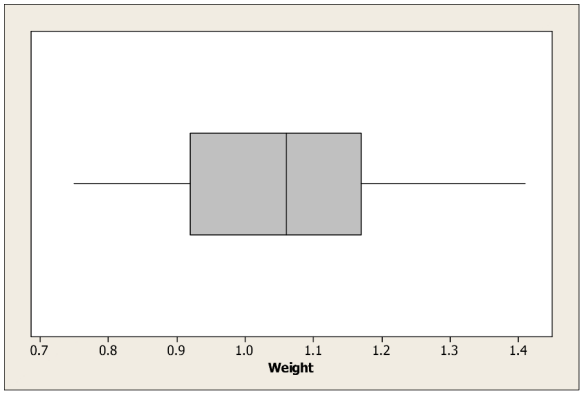
c For $n = 27$, the position of the median is $0.5(n+1) = 0.5(27+1) = 14$ and $m = 1.06$. The positions of the quartiles are $0.25(n+1) = 7$ and $0.75(n+1) = 21,$ so that $Q_1 = 0.92,\ Q_3 = 1.17,$ and $IQR = 1.17 - 0.92 = 0.25.$

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 centre of the box, the whiskers are nearly the same lengths, and the data set is relatively symmetric.

2.47 **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 $0.75(n+1) = 12.$ The sorted measurements are shown below.

Lemieux: 1, 6, 7, 17, 19, 28, 35, 44, 45, 50, 54, 69, 69, 70, 85
Hull: 0, 1, 25, 29, 30, 32, 37, 39, 41, 42, 54, 57, 70, 72, 86

For Mario Lemieux, $m = 44,\ Q_1 = 17,\ Q_3 = 69$

For Brett Hull, $m = 39,\ Q_1 = 29,\ Q_3 = 57$

Then the five-number summaries are

	Min	Q_1	**Median**	Q_3	**Max**
Lemieux	1	17	44	69	85
Hull	0	29	39	57	86

b For Mario Lemieux, calculate $IQR = Q_3 - Q_1 = 69 - 17 = 52.$ Then the *lower and upper fences* are:

$Q_1 - 1.5IQR = 17 - 78 = -61$

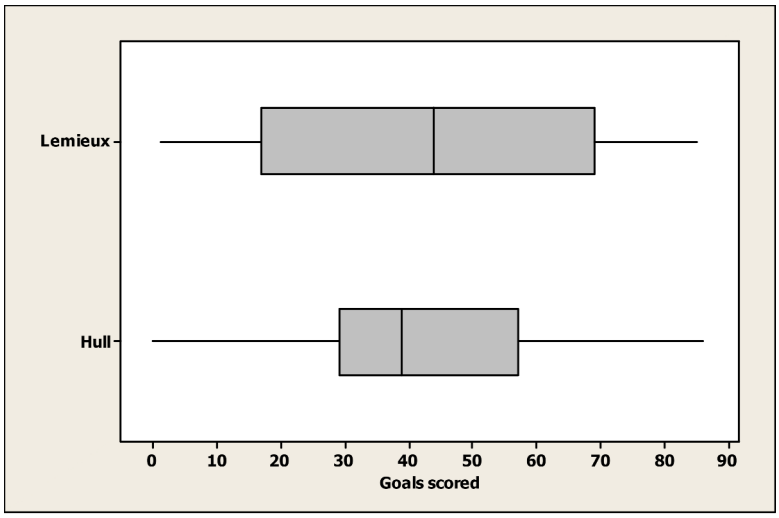
$Q_3 + 1.5IQR = 69 + 78 = 147$

For Brett Hull, $IQR = Q_3 - Q_1 = 57 - 29 = 28.$ Then the *lower and upper fences* are:

$Q_1 - 1.5IQR = 29 - 42 = -13$

$Q_3 + 1.5IQR = 57 + 42 = 99$

There are no outliers, and the box plots are shown below.



c Answers will vary. The Lemieux distribution is roughly symmetric, while the Hull distribution seems little skewed. The Lemieux distribution is slightly more variable; it has a higher *IQR* and a higher median number of goals scored.

2.48 The distribution is fairly symmetric with two outliers (24th and 33rd general elections).

2.49 **a** Just by scanning through the 25 measurements, it seems that there are a few unusually large measurements, which would indicate a distribution that is skewed to the right.

b The position of the median is $0.5(n+1) = 0.5(25+1) = 13$ and $m = 24.4$. The mean is

$$\bar{x} = \frac{\sum x_i}{n} = \frac{960}{25} = 38.4$$

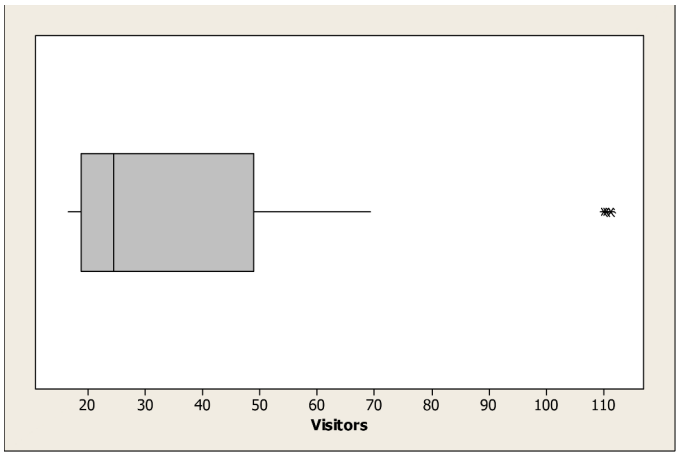
which is larger than the median, indicating a distribution skewed to the right.

c The positions of the quartiles are $0.25(n+1) = 6.5$ and $0.75(n+1) = 19.5,$ so that $Q_1 = 18.7,\ Q_3 = 48.9,$ and $IQR = 48.9 - 18.7 = 30.2.$ The *lower and upper fences* are:

$Q_1 - 1.5IQR = 18.7 - 45.3 = -26.6$

$Q_3 + 1.5IQR = 48.9 + 45.3 = 94.2$

The box plot is shown on the next page. There are three outliers in the upper tail of the distribution, so the upper whisker is connected to the point $x = 69.2$. The long right whisker and the median line located to the left of the centre of the box indicates that the distribution that is skewed to the right.



2.50 **a** The sorted data is 165.12, 176.43, 178.23, 180.00, 204.94, 222.23, 225.47, 236.72, 238.66, 276.70, 309.70, 312.40.

The positions of the median and the quartiles are

$0.5(n+1) = 6.5,\ 0.25(n+1) = 3.25$ and $0.75(n+1) = 9.75,$

$m = (222.23 + 225.47)/2 = 223.85$

$Q_1 = 178.23 + .4425 = 178.67$

so that $Q_3 = 238.66 + 28.53 = 267.19$

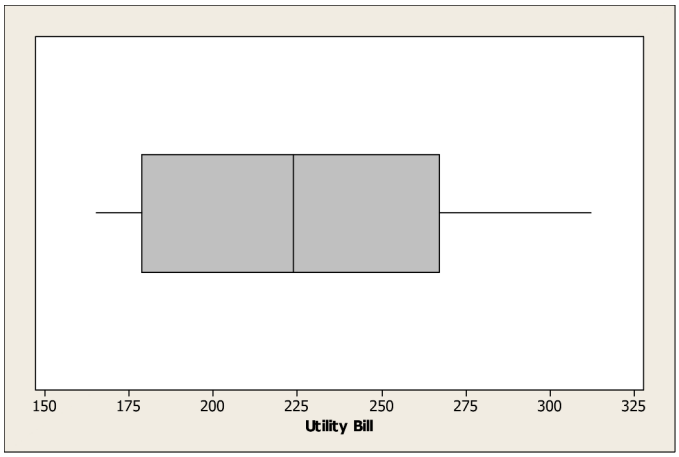
and $IQR = 267.19 - 178.67 = 88.52$

The *lower and upper fences* are:

$Q_1 - 1.5IQR = 178.67 - 90.02 = 88.65$

$Q_3 + 1.5IQR = 267.19 + 90.02 = 357.21$

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



b Because of the long right whisker, the distribution is slightly skewed to the right.

2.51 Answers will vary. Students should notice the outliers in the female group, that the median female temperature is higher than the median male temperature.

2.52 **a** The text proposes the following way to find the first quartile: after arranging the data in increasing order, take "the value of x in position $0.25(n+1)$ … When $0.25(n+1)$ is not an integer, the quartile is found by interpolation, using the values in the two adjacent positions." For this question, the location

is at $0.25(n + 1) = 0.25(14) = 3.5$. The average of the third and fourth ordered points is (110.719 + 115.678)/2 = 113.20 = Q_1.

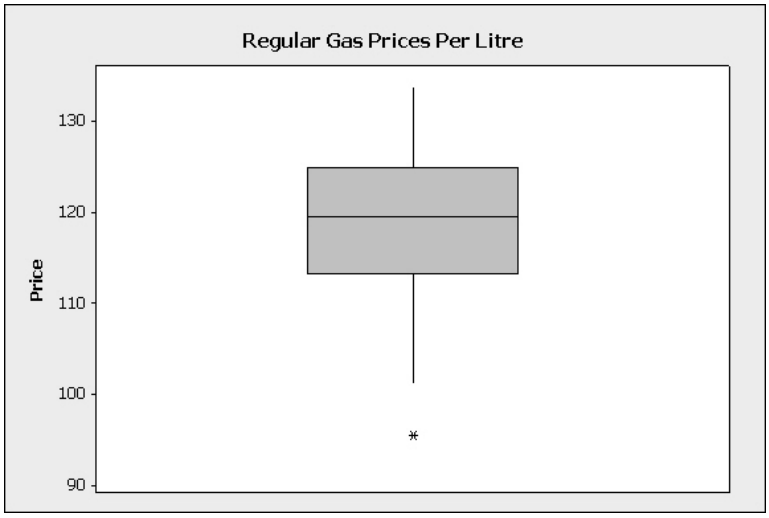
The text proposes the following way to find the third quartile: after putting the data in increasing order, take "the value of x in position $0.75(n + 1)$ … When $0.75(n + 1)$ is not an integer, the quartile is found by interpolation, using the values in the two adjacent positions." For this question, the location is at $0.75(n + 1) = 0.75(14) = 10.5$. The average of the tenth and eleventh ordered points is 124.39 + 125.51)/2 = 124.95 = Q_3.

b The interquartile range is $IQR = Q_3 - Q_1 = 124.95 - 113.20 = 11.75$.

c The text defines the formula for the lower fence as: $Q_1 - 1.5(IQR) = 113.20 - 1.5(11.75) = 95.575$.

d The text defines the formula for the upper fence as: $Q_3 + 1.5(IQR) = 124.95 + 1.5(11.75) = 142.75$.

e The box plot is as follows.



f Yes, there appears to be one outlier.

g As defined in the text, the z-score is $z = \dfrac{x - \overline{x}}{s}$. For the smallest observation (95.517), the z-score is $z = \dfrac{97.517 - 117.7445}{9.9585} = -2.03$. For the largest observation (133.79), $z = \dfrac{133.79 - 117.7445}{9.9585} = 1.61$. For the smallest observation, the z-score of −2.03 maybe somewhat unusually small.

h For Hamilton, the z-score is $z = \dfrac{119.524 - 117.7445}{9.9585} = 0.18$, which is not unusual.

i I would live where it is most expensive, so that people would drive their car less (on average).

2.53 **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 centres are roughly the same; the Sunmaid raisins appear slightly more variable.

2.54 **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 $\bar{x} = \dfrac{\sum x_i}{n} = \dfrac{155.5}{25} = 6.22$ and

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

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

c Calculate $\bar{x} \pm 2s = 6.22 \pm 6.994$ 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.

2.55 **a** The largest observation found in the data from Exercise 1.25 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

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

2.56 **a** Refer to Exercise 2.55. Since $\sum x_i = 418.4,$ the sample mean is

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

The three intervals of interest are shown in the following table, along with the number of observations that 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 1.25), but is skewed to the right.

2.57 The ordered data are shown below.

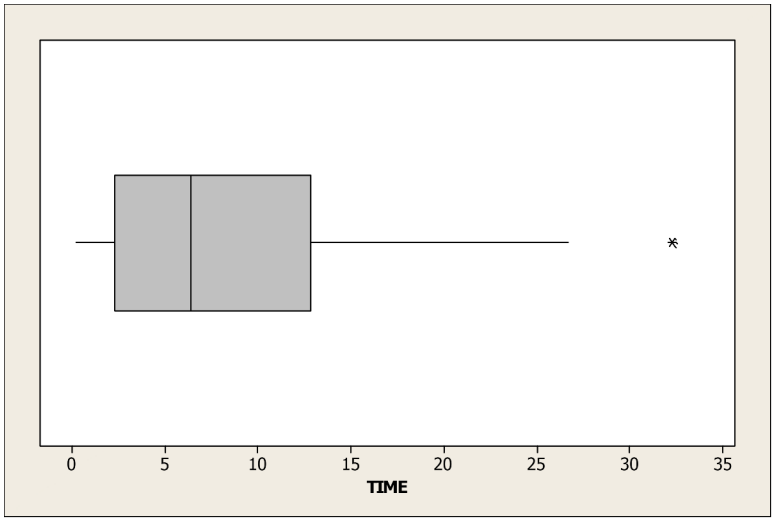
0.2	2.0	4.3	8.2	14.7
0.2	2.1	4.4	8.3	16.7
0.3	2.4	5.6	8.7	18.0
0.4	2.4	5.8	9.0	18.0
1.0	2.7	6.1	9.6	18.4
1.2	3.3	6.6	9.9	19.2
1.3	3.5	6.9	11.4	23.1
1.4	3.7	7.4	12.6	24.0

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 $m = \left(6.1 + 6.6\right)/2 = 6.35,\ Q_1 = 2.1 + 0.75(2.4 - 2.1) = 2.325,$ 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.



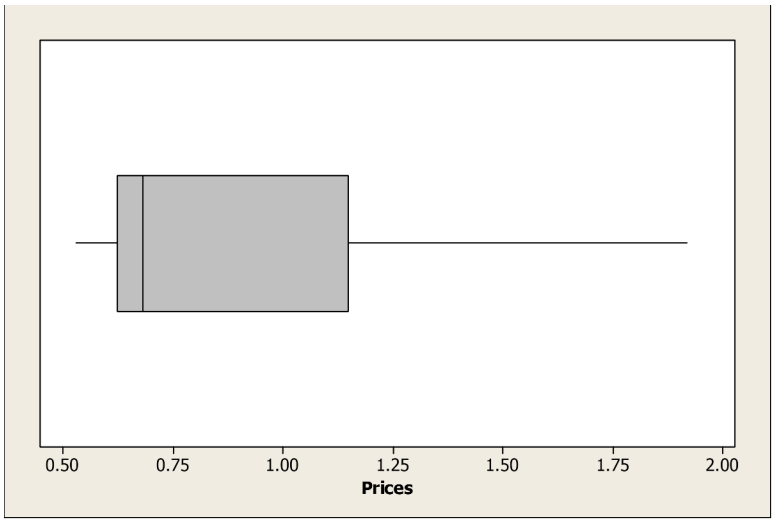
2.58 **a** For $n = 14$, the position of the median is $0.5(n+1) = 7.5$ and the positions of the quartiles are $0.25(n+1) = 3.75$ and $0.75(n+1) = 11.25.$ The lower quartile is three-fourths the way between the third and fourth measurements or $Q_1 = 0.60 + 0.75(0.63 - 0.60) = 0.6225$ and the upper quartile is one-fourth the way between the eleventh and twelveth 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 on the next page. 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

$$\bar{x} = \frac{\sum x_i}{n} = \frac{12.55}{14} = 0.896$$

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.

2.59 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; and 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; and 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.

c If the chemist had used a sample size of four for this experiment, the distribution would not be mound-shaped. Any possible histogram we could construct would be non-mound-shaped. We can use at most four classes, each with frequency 1, and we will not obtain a histogram that is even close to mound-shaped. Therefore, the Empirical Rule would not be suitable for describing $n = 4$ measurements.

2.60 Since it is not obvious that the distribution of amount of chloroform per litre of water in various water sources is mound-shaped, we cannot make this assumption. Tchebysheff's Theorem can be used, however, and the necessary intervals and fractions falling in these intervals are given in the table.

k	$\bar{x} \pm ks$	**Interval**	**Tchebysheff**
1	34 ± 53	–19 to 87	at least 0
2	34 ± 106	–72 to 140	at least 0.75
3	34 ± 159	–125 to 193	at least 0.89

2.61 The following information is available:

$$n = 400,\ \bar{x} = 600,\ s^2 = 4900$$

The standard deviation of these scores is then 70, and the results of Tchebysheff's Theorem follow:

k	$\bar{x} \pm ks$	**Interval**	**Tchebysheff**
1	600 ± 70	530 to 670	at least 0
2	600 ± 140	460 to 740	at least 0.75
3	600 ± 210	390 to 810	at least 0.89

If the distribution of scores is mound-shaped, we use the Empirical Rule, and conclude that approximately 68% of the scores would lie in the interval 530 to 670 (which is $\bar{x} \pm s$). Approximately 95% of the scores would lie in the interval 460 to 740.

2.62 **a** Calculate $n = 10,\ \sum x_i = 68.5,\ \sum x_i^2 = 478.375.$ Then

$$\bar{x} = \frac{\sum x_i}{n} = \frac{68.5}{10} = 6.85 \quad \text{and} \quad s = \sqrt{\frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1}} = \sqrt{\frac{478.375 - \frac{(68.5)^2}{10}}{9}} = 1.008$$

b The z-score for $x = 8.5$ is

$$z = \frac{x - \bar{x}}{s} = \frac{8.5 - 6.85}{1.008} = 1.64$$

This is not an unusually large measurement.

c The most frequently recorded measurement is the mode or $x = 7$ hours of sleep.

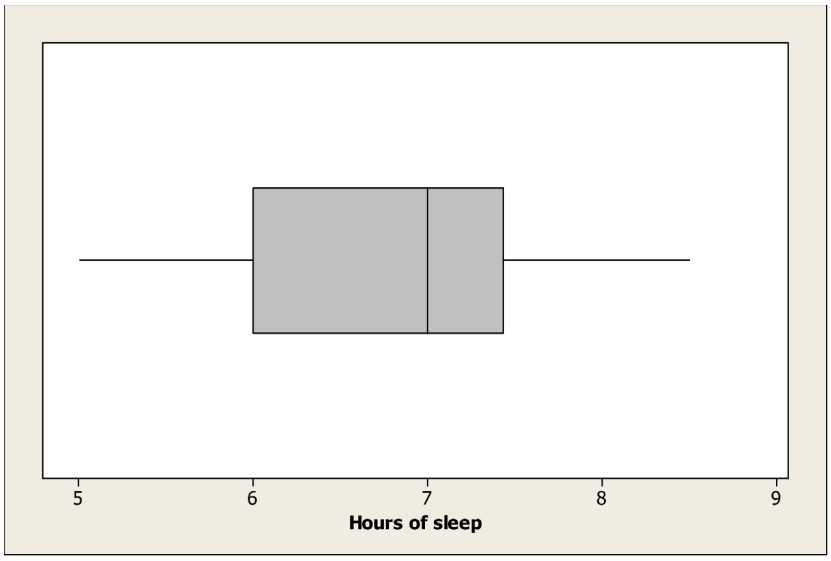
d For $n = 10$, the position of the median is $0.5(n+1) = 5.5$ and the positions of the quartiles are $0.25(n+1) = 2.75$ and $0.75(n+1) = 8.25.$ The sorted data are 5, 6, 6, 6.75, 7, 7, 7, 7.25, 8, 8.5. Then $m = \left(7 + 7\right)/2 = 7,\ Q_1 = 6 + 0.75(6 - 6) = 6$ and $Q_3 = 7.25 + 0.25(8 - 7.25) = 7.4375.$

Then $IQR = 7.4375 - 6 = 1.4375$ and the *lower and upper fences* are

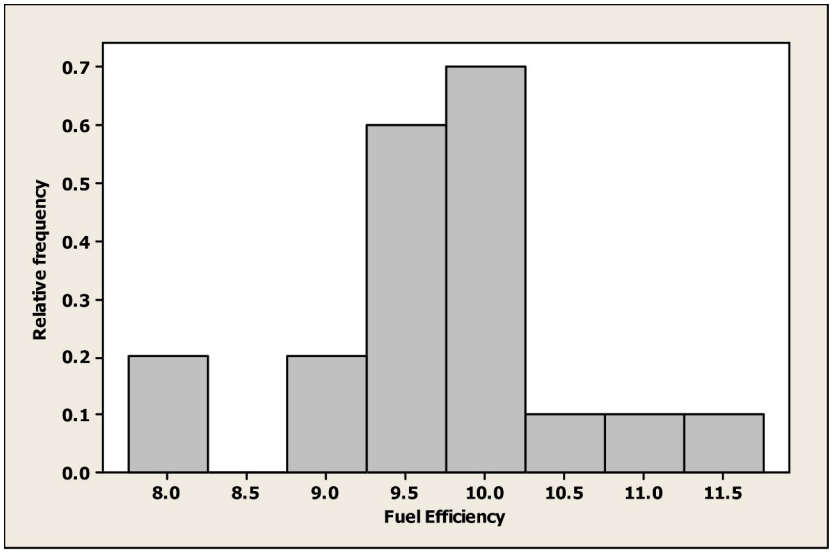
$$Q_1 - 1.5IQR = 6 - 2.15625 = 3.84$$

$$Q_3 + 1.5IQR = 7.4375 + 2.15625 = 9.59$$

There are no outliers (confirming the results of part **b**) and the box plot is shown on the next page.



2.63 **a** Answers will vary. A typical histogram is shown below. The distribution is slightly skewed to the left.



b Calculate $n = 20,\ \sum x_i = 193.1,\ \sum x_i^2 = 1876.65.$ Then

$$\bar{x} = \frac{\sum x_i}{n} = 9.655$$

$$s = \sqrt{\frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1}} = \sqrt{\frac{1876.65 - \frac{(193.1)^2}{20}}{19}} = \sqrt{0.646} = 0.804$$

c The sorted data is shown below:

```
7.9     8.1     8.9     8.9     9.4     9.4     9.5     9.5     9.6     9.7
9.8     9.8     9.9     9.9     10.0    10.1    10.2    10.3    10.9    11.3
```

The z-scores for $x = 7.9$ and $x = 11.3$ are

$$z = \frac{x - \bar{x}}{s} = \frac{7.9 - 9.655}{0.804} = -2.18 \text{ and } z = \frac{x - \bar{x}}{s} = \frac{11.3 - 9.655}{0.804} = 2.05$$

Since neither of the *z*-scores are greater than 3 in absolute value, the measurements are not judged to be outliers.

d The position of the median is $0.5(n+1)=10.5$ and the median is $m = (9.7 + 9.8)/2 = 9.75$.

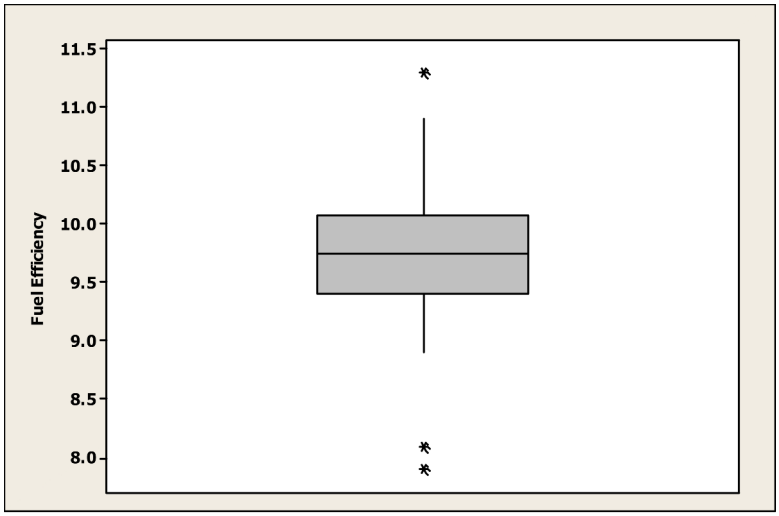
e The positions of the quartiles are $0.25(n+1)=5.25$ and $0.75(n+1)=15.75.$ Then $Q_1 = 9.4+0.25(9.4-9.4)=9.4$ and $Q_3 = 10.0+0.75(10.1-10.0)=10.075.$

2.64 Refer to Exercise 2.63. Calculate $IQR = 10.075 - 9.4 = 0.675.$ The *lower and upper fences* are

$$Q_1 - 1.5IQR = 9.4 - 1.5(.675) = 8.3875$$

$$Q_3 + 1.5IQR = 10.075 + 1.5(.675) = 11.0875$$

There are three outliers. The box plot is shown below.



2.65 **a** The range is $R = 71 - 40 = 31$ and the range approximation is $s \approx R/4 = 31/4 = 7.75.$

b Calculate $n=10,\ \sum x_i = 592,\ \sum x_i^2 = 36{,}014.$ Then

$$\bar{x} = \frac{\sum x_i}{n} = \frac{592}{10} = 59.2$$

$$s = \sqrt{\frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1}} = \sqrt{\frac{36{,}014 - \frac{(592)^2}{10}}{9}} = \sqrt{107.5111} = 10.369$$

The sample standard deviation calculated above is of the same order as the approximated value found in part **a**.

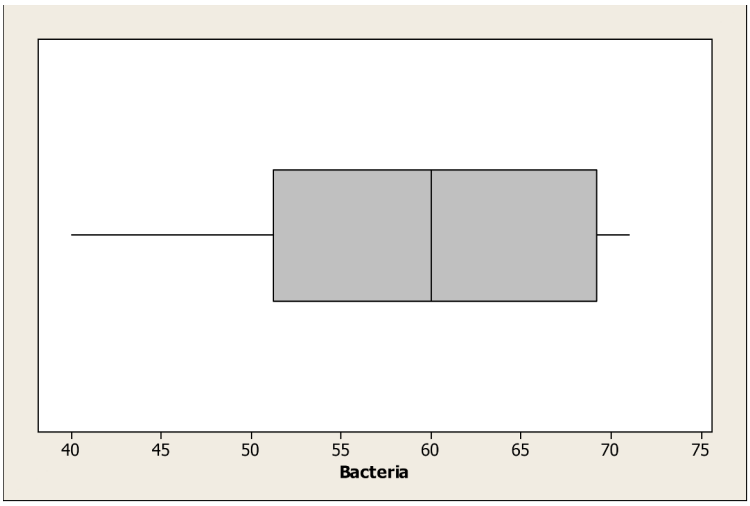
c The ordered set is 40, 49, 52, 54, 59, 61, 67, 69, 70, 71.
Since $n = 10$, the positions of m, Q_1, and Q_3 are 5.5, 2.75, and 8.25, respectively, and $m = (59+61)/2 = 60,\ Q_1 = 49 + 0.75(52-49) = 51.25,\ Q_3 = 69.25,$ and $IQR = 69.25 - 51.25 = 18.0.$

The *lower and upper fences* are

$$Q_1 - 1.5IQR = 51.25 - 27.00 = 24.25$$

$$Q_3 + 1.5IQR = 69.25 + 27.00 = 96.25$$

and the box plot is shown on the next page. There are no outliers and the data set is slightly skewed left.



2.66 The results of the Empirical Rule follow:

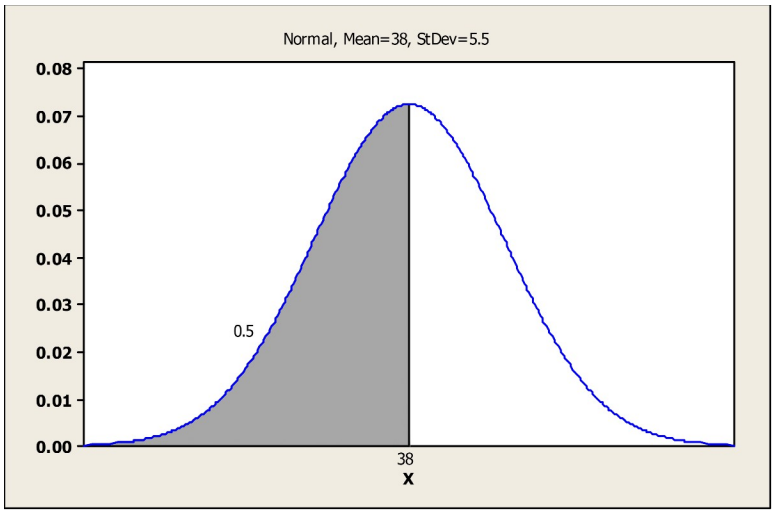
k	$\bar{x} \pm ks$	**Interval**	**Empirical Rule**
1	420 ± 5	415 to 425	approximately 0.68
2	420 ± 10	410 to 430	approximately 0.95
3	420 ± 15	405 to 435	approximately 0.997

Notice that we are assuming that attendance follows a mound-shaped distribution and hence that the Empirical Rule is appropriate.

2.67 If the distribution is mound-shaped with mean $\mu,$ then almost all of the measurements will fall in the interval $\mu \pm 3\sigma,$ which is an interval 6σ in length. That is, the range of the measurements should be approximately $6\sigma.$ In this case, the range is 800 − 200 = 600, so that $\sigma \approx 600/6 = 100.$

2.68 The stem lengths are approximately normal with mean 38 and standard deviation 5.5.

a In order to determine the percentage of roses with length less than 38, we must determine the proportion of the curve which lies within the shaded area (the lower half) in the figure below. Hence, the fraction below 38 would be 50%.



b The proportion of the area between 31 and 51 is

$$P(31 < x < 51) = P\left(\frac{31-38}{5.5} < z < \frac{51-38}{5.5}\right) = P(-1.27 < z < 2.36) \approx 89\%$$

the proportion of the area between 31 and 51 is approximately 89%.

2.69 **a** The range is $R = 172 - 108 = 64$ and the range approximation is $s \approx R/4 = 64/4 = 16.$

b Calculate $n = 15,\ \sum x_i = 2041,\ \sum x_i^2 = 281{,}807.$ Then

$$\bar{x} = \frac{\sum x_i}{n} = \frac{2041}{15} = 136.07$$

$$s = \sqrt{\frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1}} = \sqrt{\frac{281{,}807 - \frac{(2041)^2}{15}}{14}} = \sqrt{292.495238} = 17.102$$

c According to Tchebysheff's Theorem, with $k = 2$, at least 3/4 or 75% of the measurements will lie within $k = 2$ standard deviations of the mean. For this data, the two values, a and b, are calculated as $\bar{x} \pm 2s \Rightarrow 136.07 \pm 2(17.10) \Rightarrow 136.07 \pm 34.20$ or $a = 101.87$ and $b = 170.27$

2.70 The diameters of the trees are approximately mound-shaped with mean 35 and standard deviation 7.

a The value $x = 21$ lies two standard deviations below the mean, while the value $x = 55$ is approximately three standard deviations above the mean. Using the Empirical Rule, te fraction of trees with diameters between 21 and 35 is half of 0.95 or 0.475, while the fraction of trees with diameters between 35 and 55 is half of 0.997 or 0.4985. The total fraction of trees with diameters between 21 and 55 is 0.475 + 0.4985 = 0.9735.

b The value $x = 43$ lies approximately one standard deviation above the mean. Using the Empirical Rule, the fraction of trees with diameters between 35 and 43 is half of 0.68 or 0.34, and the fraction of trees with diameters greater than 43 is approximately 0.5 − 0.34 = 0.16.

2.71 **a** The range is $R = 19 - 4 = 15$ and the range approximation is $s \approx R/4 = 15/4 = 3.75.$

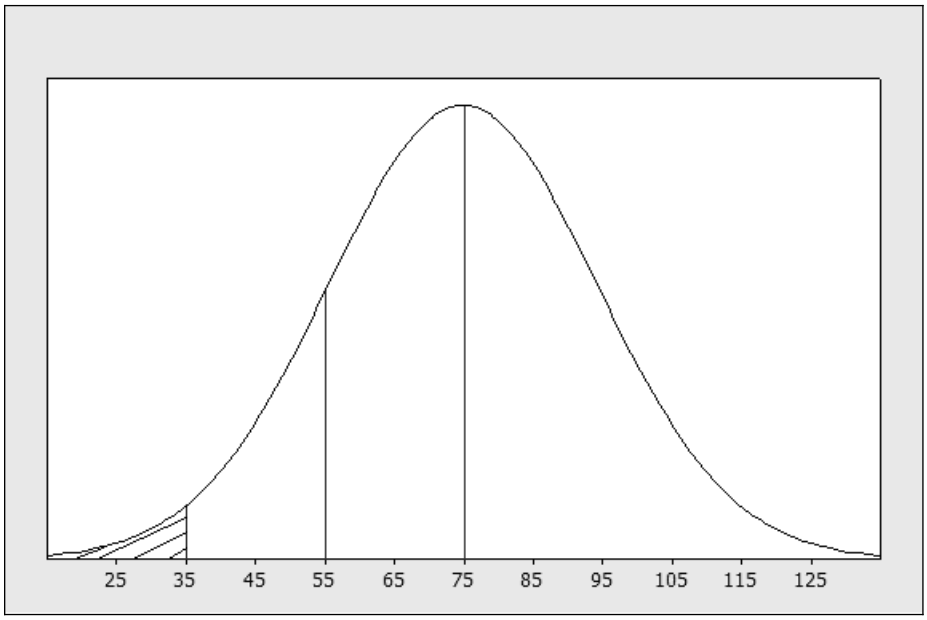
b Calculate $n = 15,\ \sum x_i = 175,\ \sum x_i^2 = 2237.$ Then

$$\bar{x} = \frac{\sum x_i}{n} = \frac{175}{15} = 11.67$$

$$s = \sqrt{\frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1}} = \sqrt{\frac{2237 - \frac{(175)^2}{15}}{14}} = \sqrt{13.95238} = 3.735$$

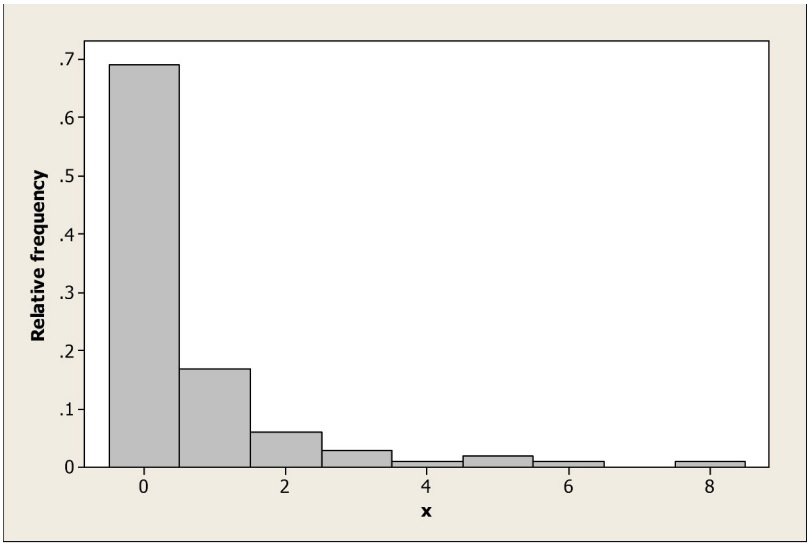
c Calculate the interval $\bar{x} \pm 2s \Rightarrow 11.67 \pm 2(3.735) \Rightarrow 11.67 \pm 7.47$ or 4.20 to 19.14. Referring to the original data set, the fraction of measurements in this interval is 14/15 = 0.93.

2.72 **a** It is known that duration times are approximately normal, with mean 75 and standard deviation 20. In order to determine the probability that a commercial lasts less than 35 seconds, we must determine the fraction of the curve that lies within the shaded area in the figure on the next page. Using the Empirical Rule, the fraction of the area between 35 and 75 is half of 0.95 or 0.475. Hence, the fraction below 35 would be 0.5 − 0.475 = 0.025.



b The fraction of the curve area that lies above the 55-second mark may again be determined by using the Empirical Rule. Refer to the figure in part **a**. The fraction between 55 and 75 is 0.34 and the fraction above 75 is 0.5. Hence, the probability that a commercial lasts longer than 55 seconds is 0.5 + 0.34 = 0.84.

2.73 **a** The relative frequency histogram for these data is shown below.



b Refer to the formulas given in Exercise 2.37. Using the frequency table and the grouped formulas, calculate $n = 100, \sum x_i f_i = 66, \sum x_i^2 f_i = 234.$ Then

$$\bar{x} = \frac{\sum x_i f_i}{n} = \frac{66}{100} = 0.66$$

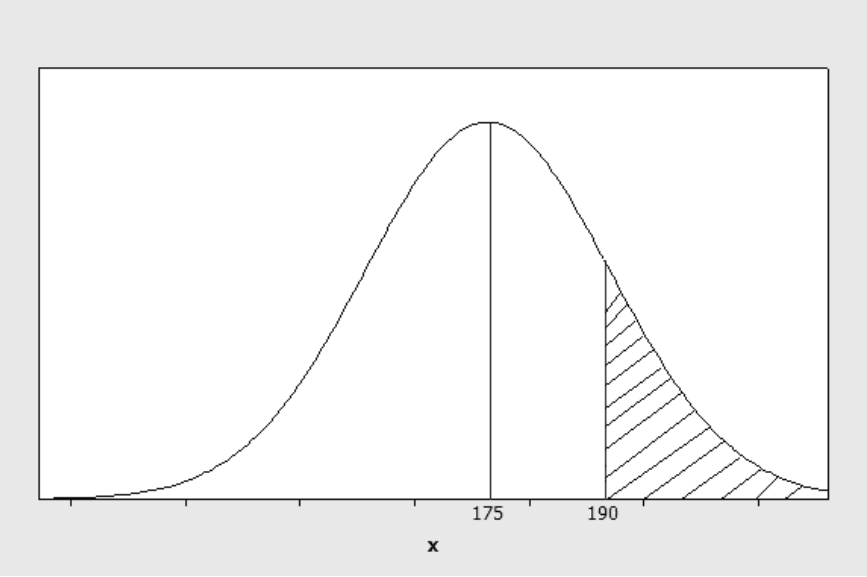
$$s^2 = \frac{\sum x_i^2 f_i - \frac{\left(\sum x_i f_i\right)^2}{n}}{n-1} = \frac{234 - \frac{(66)^2}{100}}{99} = 1.9236 \quad \text{and} \quad s = \sqrt{1.9236} = 1.39$$

c The three intervals, $\bar{x} \pm ks$ for $k = 2, 3$ 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	$\bar{x} \pm ks$	Interval	Fraction in Interval	Tchebyshef f	Empirical Rule
2	0.66 ± 2.78	–2.12 to 3.44	95/100 = 0.95	at least 0.75	≈ 0.95
3	0.66 ± 4.17	–3.51 to 4.83	96/100 = 0.96	at least 0.89	≈ 0.997

2.74 **a** The percentage of universities that have between 145 and 205 professors corresponds to the fraction of measurements expected to lie within two standard deviations of the mean. Tchebysheff's Theorem states that this fraction will be at least three-fourths or 75%.

b If the population is normally distributed, the Empirical Rule is appropriate and the desired fraction is calculated. Referring to the normal distribution shown below, the fraction of area lying between 175 and 190 is 0.34, so that the fraction of universities having more than 190 professors is 0.5 − 0.34 = 0.16.



2.75 We must estimate s and compare with the student's value of 0.263. In this case, $n = 20$ and the range is $R = 17.4 - 16.9 = 0.5$. The estimated value for s is then $s \approx R/4 = 0.5/4 = 0.125,$ which is less than 0.263. It is important to consider the magnitude of the difference between the "rule of thumb" and the calculated value. For example, if we were working with a standard deviation of 100, a difference of 0.142 would not be great. However, the student's calculation is twice as large as the estimated value. Moreover, two standard deviations, or $2(0.263) = 0.526,$ already exceeds the range. Thus, the value $s = 0.263$ is probably incorrect. The correct value of s is

$$s = \sqrt{\frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1}} = \sqrt{\frac{5{,}851.95 - \frac{117{,}032.41}{20}}{19}} = \sqrt{0.0173} = 0.132$$

2.76 Notice that Brett Hull has a relatively symmetric distribution. The whiskers are the same length and the median line is close to the middle of the box. There is an outlier to the right meaning that there is an extremely large number of goals during one of his seasons. The distributions for Mario and Bobby are skewed left, whereas the distribution for Wayne is slightly skewed right. The variability of the distributions is similar for Brett and Bobby. The variability for Mario and Wayne is similar and is much higher than the variability for Brett and Bobby. Wayne has long right whisker, meaning that there may be an unusually large number of goals during one of his seasons. Mario has the highest *IQR* and the highest median number of goals. The median number of goals for Brett Hull is the lowest (close to 38); the other three players are all about 41–44.

2.77 **a** Use the information in the exercise. For 1957 – 80, $IQR = 21,$ and the upper fence is
$Q_3 + 1.5IQR = 52 + 1.5(21) = 83.5$

For 1957 – 75, $IQR = 16,$ and the upper fence is
$Q_3 + 1.5IQR = 52.25 + 1.5(16) = 76.25$

b Although the maximum number of goals in both distribution is the same (77 goals), the upper fence is different in 1957 – 80, so that the record number of goals, $x = 77$ is no longer an outlier.

2.78 **a** Calculate $n = 50, \ \sum x_i = 418,$ so that $\bar{x} = \frac{\sum x_i}{n} = \frac{418}{50} = 8.36.$

b The position of the median is $0.5(n + 1) = 25.5$ and $m = (4 + 4)/2 = 4$.

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

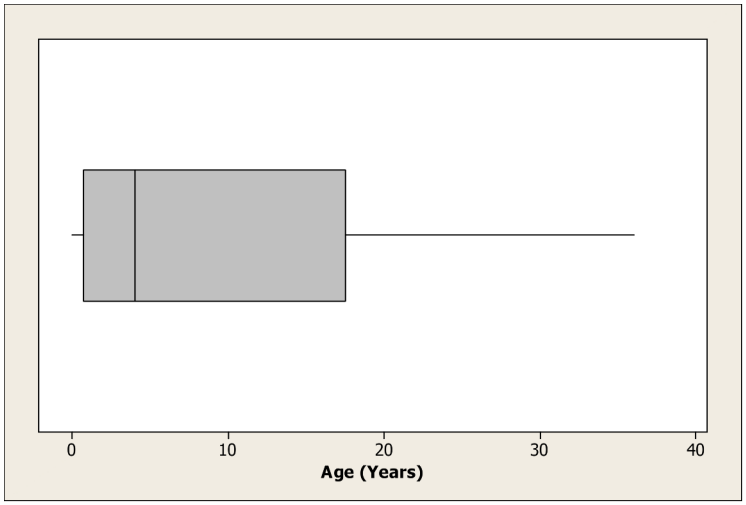
d Since $n = 50$, the positions of Q_1 and Q_3 are 0.25(51) = 12.75 and 0.75(51) = 38.25, respectively. Then $Q_1 = 0 + 0.75(1-0) = 0.75$, $Q_3 = 17 + 0.25(19-17) = 17.5$, and $IQR = 17.5 - 0.75 = 16.75$.

The *lower and upper fences* are

$$Q_1 - 1.5IQR = .75 - 25.125 = -24.375$$

$$Q_3 + 1.5IQR = 17.5 + 25.125 = 42.625$$

and the box plot is shown below. There are no outliers and the data is skewed to the right.



2.79 The variable of interest is the environmental factor in terms of the threat it poses to Canada. Each bulleted statement produces a percentile.

- x = toxic chemicals is the 61st percentile.
- x = air pollution and smog is the 55th percentile.
- x = global warming is the 52nd percentile.

2.80 Answers will vary. Students should notice that the distribution of baseline measurements is relatively mound-shaped. Therefore, the Empirical Rule will provide a very good description of the data. A measurement that is further than two or three standard deviations from the mean would be considered unusual.

2.81 **a** Calculate $n = 25,\ \sum x_i = 104.9$, $\sum x_i^2 = 454.810.$ Then

$$\bar{x} = \frac{\sum x_i}{n} = \frac{104.9}{25} = 4.196$$

$$s = \sqrt{\frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1}} = \sqrt{\frac{454.810 - \frac{(104.9)^2}{25}}{24}} = \sqrt{0.610} = 0.781$$

b The ordered data set is shown below:

```
2.5   3.0   3.1   3.3   3.6
3.7   3.8   3.8   3.9   3.9
4.1   4.2   4.2   4.2   4.3
4.3   4.4   4.7   4.7   4.8
4.8   5.2   5.3   5.4   5.7
```

c The *z*-scores for $x = 2.5$ and $x = 5.7$ are

$$z = \frac{x - \bar{x}}{s} = \frac{2.5 - 4.196}{0.781} = -2.17 \text{ and } z = \frac{x - \bar{x}}{s} = \frac{5.7 - 4.196}{0.781} = 1.93$$

Since neither of the *z*-scores are greater than 3 in absolute value, the measurements are not judged to be unusually large or small.

2.82 **a** For $n = 25$, the position of the median is $0.5(n+1) = 13$ and the positions of the quartiles are $0.25(n+1) = 6.5$ and $0.75(n+1) = 19.5.$ Then $m = 4.2,$ $Q_1 = (3.7 + 3.8)/2 = 3.75,$ and $Q_3 = (4.7 + 4.8)/2 = 4.75.$ Then the five-number summary is

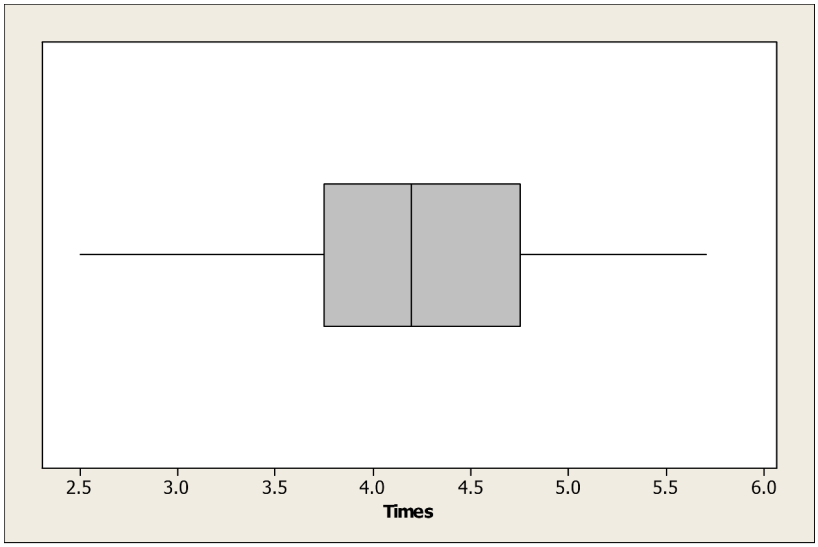
Min	Q_1	Median	Q_3	Max
2.5	3.75	4.2	4.75	5.7

b–c Calculate $IQR = Q_3 - Q_1 = 4.75 - 3.75 = 1.$ Then the *lower and upper fences* are

$$Q_1 - 1.5IQR = 3.75 - 1.5 = 2.25$$

$$Q_3 + 1.5IQR = 4.75 + 1.5 = 6.25$$

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



d Answers will vary. A stem and leaf plot, generated by *MINITAB,* is shown below. The data is roughly mound-shaped.

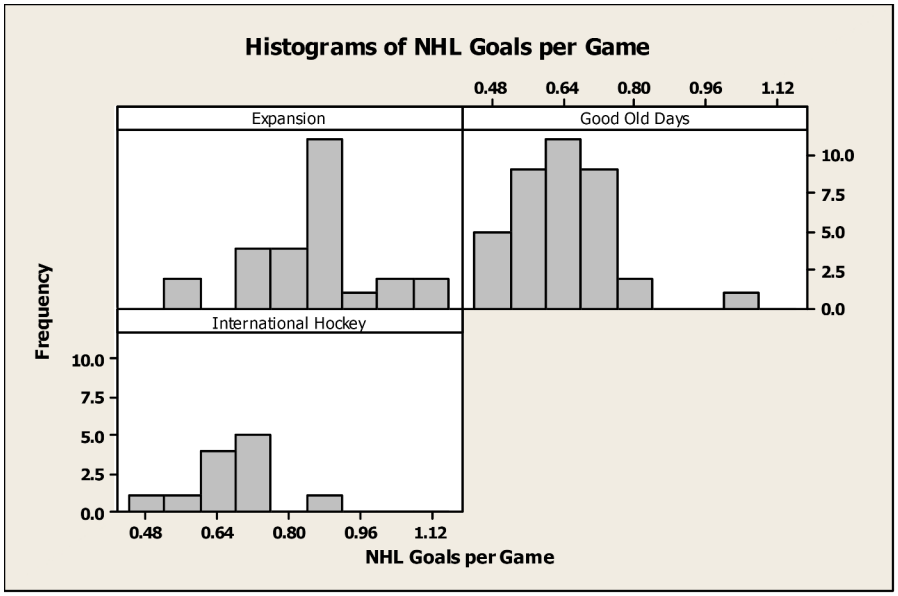
Stem and Leaf Plot: Times

```
Stem and leaf of Times  N  = 25
Leaf Unit = 0.10

 1   2  5
 4   3  013
 10  3  678899
(7)  4  1222334
 8   4  7788
 4   5  234
 1   5  7
```


Case Study: The Boys of Winter

1 The *MINITAB* computer package was used to analyze the data. In the printout below, various descriptive statistics as well as histograms and box plots are shown.

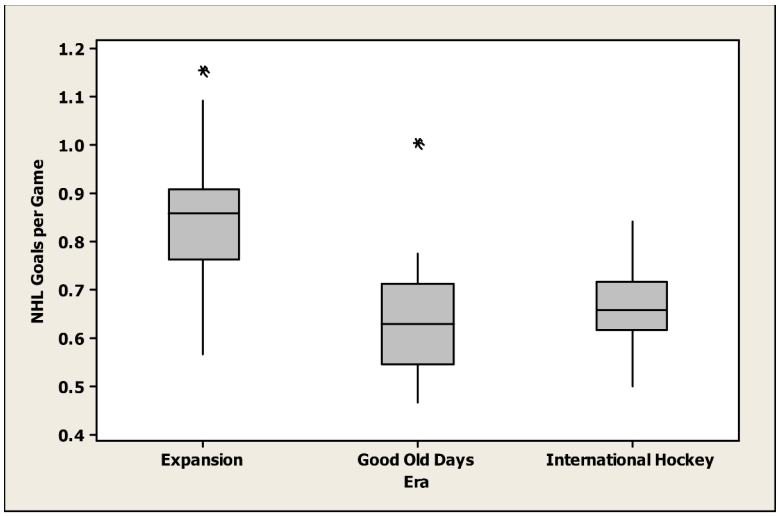


Descriptive Statistics: Average

Variable	Era	Total Count	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3
Average	1	37	0.6345	0.0180	0.1095	0.4667	0.5423	0.6286	0.7113
	2	26	0.8505	0.0280	0.1426	0.5658	0.7594	0.8562	0.9067
	3	12	0.6633	0.0249	0.0864	0.5000	0.6159	0.6585	0.7128

Variable	Era	Maximum	Range	IQR
Average	1	1.0000	0.5333	0.1690
	2	1.1500	0.5842	0.1473
	3	0.8415	0.3415	0.0969

2 Notice that the average goals per game is the least in the Good Old Days era (1 = 1931–1967) and the highest in the Expansion era (2 = 1968–1993). The Expansion era is the most variable. Although the International Hockey era (3 = 1994–2006) has slightly higher average goals per game than Good Old Days era, it is noticeably less variable.



3 The box plot shows that each of the Expansion and Good Old Days eras has one outlier. There is no outlier in the International Hockey era, the least variable era.

4 In summary, the Expansion era is quite different than the other two eras; it has higher mean and median number of goals per game. The outlier in the Expansion era indicates the season with the record-high goals per game. Notice that there is very little difference between the Good Old Days and International Hockey eras.

Project 2: Ignorance Is Not Bliss (Project 1-B continued)

a The sample mean is $\bar{x} = \frac{\Sigma x_i}{n}$. For this example, we have

$$\bar{X} = \frac{22 + 19 + 21 + \cdots + 27 + 33}{25} = \frac{544}{25} = 21.76$$

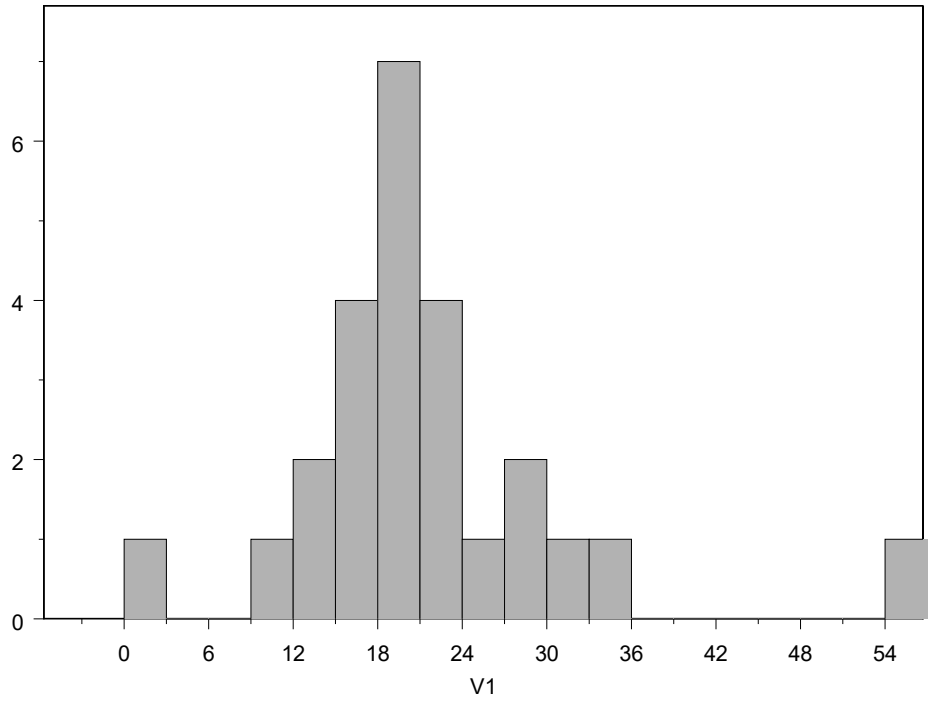
The sample mode is the value that occurs most often. For our data set, there are two: at 19 and at 22. Thus, there are two modes.

To find the median, first put the data into increasing order, as follows:

0	10	14	15	16	17	17	18	19	19	19	20
20	21	21	22	22	22	23	27	29	30	33	35
55											

The median is the value that is found in the middle position. For 25 data points, the middle value is the thirteenth largest, which in this case is 20.

From the histogram below, it can be concluded that the data is essentially mound-shaped.



b Since we have a small data set with a large outlier, the median would be the best choice in this situation. The sample median is less sensitive to outliers and thus gives a more accurate representation of the centre of this distribution. The mean of small samples, such as this one, is heavily influenced by outliers, such as the point $x = 55$ here, and therefore does not give an accurate representation of centre.

c The sample standard deviation can be calculated as follows:

$$s = \sqrt{\frac{\Sigma\left(x_i^2\right) - \frac{\left(\Sigma x_i\right)^2}{n}}{n-1}} = \sqrt{\frac{14,234 - \frac{\left(544\right)^2}{25}}{25-1}} = \sqrt{\frac{2,396.56}{24}} = 9.992831$$

The range (R) is simply the difference between the maximum value and the minimum value. For our data set, $R = 55 - 0 = 55$. The approximation for s based on R (as espoused in Section 2.5 of the text) is $s \approx \frac{R}{4}$ or 55/4 = 13.75, which is a decent approximation, but certainly not very good.

d If all data points were increased by 4%, the mean would also increase by 4%, or be multiplied by 1.04. This can be proven as follows. Assume that c is a constant. Then, we obtain:

$$\overline{x} = \frac{\Sigma x_i}{n} = \frac{\Sigma c x_i}{n} = \frac{c\Sigma x_i}{n} = c\overline{x}$$

Thus, the result follows if we let c equal 1.04.

e If all data points were raised by 5%, the standard deviation would also be raised by 5%. Recall the formula for the standard deviation and let c be a constant. Then

$$s = \sqrt{\frac{\Sigma\left(x_i^2\right) - \frac{\left(\Sigma x_i\right)^2}{n}}{n-1}}$$

$$\Rightarrow \sqrt{\frac{\Sigma\left(\left[cx_i\right]^2\right) - \frac{\left(\Sigma c x_i\right)^2}{n}}{n-1}} = \sqrt{\frac{\Sigma\left(c^2 x_i^2\right) - \frac{\left(c\Sigma x_i\right)^2}{n}}{n-1}} = \sqrt{\frac{c^2\Sigma\left(x_i^2\right) - \frac{c^2\left(\Sigma x_i\right)^2}{n}}{n-1}}$$

$$\Rightarrow \sqrt{\frac{c^2\left\{\Sigma\left(x_i^2\right) - \frac{\left(\Sigma x_i\right)^2}{n}\right\}}{n-1}} = \sqrt{c^2}\sqrt{\frac{\Sigma\left(x_i^2\right) - \frac{\left(\Sigma x_i\right)^2}{n}}{n-1}} = c \cdot s$$

Thus, we see that the standard deviation will be multiplied by the constant multiple c, in this case, 1.05.

f From part **a**, it was calculated that $\overline{x} = 21.76$ and $s = 9.992831.$ Then, the interval $\overline{x} \pm s$ becomes $x \in \left[11.767169, 31.752831\right].$ Counting the number of x values between 12 and 31 (inclusive), there are 20 entries. Thus, $\frac{20}{25} = 80\%$ of the entries are within the interval $\overline{x} \pm s.$ Now, computing the domain of the interval $\overline{x} \pm 2s,$ it can be seen that $x \in \left[1.774338, 41.745662\right].$ There are 23 values of x between 2 and 41 (inclusive), thus, $\frac{23}{25} = 92\%.$ Comparing to the Empirical Rule, normal distributions should have approximately 68% of the total values of x within $\overline{x} \pm s$ and 95% of the total values within $\overline{x} \pm 2s.$ Thus, it can be seen that there are more measurements than predicted for the first interval and slightly less than predicted for the second interval. This discrepancy can be accounted for by the non-normal behaviour in the distribution and the small amount of data points in the sample. Finally, Tchebysheff's Theorem predicts that at least 0% of the measurements are within $\overline{x} \pm s$ and that at least 75% of the measurements are within $\overline{x} \pm 2s.$ As such, this sample follows Tchebysheff's Theorem.

g Yes, Tchebysheff's Theorem can be used to describe this data set, since it can be used for *any* distribution.

h The Empirical Rule has a limited use in describing this sample. The data is relatively mound-shaped, and so the Empirical Rule is somewhat appropriate. However, due to the outliers in the data set, the Empirical

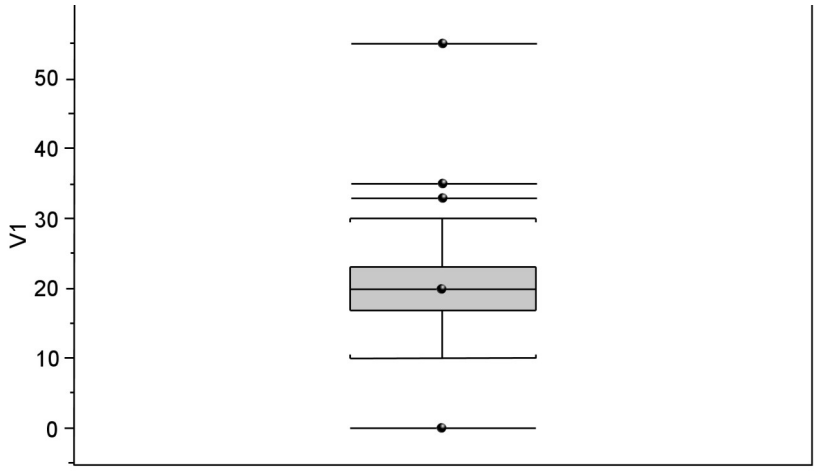
Rule fails to accurately predict the percentage of measurements within the interval $\bar{x} \pm s$. It provides a better approximation for the interval $\bar{x} \pm 2s$.

i Referring to plot made in part **a**, it seems as though points 0 and 55 could be outliers, as these are far from the bulk of the data in the centre.

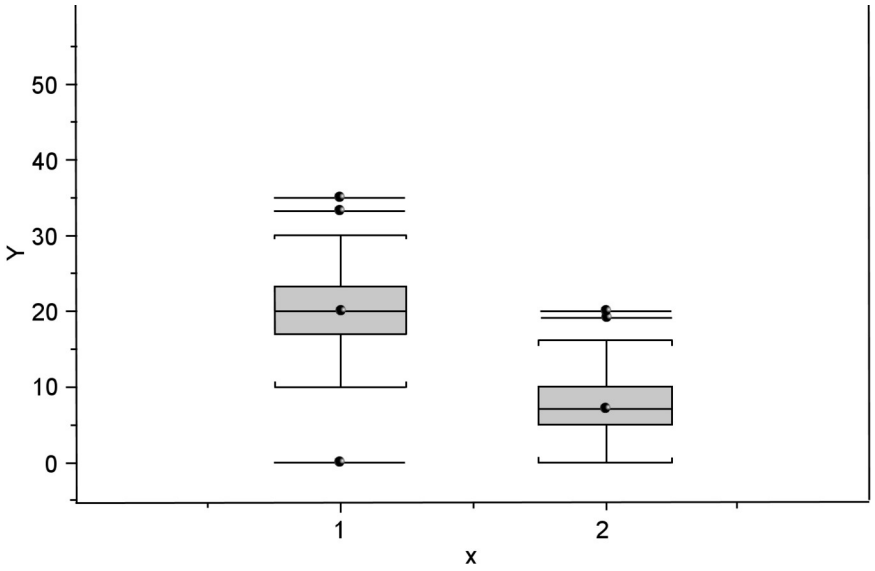
j Given that $n = 25$, the 25th percentile is found at $x = 0.25(n+1) = 0.25(26) = 6.5$. Thus, the 25th percentile is the average of the sixth and seventh measurements when they are arranged in increasing order, $Q_1 = \frac{17+17}{2} = 17.$ Likewise, $x = 0.50(26) = 13$ and thus $Q_2 = 20$. Lastly, $x = 0.75(26) = 19.5$, thus $Q_3 = \frac{23+27}{2} = 25.$ The interquartile range is therefore $IQR = Q_3 - Q_1 = 25 - 17 = 8$.

k The range is 55 as calculated in part **c**, whereas the $IQR = 8$. Thus, about 50% of the data can be found in a very narrow middle area of the data range, suggesting that a mound-shaped distribution is likely.

l A box plot is shown below. The box plot shows four outliers, as indicated by the points beyond the whiskers.



m The side-by-side box plots for the Hand Washing Time after (box 1) and before (box 2) the training session are given below:



n It is clear from the box plots that the average time spent washing hands improved after the training session.

o Yes, we can conclude that the session was useful, and that the conjecture was true. A difference in median times of 20 seconds and 7 seconds is substantial.

p The histogram for the both data sets taken together is shown below. There is clear evidence of two distinct peaks, suggesting that the data comprises of two distinct populations (which we know is true).