Chapter 1 Solutions

1.1. Exam1 = 95, Exam2 = 98, Final = 96.

1.2. For this student, TotalPoints $= 2 \cdot 88 + 2 \cdot 85 + 3 \cdot 77 + 2 \cdot 90 + 80 = 837$, so the grade is B.

1.3. The cases are apartments. There are five variables: rent (quantitative), cable (categorical), pets (categorical), bedrooms (quantitative), distance to campus (quantitative).

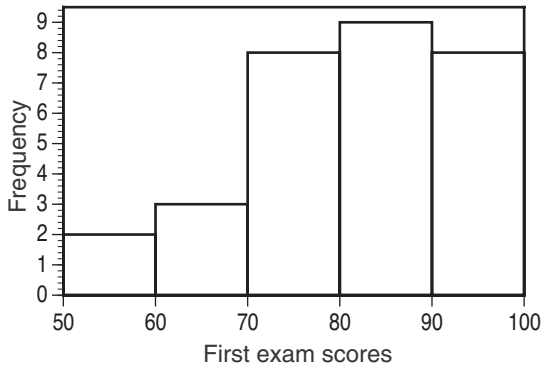
1.4. 31.3% $=$ 8.7% $+$ 22.6% of young adults have either a bachelor's degree or an associate degree.

1.5. Shown are two possible stemplots; the first uses split stems (described on page 11 of the text). The scores are slightly left-skewed; most range from 70 to the low 90s.

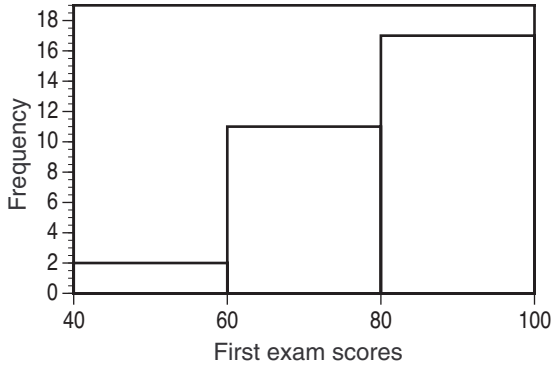
5	58
6	0
6	58
7	0023
7	5558
8	00003
8	5557
9	0002233
9	8

5	58
6	058
7	00235558
8	000035557
9	00022338

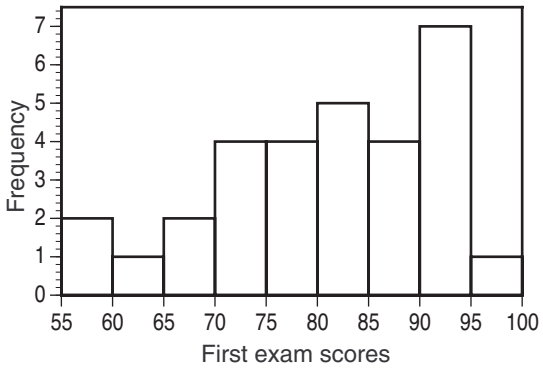
1.6. Student preferences will vary. The stemplot has the advantage of showing each individual score. Note that this histogram has the same shape as the second histogram in the previous exercise.



1.7. The larger classes hide a lot of detail.



1.8. This histogram shows more details about the distribution (perhaps more detail than is useful). Note that this histogram has the same shape as the first histogram in the solution to Exercise 1.6.



1.9. Using either a stemplot or histogram, we see that the distribution is left-skewed, centered near 80, and spread from 55 to 98. (Of course, a histogram would not show the exact values of the maximum and minimum.)

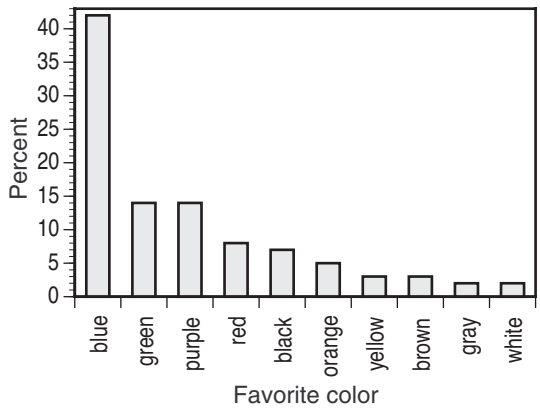
1.10. Recall that categorical variables place individuals into groups or categories, while quantitative variables "take numerical values for which arithmetic operations... make sense."

Variables (a), (d), and (e)—age, amount spent on food, and height—are quantitative. The answers to the other three questions—about dancing, musical instruments, and broccoli—are categorical variables.

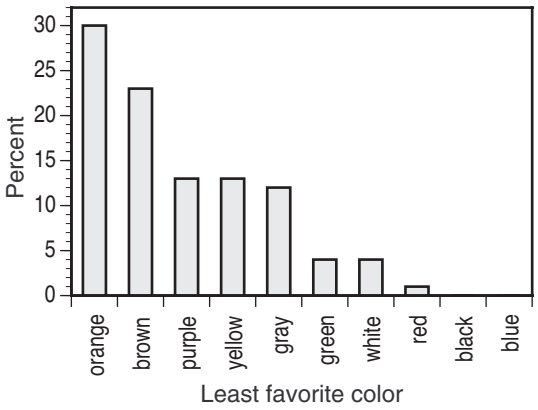
1.13. Possible responses would include heart rate before and after exercise (typically measured with a watch and fingertip), number of sit-ups (no instrument required), time to run 100 m (measured with a stopwatch).

1.14. Student answers will vary. Recent rankings in *U.S. News and World Report* used "16 measures of academic excellence," including academic reputation (measured by surveying college and university administrators), retention rate, graduation rate, class sizes, faculty salaries, student-faculty ratio, percentage of faculty with highest degree in their fields, quality of entering students (ACT/SAT scores, high school class rank, enrollment-to-admission ratio), financial resources, and the percentage of alumni who give to the school.

1.15. For example, blue is by far the most popular choice; 70% of respondents chose 3 of the 10 options (blue, green, and purple).



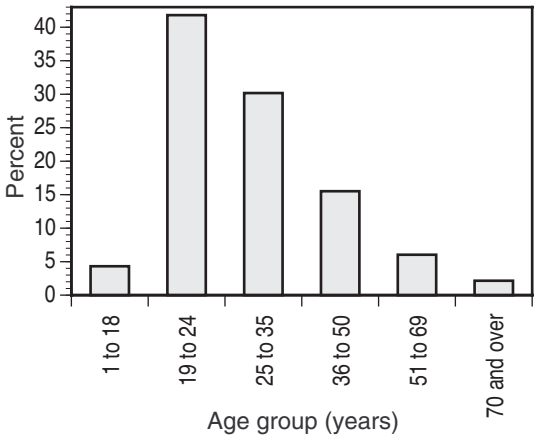
1.16. For example, opinions about least-favorite color are somewhat more varied than favorite colors. Interestingly, purple is liked and disliked by about the same fractions of people.



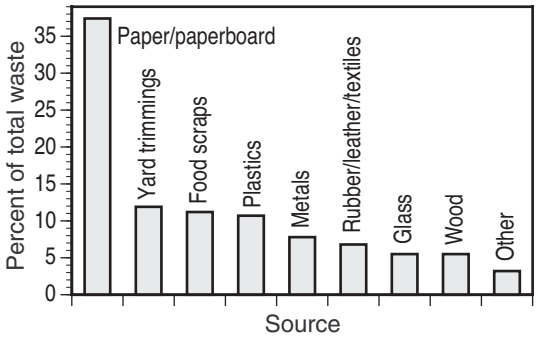
1.17. **(a)** There were 232 total respondents. The percents are given in the table below; for example, $\frac{10}{232} \doteq 4.31\%$. **(b)** The bar graph is shown below. **(c)** For example, 87.5% of the group were between 19 and 50. **(d)** The age-group classes do not have equal width: The first is 18 years wide, the second is 6 years wide, the third is 11 years wide, etc.

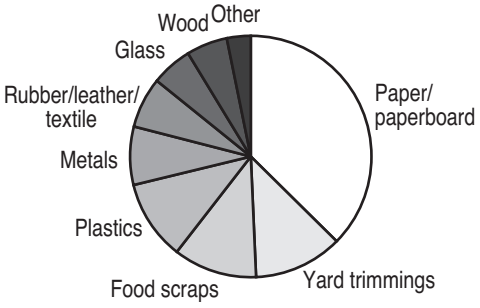
Note: *In order to produce a histogram from the given data, the bar for the first age group would have to be three times as wide as the second bar, the third bar would have to be wider than the second bar by a factor of 11/6, etc. Additionally, if we change a bar's width by a factor of x, we would need to change that bar's height by a factor of 1/x.*

Age group (years)	Percent
1 to 18	4.31%
19 to 24	41.81%
25 to 35	30.17%
36 to 50	15.52%
51 to 69	6.03%
70 and over	2.16%

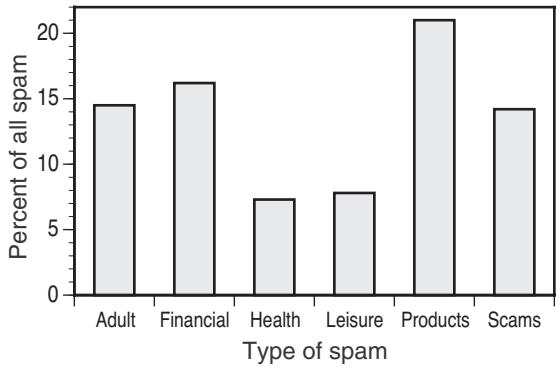


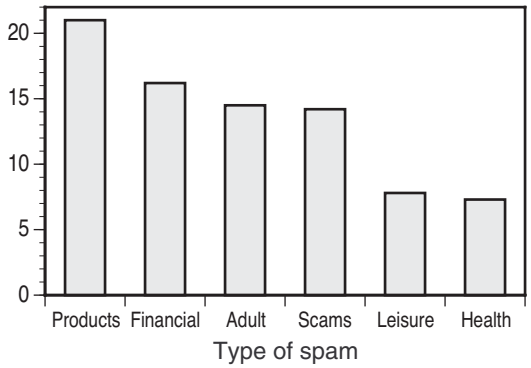
1.18. **(a)** The weights add to 231.8 million tons. **(b)** & **(c)** The bar and pie graphs are shown below.



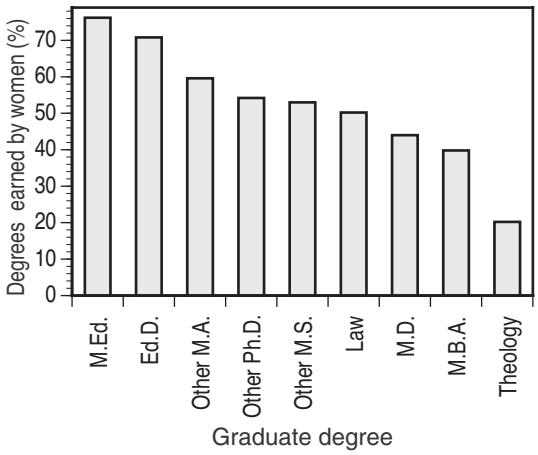


1.19. The two bar graphs are shown below.





1.20. **(a)** The given percentages refer to nine distinct groups (all M.B.A. degrees, all M.Ed. degrees, and so on) rather than one single group. **(b)** Bar graph shown on the right. Bars are ordered by height, as suggested by the text; students may forget to do this or might arrange in the opposite order (smallest to largest).



1.21. **(a)** Alaska is 5.7% (the leaf 7 on the stem 5) and Florida 17.6% (leaf 6 on stem 17). **(b)** The distribution is roughly symmetric (perhaps slightly skewed to the left), centered near 13% (the median [see Section 1.2] is 12.85%). Ignoring the outliers, the percentages are spread from 8.5% to 15.6%.

1.22. Shown on the right are the original stemplot (as given in the text for Exercise 1.21, minus Alaska and Florida) and the split-stems version students were asked to construct for this exercise. Preferences may vary between the two.

8	5
9	679
10	6
11	02233677
12	0011113445789
13	00012233345568
14	034579
15	36

8	5
9	
9	679
10	
10	6
11	02233
11	677
12	001111344
12	5789
13	0001223334
13	5568
14	034
14	579
15	3
15	6

1.23. Shown is the stemplot; as the text suggests, we have trimmed numbers (dropped the last digit) and split stems. 359 mg/dl appears to be an outlier. Overall, glucose levels are not under control: Only 4 of the 18 had levels in the desired range.

0	799
1	0134444
1	5577
2	0
2	57
3	
3	5

1.24. The back-to-back stemplot on the right suggests that the individual-instruction group was more consistent (their numbers have less spread) but not more successful (only two had numbers in the desired range).

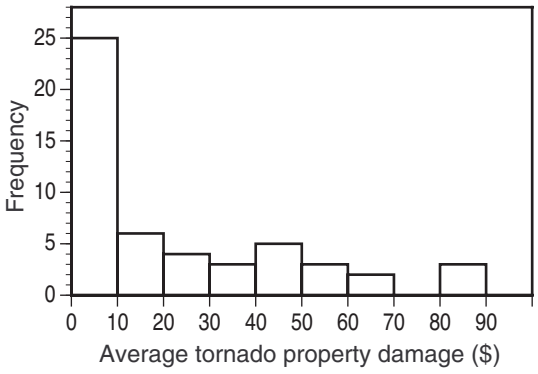
Individual		Class
	0	799
22	1	0134444
99866655	1	5577
22222	2	0
8	2	57
	3	
	3	5

1.25. The distribution is roughly symmetric, centered near 7 (or "between 6 and 7"), and spread from 2 to 13.

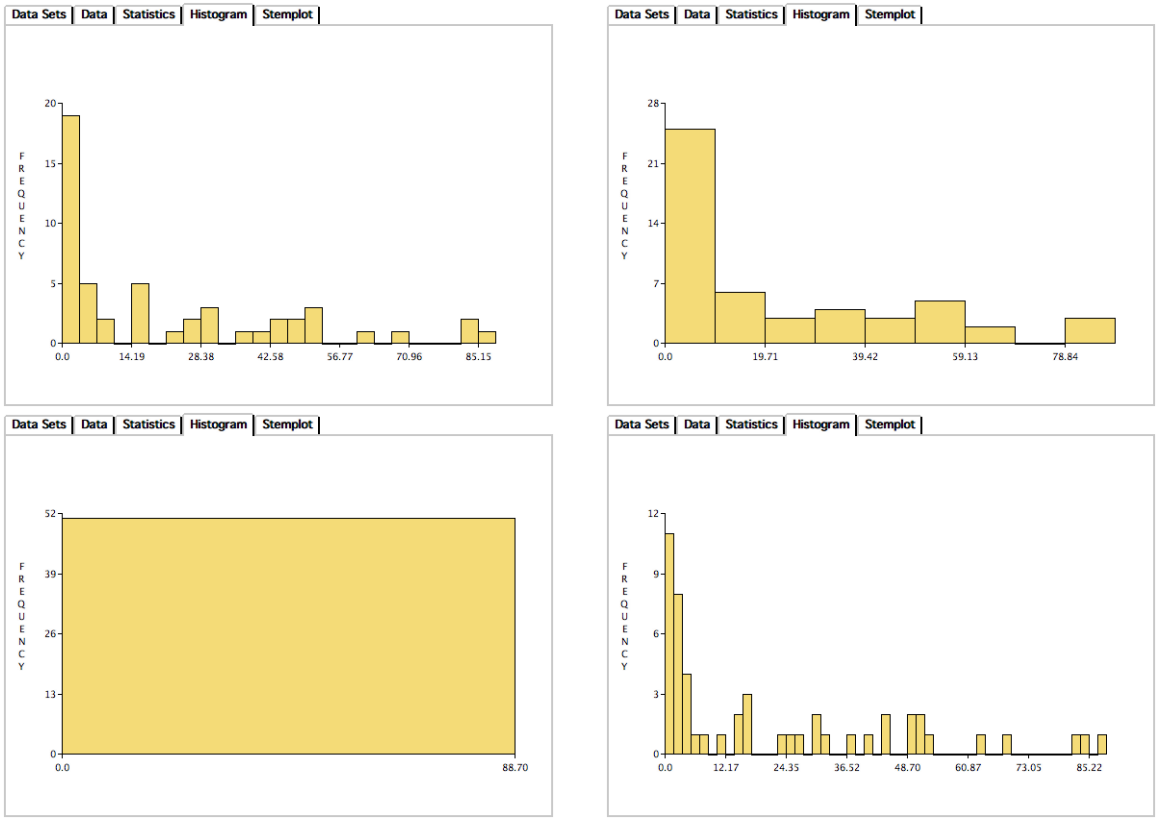
1.26. This distribution is skewed to the right, meaning that Shakespeare's plays contain many short words (up to six letters) and fewer very long words. We would probably expect most authors to have skewed distributions, although the exact shape and spread will vary.

1.27. There are three peaks in the histogram: One at \$4–6,000, one at \$18–20,000, and one at \$28–30,000. There is a clear break between the least expensive schools and the rest; the line between the middle and most expensive schools is not so clear. Presumably, the lowest group (up to \$10,000) includes public institutions, the highest group (starting around \$25,000) exclusive private schools like Harvard, and the middle group other private schools. Of course, these are generalizations; there may be a few exceptions (low-priced private schools or selective public schools).

1.28. **(a)** The top five states are Texas, Minnesota, Oklahoma, Missouri, and Illinois. The bottom five are Alaska, Puerto Rico, Rhode Island, Nevada, and Vermont. **(b)** The histogram (right) shows a sharp right skew, with a large peak (25 of the 51 numbers) in the "less than 10" category; arguably, that category is the "center" of the distribution. The distribution is spread from \$0 to about \$90; the top three states (Texas, Minnesota, Oklahoma) might be considered outliers, as that bar is separated from the rest (no states fell in the \$70–80 category). **(c)** The default histogram will vary with the software used.



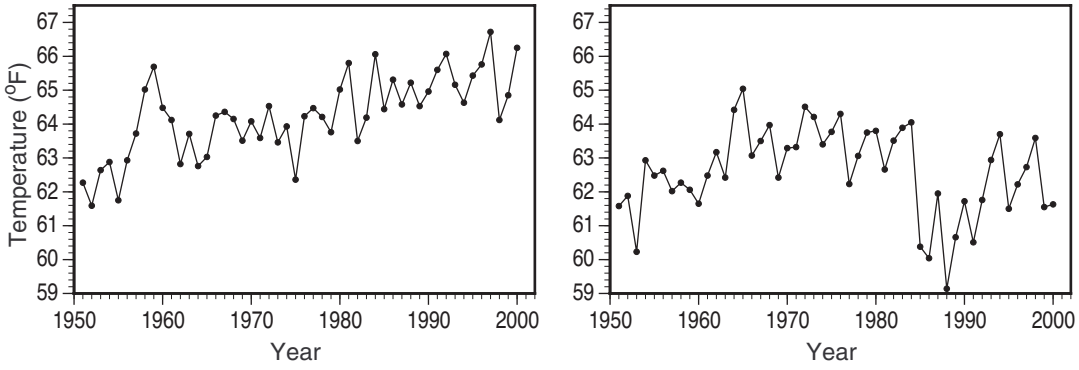
1.29. **(a)** The applet defaulted (for me) to 25 intervals. This histogram is shown below, along with the nine-class histogram. Note that the latter does not *exactly* match the histogram of the previous problem because the applet's classes are about 9.85 units wide, rather than 10 units wide. **(b)** The one-class and 51-class histograms are shown below. **(c)** Student opinions about which number of classes is best will vary, but something between 6 and 12 seems like a good range.



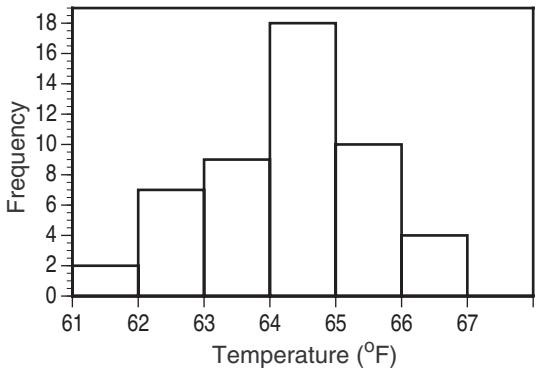
1.30. **(a)** Totals emissions would almost certainly be higher for very large countries; for example, we would expect that even with great attempts to control emissions, China (with over 1 billion people) would have higher total emissions than the smallest countries in the data set. **(b)** A stemplot is shown; a histogram would also be appropriate. We see a strong right skew with a peak from 0 to 0.2 metric tons per person and a smaller peak from 0.8 to 1. The three highest countries (the United States, Canada, and Australia) appear to be outliers; apart from those countries, the distribution is spread from 0 to 11 metric tons per person.

0	00000000000000011111
0	222233333
0	445
0	6677
0	888999
1	001
1	
1	
1	67
1	9

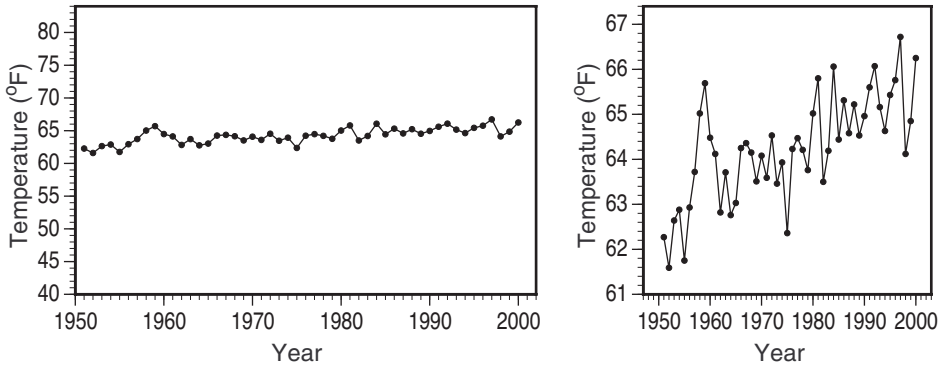
1.31. Shown below are two separate graphs (Pasadena on the left, Redding on the right); students may choose to plot both time series on a single set of axes. If two graphs are created, they should have the same vertical scale for easy comparison. Both plots show random fluctuation. Pasadena temperatures show an upward trend. Redding temperatures are initially similar to Pasadena's but dropped in the mid-1980s.



1.32. The distribution is symmetrical and mound-shaped, spread from 61°F to 67°F, with center 64–65°F. The histogram does not show what we see in the time plot from the previous exercise: That mean annual temperature has been rising over time.

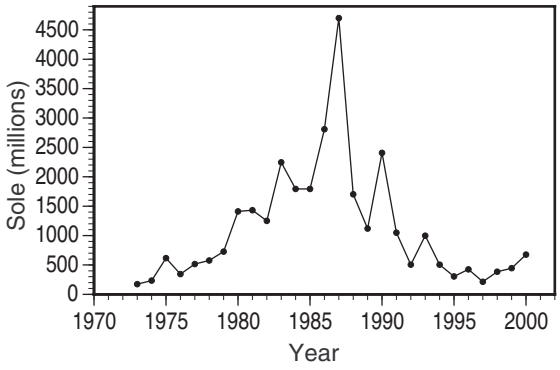


1.33. Shown below are two possible graphs.



1.34. **(a)** A stemplot is shown; a histogram would also be appropriate. The distribution is right-skewed, with a high outlier (4700 million). Other than the outlier, the numbers range from about 100 million to 2800 million sole. **(b)** The time plot shows that the number of recruits peaked in the mid-1980s and in recent years has fallen back to levels similar to those in the 1970s.

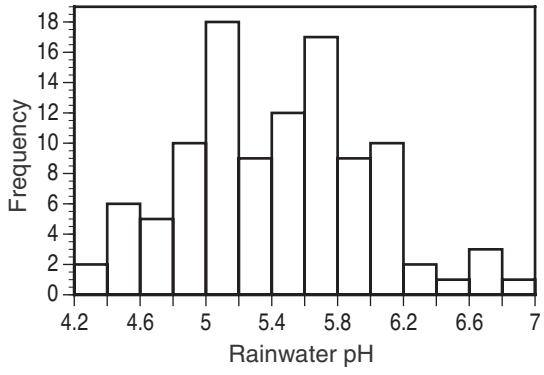
```
0 | 12233344
0 | 55556679
1 | 01244
1 | 777
2 | 24
2 | 8
3 |
3 |
4 |
4 | 7
```

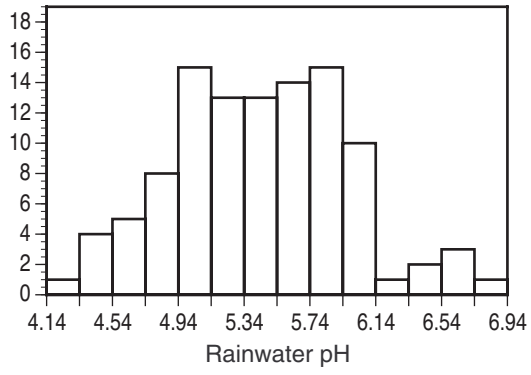


1.35. A stemplot or a histogram is appropriate for displaying the distribution. We see that the DT scores are skewed to the right, centered near 5 or 6, spread from 0 to 18. There are no outliers. We might also note that only 11 of these 264 women (about 4%) scored 15 or higher.

```
0 | 00000000000000000000000000000000000000011111111111111111111
0 | 22222222222222222333333333333333333333333
0 | 44444444444444444445555555555555555555
0 | 66666666666666666667777777777777
0 | 88888888888888999999999999999999
1 | 000000000000111111111
1 | 2222222222233333333333
1 | 444444455
1 | 66666777
1 | 8
```

1.36. **(a)** The first histogram shows two modes: 5–5.2 and 5.6–5.8. **(b)** The second histogram has peaks in locations close to those of the first, but these peaks are much less pronounced, so they would usually be viewed as distinct modes. **(c)** The results will vary with the software used.





1.37. The upper-left graph is studying time (Question 4); it is reasonable to expect this to be right-skewed (many students study little or not at all; a few study longer).

The graph in the lower right is the histogram of student heights (Question 3): One would expect a fair amount of variation but no particular skewness to such a distribution.

The other two graphs are handedness (upper right) and gender (lower left)—unless this was a particularly unusual class! We would expect that right-handed students should outnumber lefties substantially. (Roughly 10 to 15% of the population as a whole is left-handed.)

1.38. Sketches will vary. The distribution of coin years would be left-skewed because newer coins are more common than older coins.

1.39. A stemplot or a histogram is appropriate for displaying the distribution. We see that the data are skewed to the right with center near 30–40,000 barrels. At least the top two, and arguably the top three, observations are outliers; apart from these, the numbers are spread from 0 to 110,000 barrels.

0	00001111111111
0	22222223333333333333
0	44444445555555
0	6666667
0	8899
1	01
1	
1	5
1	
1	9
2	0

1.40. The stemplot gives more information than a histogram (since all the original numbers can be read off the stemplot), but both give the same impression. The distribution is roughly symmetric with one value (4.88) that is somewhat low. The center of the distribution is between 5.4 and 5.5 (the median is 5.46, the mean is 5.448).

48	8
49	
50	7
51	0
52	6799
53	04469
54	2467
55	03578
56	12358
57	59
58	5

1.41. (a) Not only are most responses multiples of 10; many are multiples of 30 and 60. Most people will "round" their answers when asked to give an estimate like this; in fact, the most striking answers are ones such as 115, 170, or 230. The students who claimed 360 minutes (6 hours) and 300 minutes (5 hours) may have been exaggerating. (Some students might also "consider suspicious" the student who claimed to study 0 minutes per night. As a teacher, I can easily believe that such students exist, and I suspect that some of your students might easily accept that claim as well.) **(b)** The stemplots suggest that women (claim to) study more than men. The approximate centers are 175 minutes for women and 120 minutes for men.

Women		Men
	0	033334
96	0	66679999
22222221	1	2222222
888888888875555	1	558
4440	2	00344
	2	
	3	0
6	3	

1.42. A stemplot is shown; a histogram would also be appropriate. The distribution is clearly right-skewed, centered near 100 days, and spread from 43 to 598 days. The split stems emphasize the skewness by showing the gaps. Some students might consider some of the highest numbers to be outliers.

```
0 | 44
0 | 555556677788888888889999999
1 | 000000000001112222333444
1 | 56777899
2 | 1144
2 |
3 | 2
3 | 8
4 | 0
4 |
5 | 12
5 | 9
```

1.43. **(a)** There are four variables: GPA, IQ, and self-concept are quantitative, while gender is categorical. (OBS is not a variable, since it is not really a "characteristic" of a student.) **(b)** Below. **(c)** The distribution is skewed to the left, with center (median) around 7.8. GPAs are spread from 0.5 to 10.8, with only 15 below 6. **(d)** There is more variability among the boys; in fact, there seems to be a subset of boys with GPAs from 0.5 to 4.9. Ignoring that group, the two distributions have similar shapes.

```
 0 | 5
 1 | 8
 2 | 4
 3 | 4689
 4 | 0679
 5 | 1259
 6 | 0112249
 7 | 22333556666666788899
 8 | 0000222223347899
 9 | 002223344556668
10 | 01678
```

Female		Male
	0	5
	1	8
	2	4
4	3	689
7	4	069
952	5	1
4210	6	129
98866533	7	223566666789
997320	8	0002222348
65300	9	2223445668
710	10	68

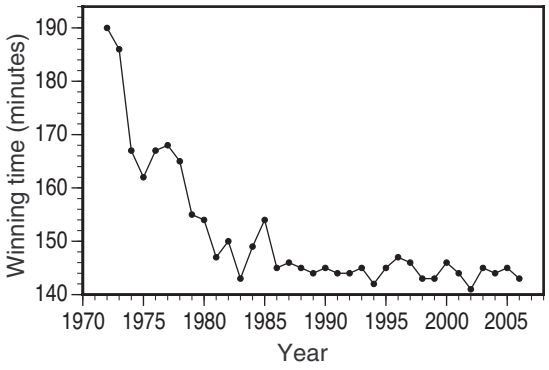
1.44. Stemplot at right, with split stems. The distribution is fairly symmetric—perhaps slightly left-skewed—with center around 110 (clearly above 100). IQs range from the low 70s to the high 130s, with a "gap" in the low 80s.

```
 7 | 24
 7 | 79
 8 |
 8 | 69
 9 | 0133
 9 | 6778
10 | 0022333344
10 | 555666777789
11 | 0000111122223334444
11 | 55688999
12 | 003344
12 | 677888
13 | 02
13 | 6
```

1.45. Stemplot at right, with split stems. The distribution is skewed to the left, with center around 59.5. Most self-concept scores are between 35 and 73, with a few below that, and one high score of 80 (but not really high enough to be an outlier).

2	01
2	8
3	0
3	5679
4	02344
4	6799
5	1111223344444
5	556668899
6	00001233344444
6	55666677777899
7	0000111223
7	
8	0

1.46. The time plot on the right shows that women's times decreased quite rapidly from 1972 until the mid-1980s. Since that time, they have been fairly consistent: All times since 1986 are between 141 and 147 minutes.



1.47. The mean score is $\bar{x} = \dfrac{821}{10} = 82.1$.

1.48. In order, the scores are:

$$55,\ 73,\ 75,\ 80,\ 80,\ 85,\ 90,\ 92,\ 93,\ 98$$

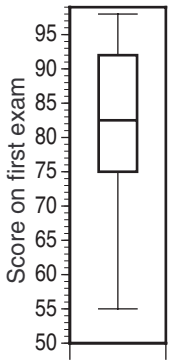
The middle two scores are 80 and 85, so the median is $M = \dfrac{80 + 85}{2} = 82.5$.

1.49. See the ordered list given in the previous solution.
The first quartile is $Q_1 = 75$, the median of the first five numbers: 55, 73, $\underline{75}$, 80, 80.
Similarly, $Q_3 = 92$, the median of the last five numbers: 85, 90, $\underline{92}$, 93, 98.

1.50. The median and quartiles were found in the previous two solutions; the minimum and maximum are easy to locate in the ordered list of scores (see the solution to Exercise 1.48), so the five-number summary is $\text{Min} = 55,\ Q_1 = 75,\ M = 82.5,\ Q_3 = 92,\ \text{Max} = 98$.

1.51. Use the five-number summary from the solution to Exercise 1.50:

$$\text{Min} = 55,\ \ Q_1 = 75,\ \ M = 82.5,\ \ Q_3 = 92,\ \ \text{Max} = 98$$



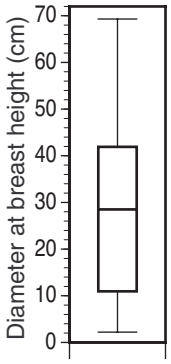
1.52. The interquartile range is $IQR = Q_3 - Q_1 = 92 - 75 = 17$, so the $1.5 \times IQR$ rule would consider as outliers scores outside the range $Q_1 - 25.5 = 49.5$ to $Q_3 + 25.5 = 117.5$. According to this rule, there are no outliers.

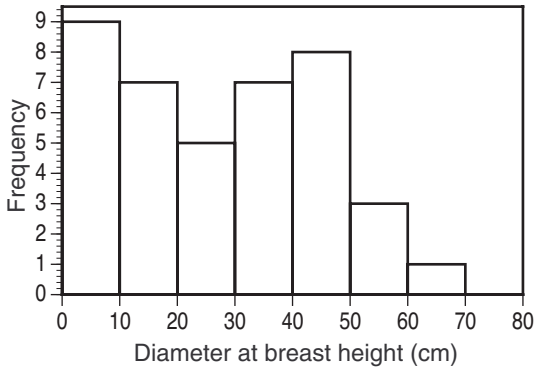
1.53. The variance *can* be computed from the formula $s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$; for example, the first term in the sum would be $(80 - 82.1)^2 = 4.41$. However, in practice, software or a calculator is the preferred approach; this yields $s^2 = \frac{1416.9}{9} = 157.4\overline{3}$ and $s = \sqrt{s^2} \doteq 12.5472$.

1.54. In order to have $s = 0$, all 5 cases must be equal; for example, 1, 1, 1, 1, 1, or 12.5, 12.5, 12.5, 12.5, 12.5. (If any two numbers are different, then $x_i - \bar{x}$ would be non-zero for some i, so the sum of squared differences would be positive, so $s^2 > 0$, so $s > 0$.)

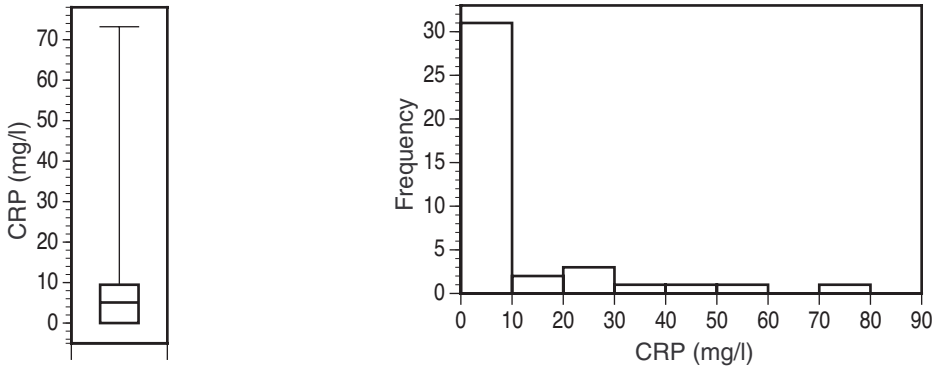
1.55. Divide total score by 4: $\frac{950}{4} = 237.5$ points.

1.56. **(a)** The five-number summary is $\text{Min} = 2.2$ cm, $Q_1 = 10.95$ cm, $M = 28.5$ cm, $Q_3 = 41.9$ cm, $\text{Max} = 69.3$ cm. **(b)** & **(c)** The boxplot and histogram are shown below. (Students might choose different interval widths for the histogram.) **(d)** Preferences will vary. Both plots reveal the right-skew of this distribution, but the boxplot does not show the two peaks visible in the histogram.

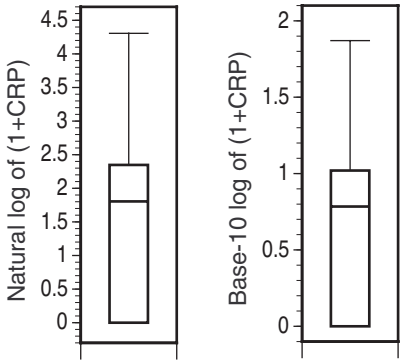




1.57. (a) The five-number summary is Min = 0 mg/l, $Q_1 = 0$ mg/l, $M = 5.085$ mg/l, $Q_3 = 9.47$ mg/l, Max = 73.2 mg/l. **(b)** & **(c)** The boxplot and histogram are shown below. (Students might choose different interval widths for the histogram.) **(d)** Preferences will vary. Both plots reveal the sharp right-skew of this distribution, but because Min = Q_1, the boxplot looks somewhat strange. The histogram seems to convey the distribution better.



1.58. Answers depend on whether natural (base-e) or common (base-10) logarithms are used. Both sets of answers are shown here. If this exercise is assigned, it would probably be best for the sanity of both instructor and students to specify which logarithm to use.

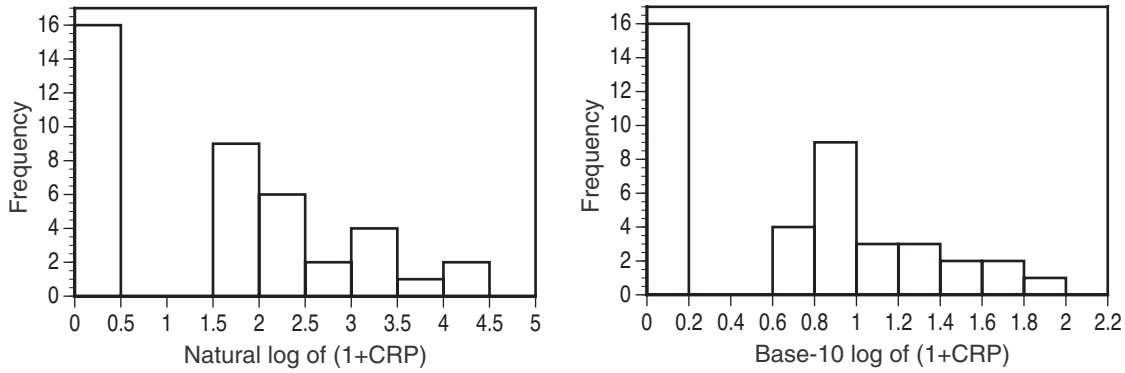


(a) The five-number summary is:

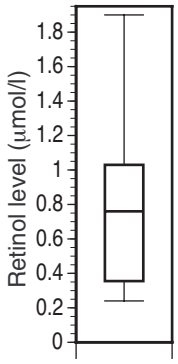
Logarithm	Min	Q_1	M	Q_3	Max
Natural	0	0	1.8048	2.3485	4.3068
Common	0	0	0.7838	1.0199	1.8704

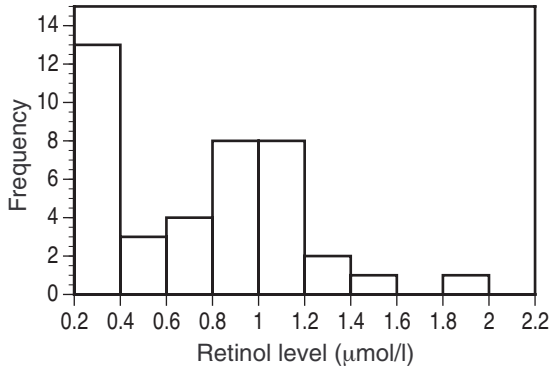
(The ratio between these answers is roughly $\ln 10 \doteq 2.3$.)

(b) & **(c)** The boxplots and histograms are shown below. (Students might choose different interval widths for the histograms.) **(d)** As for Exercise 1.57, preferences will vary.



1.59. **(a)** The five-number summary (in units of μmol/l) is Min $= 0.24$, $Q_1 = 0.355$, $M = 0.76$, $Q_3 = 1.03$, Max $= 1.9$. **(b)** & **(c)** The boxplot and histogram are shown below. (Students might choose different interval widths for the histogram.) **(d)** The distribution is right-skewed. A histogram (or stemplot) is preferable because it reveals an important feature not evident from a boxplot: This distribution has two peaks.





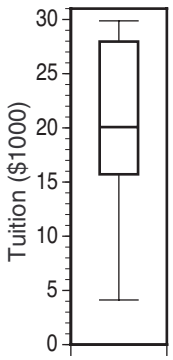
1.60. The mean and standard deviation for these ratings are $\bar{x} = 5.9$ and $s \doteq 3.7719$; the five-number summary is Min $= Q_1 = 1$, $M = 6.5$, $Q_3 =$ Max $= 10$. For a graphical presentation, a stemplot (or histogram) is better than a boxplot because the latter obscures details about the distribution. (With a little thought, one might realize that Min $= Q_1 = 1$ and $Q_3 =$ Max $= 10$ means that there are lots of 1's and lots of 10's, but this is much more evident in a stemplot or histogram.)

1	0000000000000000
2	0000
3	0
4	0
5	00000
6	000
7	0
8	000000
9	00000
10	000000000000000000

1.61. The five-number summary is:

\$4123 \$15,717 \$20,072 \$27,957.5 \$29,875

This and the boxplot on the right do not reveal the three groups of schools that are visible in the histogram. See also the solution to Exercise 1.27.



1.62. **(a)** The five-number summary is:

Min $= 5.7\%$, $Q_1 = 11.7\%$, $M = 12.75\%$, $Q_3 = 13.5\%$, Max $= 17.6\%$

(b) The *IQR* is $13.5\% - 11.7\% = 1.8\%$, so outliers are those numbers below $Q_1 - 2.7\% = 9\%$ and above $Q_3 + 2.7\% = 16.2\%$. Alaska and Florida are outliers, along with Utah (8.5%).

1.63. (a) The five-number summary (in 1999 dollars) is:

$\text{Min} = 0,\ Q_1 = 2.14,\ M = 10.64,\ Q_3 = 40.96,\ \text{Max} = 88.6$

The evidence for the skew is in the large gaps between the higher numbers; that is, the differences $Q_3 - M$ and $\text{Max} - Q_3$ are large compared to $Q_1 - \text{Min}$ and $M - Q_1$. **(b)** The IQR is $Q_3 - Q_1 = 38.82$, so outliers would be less than -56.09 or greater than 99.19. **(c)** The mean is 21.95 (1999 dollars), much greater than the median 10.64. The mean is pulled in the direction of the skew—in this case, to the right, making it larger.

1.64. See also the solution to Exercise 1.30. **(a)** The five-number summary (in units of metric tons per person) is:

$\text{Min} = 0,\ Q_1 = 0.75,\ M = 3.2,\ Q_3 = 7.8,\ \text{Max} = 19.9$

The evidence for the skew is in the large gaps between the higher numbers; that is, the differences $Q_3 - M$ and $\text{Max} - Q_3$ are large compared to $Q_1 - \text{Min}$ and $M - Q_1$. **(b)** The IQR is $Q_3 - Q_1 = 7.05$, so outliers would be less than -9.825 or greater than 18.375. According to this rule, only the United States qualifies as an outlier, but Canada and Australia seem high enough to also include them.

0	00000000000000011111
0	222233333
0	445
0	6677
0	888999
1	001
1	
1	
1	67
1	9

1.65. The distribution of household net worth would almost surely be strongly skewed to the right, perhaps more so for young households: A few would have earned (or inherited) substantial assets, but most have not had time to accumulate very much wealth. This strong skew pulls the mean to be higher than the median.

1.66. (a) $\bar{x} = 48.25$ and $M = 37.8$ thousand barrels of oil. The mean is made larger by the right skew. **(b)** The five-number summary (all measured in thousands of barrels) is:

$\text{Min} = 2,\ Q_1 = 21.505,\ M = 37.8,\ Q_3 = 60.1,\ \text{Max} = 204.9$

The evidence for the skew is in the large gaps between the higher numbers; that is, the differences $Q_3 - M$ and $\text{Max} - Q_3$ are large compared to $Q_1 - \text{Min}$ and $M - Q_1$.

1.67. The total salary is \$655,000, so the mean is $\bar{x} = \frac{\$655{,}000}{8} = \$81{,}875$. Seven of the eight employees (everyone but the owner) earned less than the mean. The median is $M = \$35{,}000$.

1.68. If three individuals earn \$0, \$0, and \$20,000, the reported median is \$20,000. If the two individuals with no income take jobs at \$14,000 each, the median decreases to \$14,000. The same thing can happen to the mean: In this example, the mean drops from \$20,000 to \$16,000.

1.69. The total salary is now \$790,000, so the new mean is $\bar{x} = \frac{\$790{,}000}{8} = \$98{,}750$. The median is unchanged.

1.70. Details at right.

$$\bar{x} = \frac{11{,}200}{7} = 1600$$

$$s^2 = \frac{214{,}872}{6} = 35{,}812 \text{ and}$$

$$s = \sqrt{35{,}812} \doteq 189.24$$

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1792	192	36864
1666	66	4356
1362	−238	56644
1614	14	196
1460	−140	19600
1867	267	71289
1439	−161	25921
11200	0	214872

1.71. The quote describes a distribution with a strong right skew: Lots of years with no losses to hurricane ($0), but very high numbers when they do occur. For example, if there is one hurricane in a 10-year period, the "average annual loss" for that period would be $100,000, but that does not adequately represent the cost for the year of the hurricane. Means are not the appropriate measure of center for skewed distributions.

1.72. **(a)** $\bar{x}$ and s are appropriate for symmetric distributions with no outliers. **(b)** Both high numbers are flagged as outliers. For women, $IQR = 60$, so the upper $1.5 \times IQR$ limit is 300 minutes. For men, $IQR = 90$, so the upper $1.5 \times IQR$ limit is 285 minutes. The table on the right shows the effect of removing these outliers.

	Women		Men	
	$\bar{x}$	s	$\bar{x}$	s
Before	165.2	56.5	117.2	74.2
After	158.4	43.7	110.9	66.9

1.73. **(a)** & **(b)** See the table on the right. In both cases, the mean and median are quite similar.

	$\bar{x}$	s	M
pH	5.4256	0.5379	5.44
Density	5.4479	0.2209	5.46

1.74. See also the solution to Exercise 1.43. **(a)** The mean of this distribution appears to be higher than 100. (There is no substantial difference between the standard deviations.) **(b)** The mean and median are quite similar; the mean is slightly smaller due to the slight left skew of the data. **(c)** In addition to the mean and median, the standard deviation is shown for reference (the exercise did not ask for it).

	$\bar{x}$	s	M
IQ	108.9	13.17	110
GPA	7.447	(2.1)	7.829

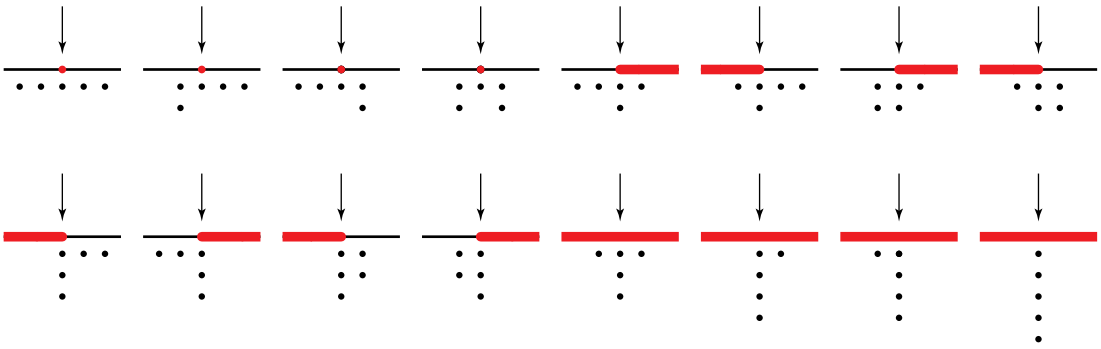
Note: *Students may be somewhat puzzled by the statement in (b) that the median is "close to the mean" (when they differ by 1.1), followed by (c), where they "differ a bit" (when $M - \bar{x} = 0.382$). It may be useful to emphasize that we judge the size of such differences relative to the spread of the distribution. For example, we can note that $\frac{1.1}{13.17} \doteq 0.08$ for (b), and $\frac{0.382}{2.1} \doteq 0.18$ for (c).*

1.75. With only two observations, the mean and median are always equal because the median is halfway between the middle two (in this case, the only two) numbers.

1.76. **(a)** The mean (green arrow) moves along with the moving point (in fact, it moves in the same direction as the moving point, at one-third the speed). At the same time, as long as the moving point remains to the right of the other two, the median (red arrow) points to the middle point (the right-most nonmoving point). **(b)** The mean follows the moving point as before. When the moving point passes the right-most fixed point, the median slides along with it until the moving point passes the leftmost fixed point, then the median stays there.

1.77. (a) There are several different answers, depending on the configuration of the first five points. *Most students* will likely assume that the first five points should be distinct (no repeats), in which case the sixth point *must* be placed at the median. This is because the median of 5 (sorted) points is the third, while the median of 6 points is the average of the third and fourth. If these are to be the same, the third and fourth points of the set of six must both equal the third point of the set of five.

The diagram below illustrates all of the possibilities; in each case, the arrow shows the location of the median of the initial five points, and the shaded region (or dot) on the line indicates where the sixth point can be placed without changing the median. Notice that there are four cases where the median does not change, regardless of the location of the sixth point. (The points need not be equally spaced; these diagrams were drawn that way for convenience.)



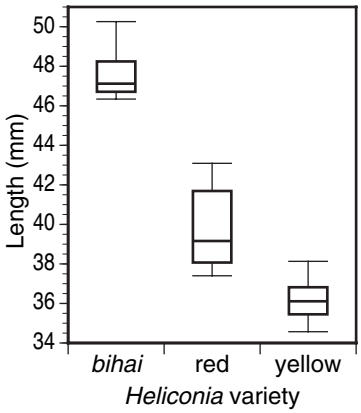
(b) Regardless of the configuration of the first five points, if the sixth point is added so as to leave the median unchanged, then in that (sorted) set of six, the third and fourth points must be equal. One of these two points will be the middle (fourth) point of the (sorted) set of seven, no matter where the seventh point is placed.

Note: *If you have a student who illustrates all possible cases above, then it is likely that the student either (1) obtained a copy of this solutions manual, (2) should consider a career in writing solutions manuals, (3) has too much time on his or her hands, or (4) both 2 and 3 (and perhaps 1) are true.*

1.78. The five-number summaries (all in millimeters) are:

	Min	Q_1	M	Q_3	Max
bihai	46.34	46.71	47.12	48.245	50.26
red	37.40	38.07	39.16	41.69	43.09
yellow	34.57	35.45	36.11	36.82	38.13

H. bihai is clearly the tallest variety—the shortest *bihai* was over 3 mm taller than the tallest red. Red is generally taller than yellow, with a few exceptions. Another noteworthy fact: The red variety is more variable than either of the other varieties.



1.79. (a) The means and standard deviations (all in millimeters) are:

Variety	$\bar{x}$	s
bihai	47.5975	1.2129
red	39.7113	1.7988
yellow	36.1800	0.9753

bihai		red		yellow	
46	3466789	37	4789	34	56
47	114	38	0012278	35	146
48	0133	39	167	36	0015678
49		40	56	37	01
50	12	41	4699	38	1
		42	01		
		43	0		

(b) *Bihai* and red appear to be right-skewed (although it is difficult to tell with such small samples). Skewness would make these distributions unsuitable for $\bar{x}$ and s.

1.80. The means and standard deviations (in units of trees) are:

Group	$\bar{x}$	s
1	23.7500	5.06548
2	14.0833	4.98102
3	15.7778	5.76146

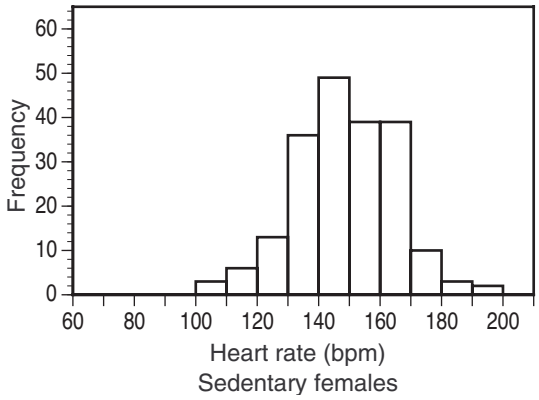
Never logged		1 year ago		8 years ago	
0		0	2	0	4
0		0	9	0	
1		1	2244	1	22
1	699	1	57789	1	5889
2	0124	2	0	2	22
2	7789	2		2	
3	3	3		3	

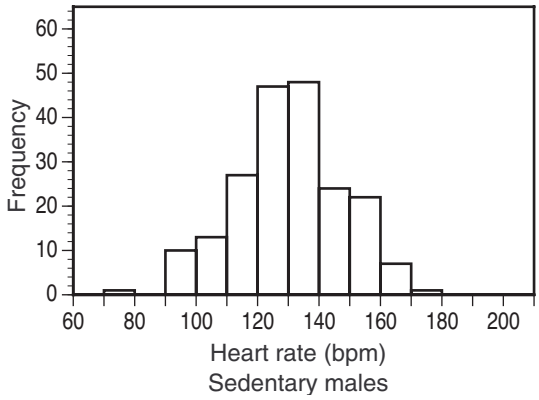
The means, along with the stemplots on the right, appear to suggest that logging reduces the number of trees per plot and that recovery is slow (the 1-year-after and 8-years-after means and stemplots are similar). Use of $\bar{x}$ and s should be acceptable, as the distributions show no extreme outliers or strong skewness (given the small sample sizes).

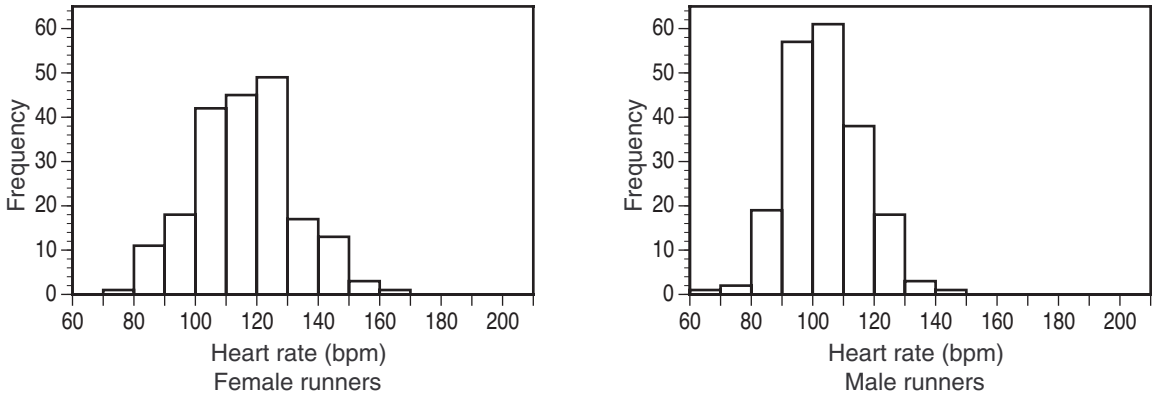
1.81. Either stemplots or histograms could be used to display the distributions, although with four sets of 200 subjects each, histograms are simpler. All four distributions are symmetric with no outliers, so means and standard deviations are appropriate; they are in the table on the right (in units of bpm). The average heart rate for runners is about 30 bpm less than the average sedentary rate.

Group	$\bar{x}$	s
Sedentary females	148.00	16.27
Sedentary males	130.00	17.10
Female runners	115.99	15.97
Male runners	103.97	12.50

Note: *Students might also observe that women generally have higher heart rates than men in the same activity-level group, but that is not an effect of running.*







1.82. Note that estimates for (a) and (b) will vary. **(a)** The median would be in position $\frac{14{,}959+1}{2} = 7480$ in the list; from the boxplot, we estimate it to be about \$45,000. **(b)** The quartiles would be in positions 3740 and 11,220, and we estimate their values to be about \$32,000 and \$65,000. **(c)** Omitting these observations should have *no* effect on the median and quartiles. (The quartiles are computed from the entire set of data; the extreme 5% are omitted only in locating the ends of the lines for the boxplot.)

Note: *The positions of the quartiles were found according to the text's method; that is, these are the locations of the medians of the first and second halves of the list. Students might instead compute* $0.25 \times 14,959$ *and* $0.75 \times 14,959$ *to obtain the answers 3739.75 and 11,219.25.*

1.83. The 5th and 95th percentiles would be approximately in positions 748 and 14,211. The "whiskers" on the box extend to approximately \$13,000 and \$137,000. (Estimates may vary.)

1.84. All five income distributions are skewed to the right. As highest education level rises, the median, quartiles, and extremes rise—that is, all five points on the boxplot increase. Additionally, the width of the box (the *IQR*) and the distance from one extreme to the other (the difference between the 5th and 95th percentiles) also increase, meaning that the distributions become more and more spread out.

1.85. The minimum and maximum are easily determined to be 1 and 12 letters, and the quartiles and median can be found by adding up the bar heights. For example, the first two bars have total height about 22% or 23% (less than 25%); adding the third bar brings the total to about 45%, so Q_1 must equal 3 letters. Continuing this way, we find that the five-number summary, in units of letters, is:

$$\text{Min} = 1,\ Q_1 = 3,\ M = 4,\ Q_3 = 5,\ \text{Max} = 12$$

1.86. Because the mean is to be 7, the five numbers must add up to 35. Also, the third number (in order from smallest to largest) must be 10 because that is the median. Beyond that, there is some freedom in how the numbers are chosen.

Note: *It is likely that many students will interpret "positive numbers" as meaning positive integers only, which leads to eight possible solutions, shown below.*

1	1	10	10	13	1	1	10	11	12	1	2	10	10	12	1	2	10	11	11
1	3	10	10	11	1	4	10	10	10	2	2	10	10	11	2	3	10	10	10

1.87. The simplest approach is to take (at least) six numbers—say, a, b, c, d, e, f in increasing order. For this set, $Q_3 = e$; we can cause the mean to be larger than e by simply choosing f to be *much* larger than e. For example, if all numbers are nonnegative, $f > 5e$ would accomplish the goal because then

$$\bar{x} = \frac{a+b+c+d+e+f}{6} > \frac{e+f}{6} > \frac{e+5e}{6} = e.$$

1.88. The algebra might be a bit of a stretch for some students:

$$\begin{aligned}
& (x_1-\bar{x}) + \quad (x_2-\bar{x}) + \quad (x_3-\bar{x}) + \cdots + (x_{n-1}-\bar{x}) + \quad (x_n-\bar{x}) \\
= \quad & x_1-\bar{x} + \quad x_2-\bar{x} + \quad x_3-\bar{x} + \cdots + x_{n-1}-\bar{x} + \quad x_n-\bar{x} \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \textit{(drop all the parentheses)} \\
= \quad & x_1+x_2+x_3+\cdots+x_{n-1}+x_n \qquad -\bar{x}-\bar{x}-\bar{x}-\cdots-\bar{x}-\bar{x} \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \textit{(rearrange the terms)} \\
= \quad & x_1+x_2+x_3+\cdots+x_{n-1}+x_n \qquad\qquad\qquad -n\cdot\bar{x}
\end{aligned}$$

Next, simply observe that $n \cdot \bar{x} = x_1 + x_2 + x_3 + \cdots + x_{n-1} + x_n$.

1.89. **(a)** One possible answer is 1, 1, 1, 1. **(b)** 0, 0, 20, 20. **(c)** For (a), any set of four identical numbers will have $s = 0$. For (b), the answer is unique; here is a rough description of why. We want to maximize the "spread-out"-ness of the numbers (which is what standard deviation measures), so 0 and 20 seem to be reasonable choices based on that idea. We also want to make each individual squared deviation—$(x_1 - \bar{x})^2$, $(x_2 - \bar{x})^2$, $(x_3 - \bar{x})^2$, and $(x_4 - \bar{x})^2$—as large as possible. If we choose 0, 20, 20, 20—or 20, 0, 0, 0—we make the first squared deviation 15^2, but the other three are only 5^2. Our best choice is two at each extreme, which makes all four squared deviations equal to 10^2.

1.90. Answers will vary. Typical calculators will carry only about 12 to 15 digits; for example, a TI-83 fails (gives $s = 0$) for 14-digit numbers. *Excel* (at least the version I checked) also fails for 14-digit numbers, but it gives $s = 262{,}144$ rather than 0. The version of Minitab used to prepare these answers fails at 20,000,001 (eight digits), giving $s = 2$.

1.91. See Exercise 1.42 for the stemplot, which shows the expected right skew. The five-number summary is a good choice: Min = 43, $Q_1 = 82.5$, $M = 102.5$, $Q_3 = 151.5$, Max = 598 days. Half the guinea pigs lived less than 102.5 days; typical lifetimes were 82.5 to 151.5 days. The longest-lived guinea pig died just short of 600 days, while one guinea pig lived only 43 days.

1.92. Convert from kilograms to pounds by multiplying by 2.2: $\bar{x} = (2.39 \text{ kg})(2.2 \text{ lb/kg}) = 5.26 \text{ lb}$ and $s = (1.14 \text{ kg})(2.2 \text{ lb/kg}) = 2.51 \text{ lb}$.

1.93. The table on the right reproduces the means and standard deviations from the solution to Exercise 1.79 and shows those values expressed in inches. For each conversion, multiply by $39.37/1000 = 0.03937$ (or divide by 25.4—an inch is defined as 25.4 millimeters). For example, for the *bihai* variety, $\bar{x} = (47.5975 \text{ mm})(0.03937 \text{ in/mm}) = (47.5975 \text{ mm}) \div (25.4 \text{ mm/in}) = 1.874 \text{ in}$.

	(in mm)		(in inches)	
Variety	$\bar{x}$	s	$\bar{x}$	s
bihai	47.5975	1.2129	1.874	0.04775
red	39.7113	1.7988	1.563	0.07082
yellow	36.1800	0.9753	1.424	0.03840

1.94. (a) $\bar{x} = 5.4479$ and $s = 0.2209$. **(b)** The first measurement corresponds to $5.50 \times 62.43 = 343.365$ pounds per cubic foot. To find $\bar{x}_{\text{new}}$ and s_{new}, we similarly multiply by 62.43: $\bar{x}_{\text{new}} \doteq 340.11$ and $s_{\text{new}} \doteq 13.79$.

Note: *The conversion from cm to feet is included in the multiplication by 62.43; the step-by-step process of this conversion looks like this:*

$$(1 \ g/cm^3)(0.001 \ kg/g)(2.2046 \ lb/kg)(30.48^3 \ cm^3/ft^3) = 62.43 \ lb/ft^3$$

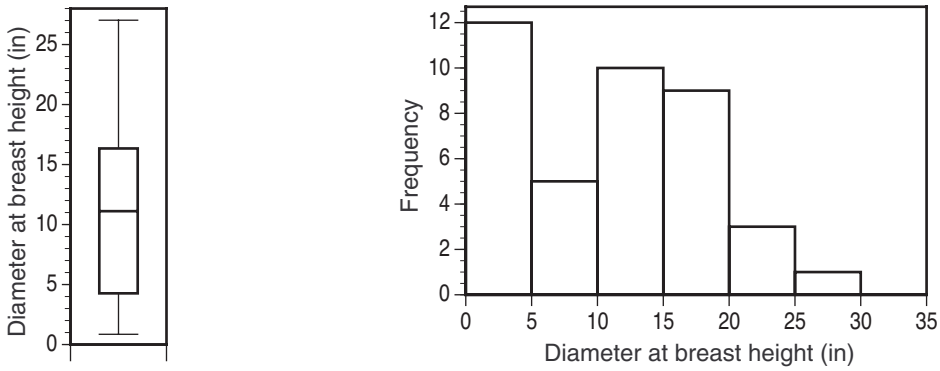
1.95. Multiplying 72 by 0.2, 0.4, 0.6, and 0.8, we find that the quintiles are located at positions 14.4, 28.8, 43.2, and 57.6. The 14th, 29th, 43rd, and 58th numbers in the list are 80, 99, 114, and 178 days.

1.96. Variance is changed by a factor of $2.54^2 = 6.4516$; generally, for a transformation $x_{\text{new}} = a + bx$, the new variance is b^2 times the old variance.

1.97. There are 72 survival times, so to find the 10% trimmed mean, remove the highest and lowest seven values (leaving 58). Remove the highest and lowest 14 values (leaving 44) for the 20% trimmed mean.

The mean and median for the full data set are $\bar{x} = 141.8$ and $M = 102.5$. The 10% trimmed mean is $\bar{x}^* = 118.16$, and the 20% trimmed mean is $\bar{x}^{**} = 111.68$. Since the distribution is right-skewed, removing the extremes lowers the mean.

1.98. After changing the scale from centimeters to inches, the five-number summary values change by the same ratio (that is, they are multiplied by 0.39). The shape of the histogram might change slightly because of the change in class intervals. **(a)** The five-number summary (in inches) is Min $= 0.858$, $Q_1 = 4.2705$, $M = 11.115$, $Q_3 = 16.341$, Max $=$ 27.027. **(b)** & **(c)** The boxplot and histogram are shown below. (Students might choose different interval widths for the histogram.) **(d)** As in Exercise 1.56, the histogram reveals more detail about the shape of the distribution.



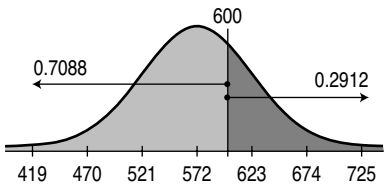
1.99. Take the mean plus or minus two standard deviations: $572 \pm 2(51) = 470$ to 674.

1.100. Take the mean plus or minus three standard deviations: $572 \pm 3(51) = 419$ to 725.

1.101. The z-score is $z = \frac{600 - 572}{51} \doteq 0.55$.

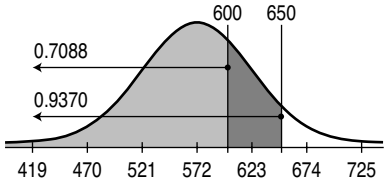
1.102. The z-score is $z = \frac{500 - 572}{51} \doteq -1.41$. This is negative because an ISTEP score of 500 is below average; specifically, it is 1.41 standard deviations below the mean.

1.103. Using Table A, the proportion below 600 ($z \doteq 0.55$) is 0.7088 and the proportion at or above is 0.2912; these two proportions add to 1. The graph on the right illustrates this with a single curve; it conveys essentially the same idea as the "graphical subtraction" picture shown in Example 1.27.



1.104. Using Table A, the proportion below 600 ($z \doteq 0.55$) is 0.7088, and the proportion below 650 ($z \doteq 1.53$) is 0.9370. Therefore:

$$\begin{array}{ccccc} \text{area between} \\ \text{600 and 650} & = & \text{area left} \\ & & \text{of 650} & - & \text{area left} \\ & & & & \text{of 600} \\ 0.2282 & = & 0.9370 & - & 0.7088 \end{array}$$



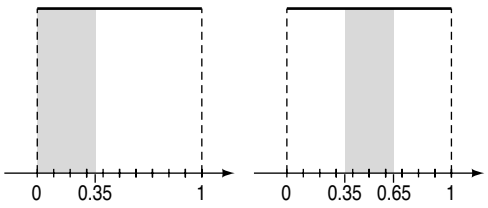
The graph on the right illustrates this with a single curve; it conveys essentially the same idea as the "graphical subtraction" picture shown in Example 1.28.

1.105. Using Table A, this ISTEP score should correspond to a standard score of $z = 1.645$, so the ISTEP score (unstandardized) is $572 + 1.645(51) \doteq 655.9$. If students use $z = 1.64$ or $z = 1.65$ instead of 1.645, the ISTEP score is about 655.6 or 656.2.

1.106. Using Table A, x should correspond to a standard score of $z \doteq 0.25$ (software gives 0.2533), so the ISTEP score (unstandardized) is about $572 + 0.25(51) \doteq 584.8$ (software: 584.9).

1.107. Sketches will vary.

1.108. **(a)** The curve forms a 1×1 square, which has area 1.
(b) $P(X < 0.35) = 0.35$.
(c) $P(0.35 < X < 0.65) = 0.3$.



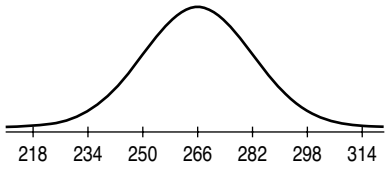
1.109. **(a)** The height should be $\frac{1}{4}$ since the area under the curve must be 1. The density curve is at right. **(b)** $P(X \leq 1) = \frac{1}{4} = 0.25$. **(c)** $P(0.5 < X < 2.5) = 0.5$.



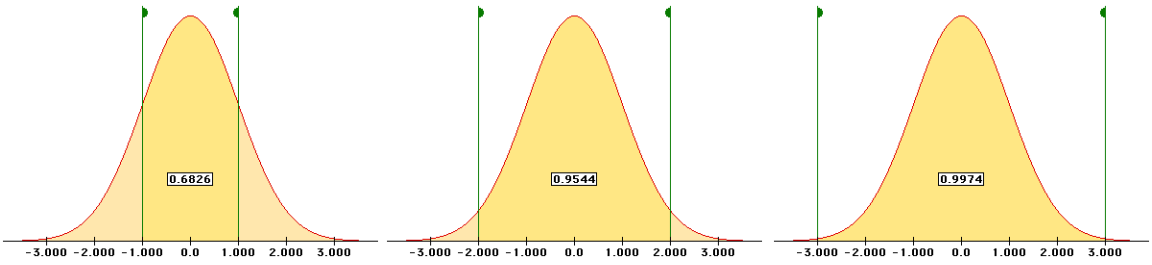
1.110. The mean and median both equal 0.5; the quartiles are $Q_1 = 0.25$ and $Q_3 = 0.75$.

1.111. **(a)** Mean is C, median is B (the right skew pulls the mean to the right). **(b)** Mean A, median A. **(c)** Mean A, median B (the left skew pulls the mean to the left).

1.112. Hint: It is best to draw the curve first, then place the numbers below it. Students may at first make mistakes like drawing a half-circle instead of the correct "bell-shaped" curve, or being careless about locating the standard deviation.

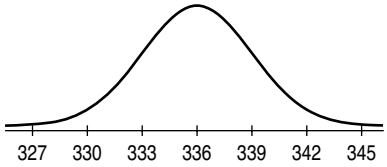


1.113. **(a)** The applet shows an area of 0.6826 between -1.000 and 1.000, while the 68–95–99.7 rule rounds this to 0.68. **(b)** Between -2.000 and 2.000, the applet reports 0.9544 (compared to the rounded 0.95 from the 68–95–99.7 rule). Between -3.000 and 3.000, the applet reports 0.9974 (compared to the rounded 0.997).



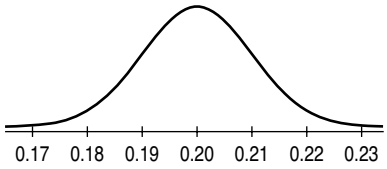
1.114. See the sketch of the curve in the solution to Exercise 1.112. **(a)** The middle 95% fall within two standard deviations of the mean: $266 \pm 2(16)$, or 234 to 298 days. **(b)** The shortest 2.5% of pregnancies are shorter than 234 days (more than two standard deviations below the mean).

1.115. **(a)** 99.7% of horse pregnancies fall within three standard deviations of the mean: $336 \pm 3(3)$, or 327 to 325 days. **(b)** About 16% are longer than 339 days since 339 days or more corresponds to at least one standard deviation above the mean.



Note: *This exercise did not ask for a sketch of the Normal curve, but students should be encouraged to make such sketches anyway.*

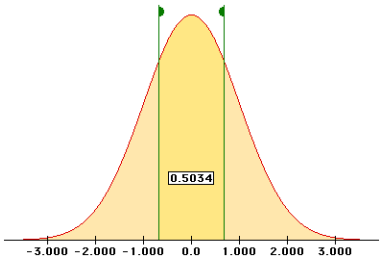
1.116. **(a)** About 50% of samples give values above the mean (0.20). Since 0.22 is two standard deviations above the mean, about 2.5% of sample values are above 0.22. **(b)** 0.18 to 0.22—that is, $0.2 \pm 2(0.01)$.



Note: *As the text models, it is probably best to use decimals for these proportions rather than percentages (0.22 instead of 22%) to lessen the confusion with, for example, 95%.*

1.117. The z-scores are $z_w = \frac{72-64}{2.7} \doteq 2.96$ for women and $z_m = \frac{72-69.3}{2.8} \doteq 0.96$ for men. The z-scores tell us that 6 feet is quite tall for a woman, but not at all extraordinary for a man.

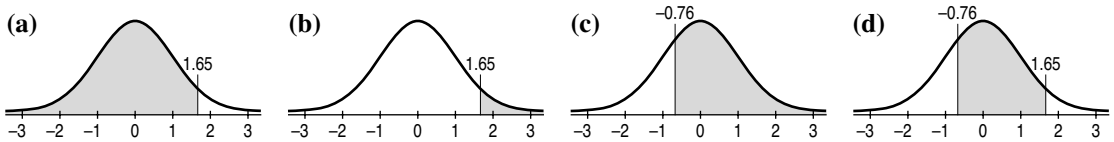
1.118. Because the quartiles of any distribution have 50% of observations between them, we seek to place the flags so that the reported area is 0.5. The closest the applet gets is an area of 0.5034, between -0.680 and 0.680. Thus, the quartiles of any Normal distribution are about 0.68 standard deviations above and below the mean.



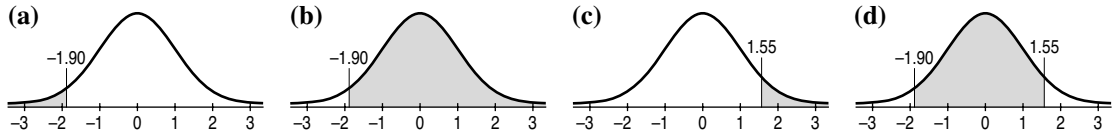
Note: *Table A places the quartiles at about ±0.67; other statistical software gives ±0.6745.*

1.119. The mean and standard deviation are $\bar{x} = 5.4256$ and $s = 0.5379$. About 67.62% ($71/105 \doteq 0.6476$) of the pH measurements are in the range $\bar{x} \pm s = 4.89$ to 5.96. About 95.24% (100/105) are in the range $\bar{x} \pm 2s = 4.35$ to 6.50. All (100%) are in the range $\bar{x} \pm 3s = 3.81$ to 7.04.

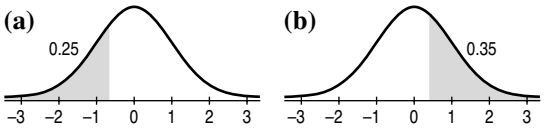
1.120. **(a)** $Z < 1.65$: 0.9505. **(b)** $Z > 1.65$: 0.0495. **(c)** $Z > -0.76$: 0.7764. **(d)** $-0.76 < Z < 1.65$: $0.9505 - 0.2236 = 0.7269$.



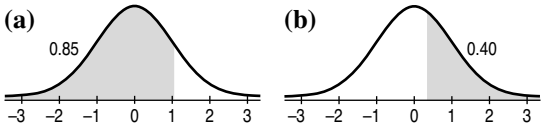
1.121. **(a)** $Z \leq -1.9$: 0.0287. **(b)** $Z \geq -1.9$: 0.9713. **(c)** $Z > 1.55$: 0.0606. **(d)** $-1.9 < Z < 1.55$: $0.9394 - 0.0287 = 0.9107$.



1.122. **(a)** 25% of the observations fall below −0.6745. (This is the 25th percentile of the standard Normal distribution). **(b)** 35% of the observations fall above 0.3853 (the 65th percentile of the standard Normal distribution).



1.123. **(a)** $z = 1.0364$ has cumulative proportion 0.85 (that is, 1.0364 is the 85th percentile of the standard Normal distribution). **(b)** If $z = 0.2533$, then $Z > z$ has proportion 0.40 (0.2533 is the 60th percentile).



1.124. 70 is two standard deviations below the mean (that is, it has standard score $z = -2$), so about 2.5% (half of the outer 5%) of adults would have WAIS scores below 70.

1.125. 130 is two standard deviations above the mean (that is, it has standard score $z = 2$), so about 2.5% of adults would score at least 130.

1.126. Tonya's score standardizes to $z = \frac{1320 - 1026}{209} \doteq 1.4067$, while Jermaine's score corresponds to $z = \frac{28 - 20.8}{4.8} = 1.5$. Jermaine's score is higher.

1.127. Jacob's score standardizes to $z = \frac{17 - 20.8}{4.8} \doteq -0.7917$, while Emily's score corresponds to $z = \frac{680 - 1026}{209} \doteq -1.6555$. Jacob's score is higher.

1.128. Jose's score standardizes to $z = \frac{1380 - 1026}{209} \doteq 1.6938$, so an equivalent ACT score is $20.8 + 1.6938 \times 4.8 \doteq 28.9$. (Of course, ACT scores are reported as whole numbers, so this would presumably be a score of 29.)

1.129. Maria's score standardizes to $z = \frac{29 - 20.8}{4.8} \doteq 1.7083$, so an equivalent SAT score is $1026 + 1.7083 \times 209 \doteq 1383$.

1.130. Tonya's score standardizes to $z = \frac{1320 - 1026}{209} \doteq 1.4067$; this is the 92nd percentile.

1.131. Jacob's score standardizes to $z = \frac{17 - 20.8}{4.8} \doteq -0.7917$; this is about the 21st percentile.

1.132. 1294 and above: The top 10% corresponds to a standard score of $z = 1.2816$, which in turn corresponds to a score of $1026 + 1.2816 \times 209 \doteq 1294$ on the SAT.

1.133. 850 and below: The bottom 20% corresponds to a standard score of $z = -0.8416$, which in turn corresponds to a score of $1026 - 0.8416 \times 209 \doteq 850$ on the SAT.

1.134. The quartiles of a Normal distribution are ± 0.6745 standard deviations from the mean, so for ACT scores, they are $20.8 \pm 0.6745 \times 4.8 = 17.6$ to 24.0.

1.135. The quintiles of the SAT score distribution are $1026 - 0.8416 \times 209 = 850$, $1026 - 0.2533 \times 209 = 973$, $1026 + 0.2533 \times 209 = 1079$, and $1026 + 0.8416 \times 209 = 1202$.

1.136. **(a)** 240 mg/dl standardizes to $z = \frac{240 - 185}{39} \doteq 1.41$, which has cumulative probability 0.9207, so about 8% of young women have levels over 240 mg/dl. **(b)** 200 mg/dl standardizes to $z = \frac{200 - 185}{39} \doteq 0.385$, which has cumulative probability 0.6499, so about 27% of young women have levels between 200 and 240 mg/dl.

1.137. 200 and 240 mg/dl standardize to $z = \frac{200 - 222}{37} \doteq -0.5946$ (cumulative probability 0.2761) and $z = \frac{240 - 222}{37} \doteq 0.4865$ (cumulative probability 0.6867). Therefore, about 31% of middle-aged men have levels over 240 mg/dl, and about 41% have levels between 200 and 240 mg/dl.

1.138. **(a)** About 0.6% of healthy young adults have osteoporosis (the cumulative probability below a standard score of -2.5 is 0.0062). **(b)** About 31% of this population of older women has osteoporosis: The BMD level which is 2.5 standard deviations below the young adult mean would standardize to -0.5 for these older women, and the cumulative probability for this standard score is 0.3085.

1.139. **(a)** About 5.2%: $x < 240$ corresponds to $z < -1.625$. Table A gives 5.16% for -1.63 and 5.26% for -1.62. Software (or averaging the two table values) gives 5.21%. **(b)** About 54.7%: $240 < x < 270$ corresponds to $-1.625 < z < 0.25$. The area to the left of 0.25 is 0.5987; subtracting the answer from part (a) leaves about 54.7%. **(c)** About 279 days or longer: Searching Table A for 0.80 leads to $z > 0.84$, which corresponds to $x > 266 + 0.84(16) = 279.44$. (Using the software value $z > 0.8416$ gives $x > 279.47$.)

1.140. **(a)** The quartiles for a standard Normal distribution are ± 0.6745. **(b)** For a $N(\mu, \sigma)$ distribution, $Q_1 = \mu - 0.6745\sigma$ and $Q_3 = \mu + 0.6745\sigma$. **(c)** For human pregnancies, $Q_1 = 266 - 0.6745 \times 16 \doteq 255.2$ and $Q_3 = 266 + 0.67455 \times 16 \doteq 276.8$ days.

1.141. **(a)** As the quartiles for a standard Normal distribution are ± 0.6745, we have $IQR = 1.3490$. **(b)** $c = 1.3490$: For a $N(\mu, \sigma)$ distribution, the quartiles are $Q_1 = \mu - 0.6745\sigma$ and $Q_3 = \mu + 0.6745\sigma$.

1.142. In the previous two exercises, we found that for a $N(\mu, \sigma)$ distribution, $Q_1 = \mu - 0.6745\sigma$, $Q_3 = \mu + 0.6745\sigma$, and $IQR = 1.3490\sigma$. Therefore, $1.5 \times IQR = 2.0235\sigma$, and the suspected outliers are below $Q_1 - 1.5 \times IQR = \mu - 2.698\sigma$, and above $Q_3 + 1.5 \times IQR = \mu + 2.698\sigma$. The percentage outside of this range is $2 \times 0.0035 = 0.70\%$.

1.143. The plot is nearly linear. Because heart rate is measured in whole numbers, there is a slight "step" appearance to the graph.

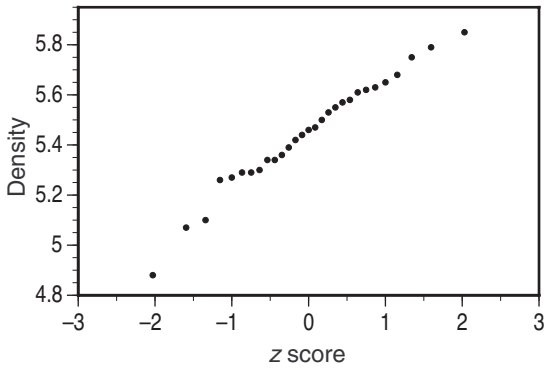
1.144. The shape of the quantile plot suggests that the data are right-skewed (as was observed in Exercises 1.24 and 1.44). This can be seen in the flat section in the lower left—these numbers were less spread out than they should be for Normal data—and the three apparent outliers (the United States, Canada, and Australia) that deviate from the line in the upper right; these were much larger than they would be for a Normal distribution.

1.145. The plot is reasonably close to a line, apart from the stair-step appearance, presumably due to limited accuracy of the measuring instrument.

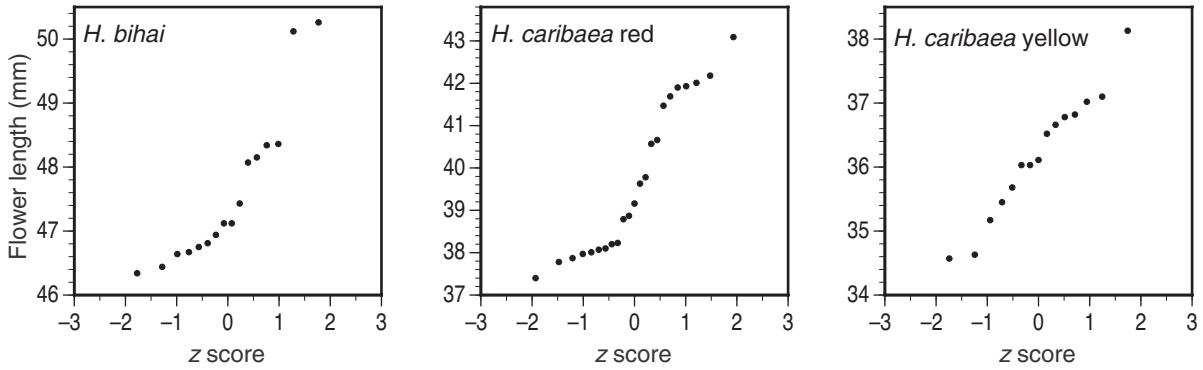
1.146. **(a)** is the graph of (3) the highway gas mileages: Aside from the Insight, these numbers are reasonably Normal, and in this graph, the points fall close to a line aside from one high outlier. **(b)** is the graph of (1) the IQ data: This distribution was the most Normal of the four, and this graph is almost a perfect line. **(c)** is the graph of (4) the call length data: The stemplot is right-skewed, with several high outliers (the outliers were not shown in the stemplot; rather they were listed after the plot). The skewness is visible in the flat section of this graph. **(d)** is the graph of (2) the tuition and fees data: The histogram showed three clusters, which are visible in the graph. The low and high clusters had peaks at their extremes; these show up in the flat sections in the lower left and upper right of the graph.

Note: *Matching (a) and (c) is probably the most difficult decision. Aside from the reasons given above, students might also observe that graph (a) shows considerably fewer points than (c), which is consistent with the 21 two-seater cars in data set (3) versus the 80 call lengths for (4).*

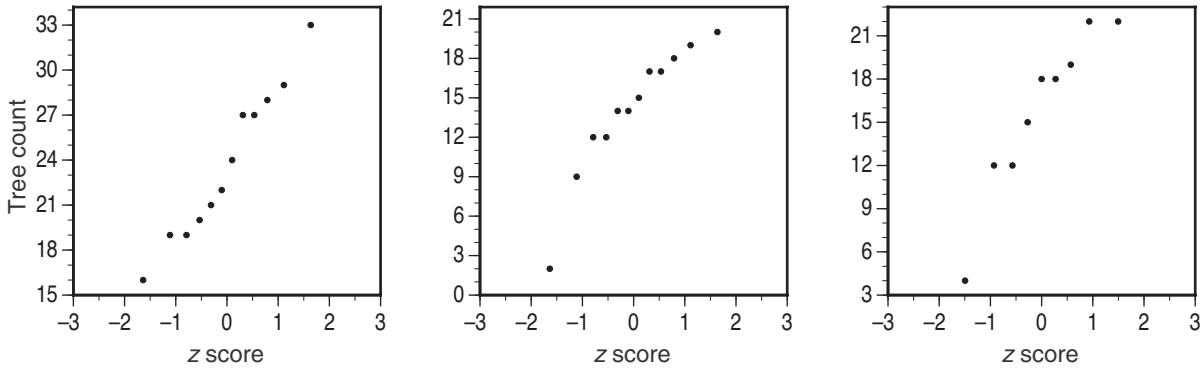
1.147. See also the solution to Exercise 1.40. The plot suggests no major deviations from Normality, although the three lowest measurements do not quite fall in line with the other points.



1.148. **(a)** All three quantile plots are below; the yellow variety is the nearest to a straight line. **(b)** The other two distributions are both slightly right-skewed (the lower-left portion of the graph is somewhat flat); additionally, the *bihai* variety appears to have a couple of high outliers.

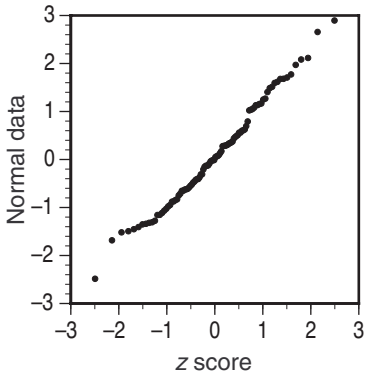


1.149. See also the solution to Exercise 1.80. The first plot (for never-logged areas) is nearly linear. The other two each show a low value, perhaps suggesting a slight skew to the left.



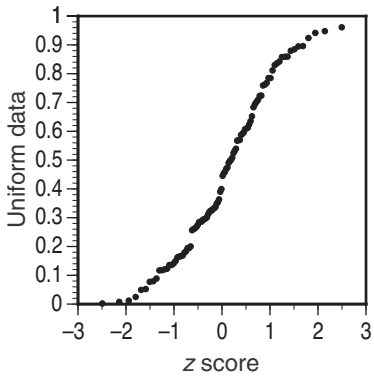
1.150. A stemplot from one sample is shown. Histograms will vary slightly but should suggest a bell curve. The Normal quantile plot shows something fairly close to a line but illustrates that, even for actual Normal data, the tails may deviate slightly from a line.

−2	4
−1	65
−1	4443333211100
−0	9988887766666555
−0	444443332111100000
0	000011222233333444
0	55556667
1	000011112244
1	55666779
2	01
2	68



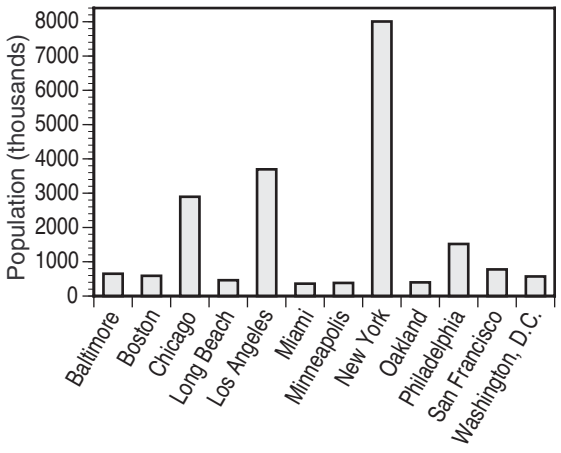
1.151. A stemplot from one sample is shown. Histograms will vary slightly but should suggest the density curve of Figure 1.35 (but with more variation than students might expect). The Normal quantile plot shows that, compared to a Normal distribution, the uniform distribution does not extend as low or as high (not surprising, since all observations are between 0 and 1).

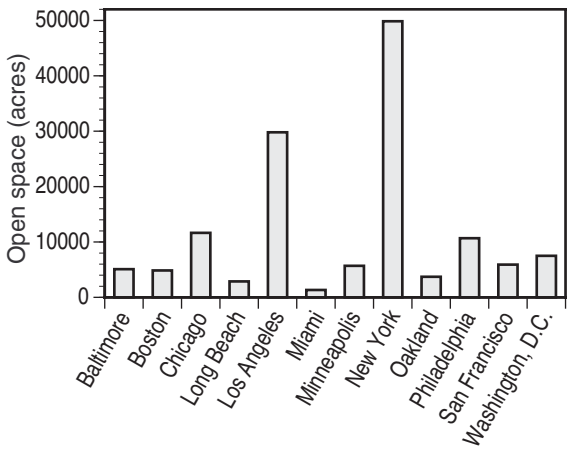
0	001245778
1	11223344666678999
2	56667888999
3	0112233345699
4	4556799
5	00233666899
6	00123589
7	002256688
8	13345557899
9	2446

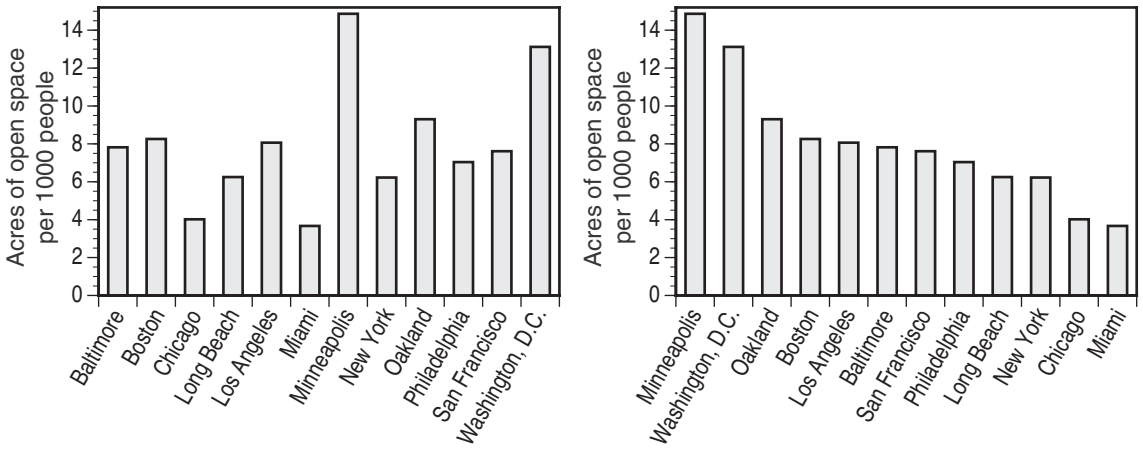


1.152. **(a)** & **(b)** These graphs are shown below and on the next page. Bars are shown in alphabetical order by city name (as the data were given in the table). **(c)** For Baltimore, for example, this rate is $\frac{5091}{651} \doteq 7.82$. The complete table is shown on the right. **(d)** & **(e)** Both of these graphs are shown below. Note that the text does not specify whether the bars should be ordered by *increasing* or *decreasing* rate. **(f)** Preferences may vary, but the ordered bars make comparisons easier.

Baltimore	7.82
Boston	8.26
Chicago	4.02
Long Beach	6.25
Los Angeles	8.07
Miami	3.67
Minneapolis	14.87
New York	6.23
Oakland	9.30
Philadelphia	7.04
San Francisco	7.61
Washington, D.C.	13.12







1.153. The given description is true on the average, but the curves (and a few calculations) give a more complete picture. For example, a score of about 675 is about the 97.5th percentile for both genders, so the top boys and girls have very similar scores.

1.154. (a) & **(b)** Answers will vary. Definitions might be as simple as "free time," or "time spent doing something other than studying." For (b), it might be good to encourage students to discuss practical difficulties; for example, if we ask Sally to keep a log of her activities, the time she spends filling it out presumably reduces her available "leisure time."

1.155. Shown is a stemplot; a histogram should look similar to this. This distribution is relatively symmetric apart from one high outlier. Because of the outlier, the five-number summary (all in hours) is preferred:

22 23.735 24.31 24.845 28.55

Alternatively, the mean and standard deviation are $\bar{x} = 24.339$ and $s = 0.9239$ hours.

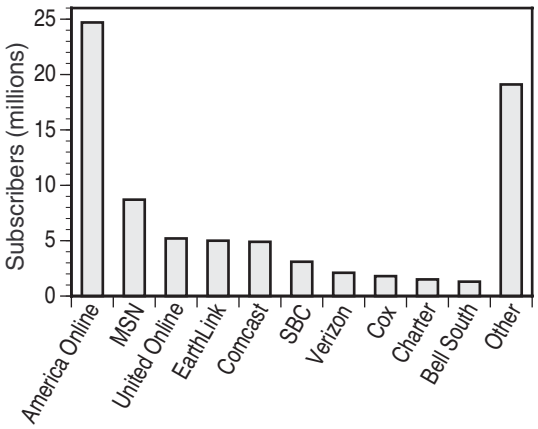
22	013
22	7899
23	000011222233344444
23	55566666667777778888888999
24	000000111111222222222333333333444444
24	55555566666666777777888888999999
25	00001111233344
25	56666889
26	2
26	56
27	2
27	
28	
28	5

1.156. Gender and automobile preference are categorical; age and household income are quantitative.

1.157. Many—but fewer than half—of these students were 19.

Note: *In fact, there had to be* at least *nine students who were 19, and no more than 111—the largest number only if the next youngest student was 43. If you have some particularly bright students, you might challenge them to prove this.*

1.158. Either a bar graph or a pie chart could be used. The given numbers sum to 58.3, so the "Other" category presumably includes the remaining 19.1 million subscribers.

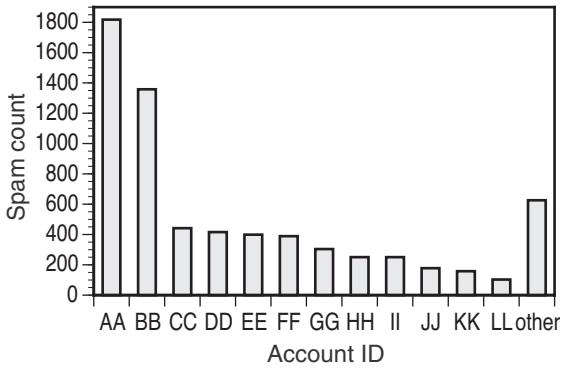


1.159. Women's weights are skewed to the right: This makes the mean higher than the median, and it is also revealed in the differences $M - Q_1 = 14.9$ lb and $Q_3 - M = 24.1$ lb.

1.160. **(a)** For car makes (a categorical variable), use either a bar graph or pie chart. For car age (a quantitative variable), use a histogram, stemplot, or boxplot. **(b)** Study time is quantitative, so use a histogram, stemplot, or boxplot. To show change over time, use a time plot (average hours studied against time). **(c)** Use a bar graph or pie chart to show radio station preferences. **(d)** Use a Normal quantile plot to see whether the measurements follow a Normal distribution.

1.161. **(a)** About 20% of low-income and 33% of high-income households consisted of two people. **(b)** The majority of low-income households, but only about 7% of high-income households, consist of one person. One-person households often have less income because they would include many young people who have no job or have only recently started working. (Income generally increases with age.)

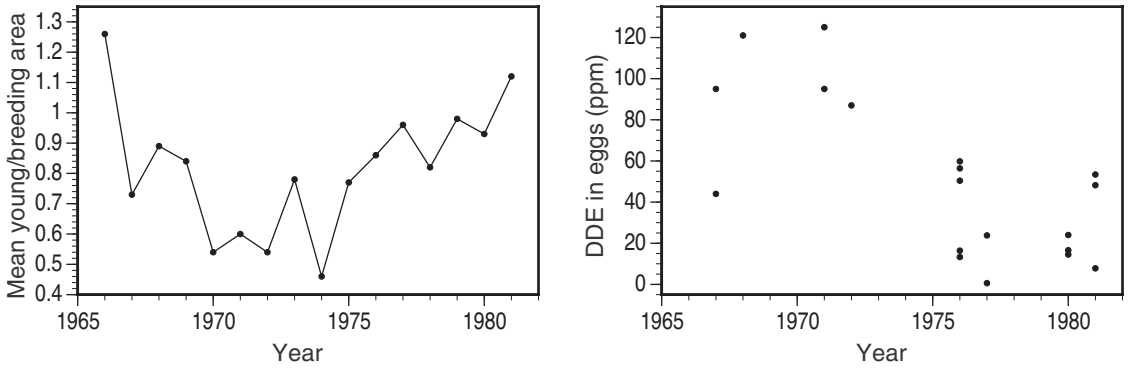
1.162. The counts given add to 6067, so the others received 626 spam messages. Either a bar graph or a pie chart would be appropriate. What students learn from this graph will vary; one observation might be that AA and BB (and perhaps some others) might need some advice on how to reduce the amount of spam they receive.



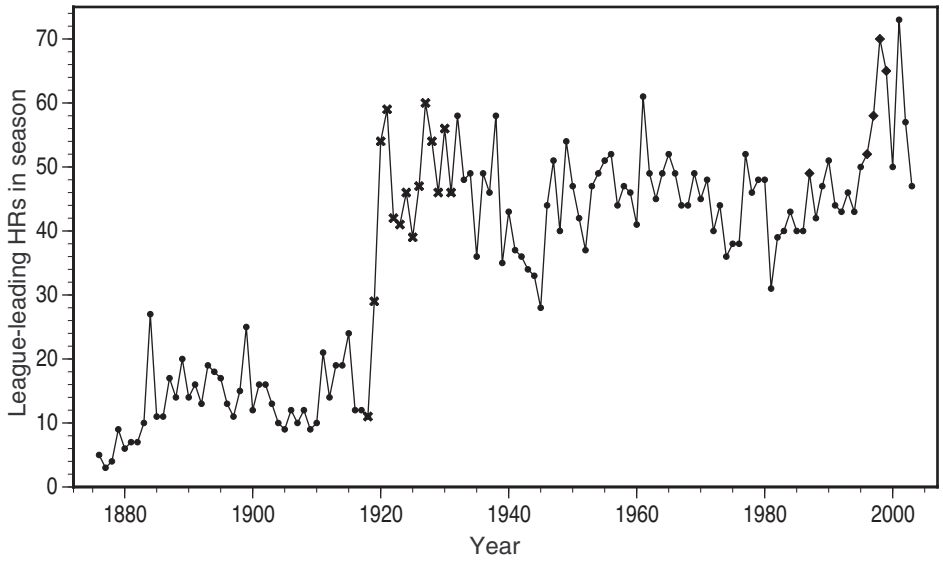
1.163. No, and no: It is easy to imagine examples of many different data sets with mean 0 and standard deviation 1—for example, $\{-1,0,1\}$ and $\{-2,0,0,0,0,0,0,0,2\}$.

Likewise, for any given five numbers $a \le b \le c \le d \le e$ (not all the same), we can create many data sets with that five-number summary, simply by taking those five numbers and adding some additional numbers in between them, for example (in increasing order): 10, __, 20, __, __, 30, __, __, 40, __, 50. As long as the number in the first blank is between 10 and 20, and so on, the five-number summary will be 10, 20, 30, 40, 50.

1.164. In the first time plot, we see that numbers of eagle young begin to rise shortly after the ban in 1972. In the second time plot, the five highest DDE numbers occurred before 1972. (Note that the points in the second time plot have not been connected here; connecting the dots is confusing when there are multiple measurements in a year.)



1.165. The time plot is shown below; because of the great detail in this plot, it is larger than other plots. Ruth's and McGwire's league-leading years are marked with different symbols. **(a)** During World War II (when many baseball players joined the military), the best home run numbers decline sharply and steadily. **(b)** Ruth seemed to set a new standard for other players; after his first league-leading year, he had 10 seasons much higher than anything that had come before, and home run production has remained near that same level ever since (even the worst post-Ruth year—1945—had more home runs than the best pre-Ruth season). While some might argue that McGwire's numbers also raised the standard, the change is not nearly as striking, nor did McGwire maintain it for as long as Ruth did. (This is not necessarily a criticism of McGwire; it instead reflects that in baseball, as in many other endeavors, rates of improvement tend to decrease over time as we reach the limits of human ability.)



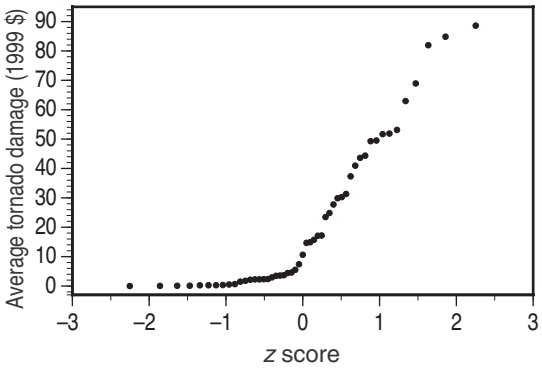
1.166. Bonds's mean changes from 36.56 to 34.41 home runs (a drop of 2.15), while his median changes from 35.5 to 34 home runs (a drop of 1.5). This illustrates that outliers affect the mean more than the median.

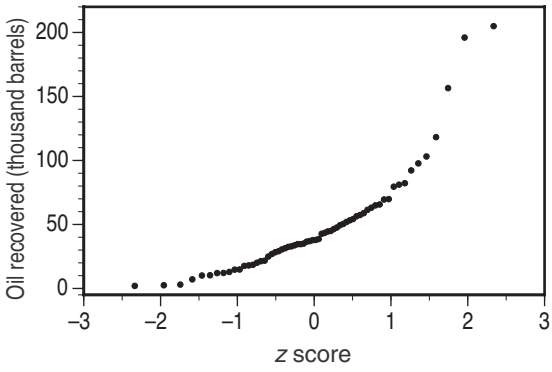
1	69
2	4
2	55
3	3344
3	77
4	02
4	5669
5	
5	
6	
6	
7	3

1.167. Recall the text's description of the effects of a linear transformation $x_{\text{new}} = a + bx$: The mean and standard deviation are each multiplied by b (technically, the standard deviation is multiplied by $|b|$, but this problem specifies that $b > 0$). Additionally, we add a to the (new) mean, but a does not affect the standard deviation. **(a)** The desired transformation is $x_{\text{new}} = -50 + 2x$; that is, $a = -50$ and $b = 2$. (We need $b = 2$ to double the standard deviation; as this also doubles the mean, we then subtract 50 to make the new mean 100.) **(b)** $x_{\text{new}} = -49.0909 + 1.8182x$; that is, $a = -49\frac{1}{11} \doteq -49.0909$ and $b = \frac{20}{11} \doteq 1.8182$. (This choice of b makes the new standard deviation 20 and the new mean $149\frac{1}{11}$; we then subtract 49.0909 to make the new mean 100.) **(c)** David's score—$2 \cdot 78 - 50 = 106$—is higher within his class than Nancy's score—$1.8182 \cdot 78 - 49.0909 \doteq 92.7$—is within her class. **(d)** From (c), we know that a third-grade score of 78 corresponds to a score of 106 from the $N(100, 20)$ distribution, which has a standard score of $z = \frac{106-100}{20} = 0.3$. (Alternatively, $z = \frac{78-75}{10} = 0.3$.) A sixth-grade score of 78 has standard score $z = \frac{92.7-100}{20} = \frac{78-82}{11} \doteq -0.36$. Therefore, about 62% of third graders and 36% of sixth graders score below 78.

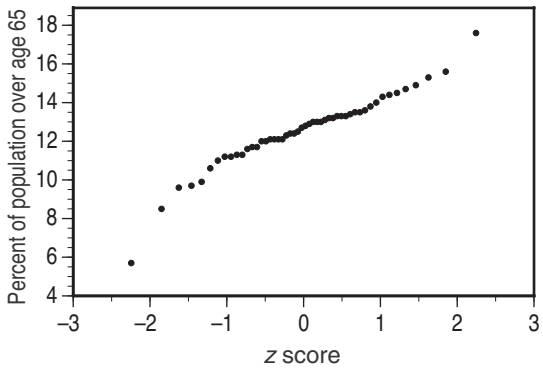
1.168. Shown below are both quantile plots. Skewness shows up in a quantile plot as a flat tail; for right-skewness, that flat portion is at the beginning (the lower left).

The tornado data shows no clear outliers (the highest points appear to fit reasonably well with the nearby points in the plot). The three highest oil-well numbers appear to be outliers. (Incidentally, the $1.5 \times IQR$ rule supports this conclusion.)

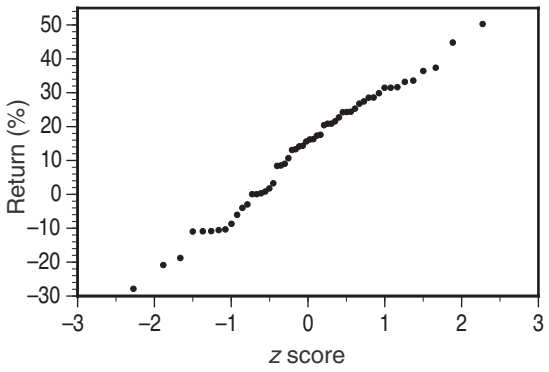




1.169. **(a)** Sketches may vary somewhat, but should be linear in the middle; the outliers would show up as a point in the lower left *below* the line (because low outliers are less than we expect them to be for a Normal distribution) and a point in the upper right *above* the line (because high outliers are greater than we expect them to be). **(b)** The quantile plot for this data agrees with the expectations noted in (a).



1.170. **(a)** Sketches should be linear in the middle. The heavy tails would show up flat sections in the lower left and upper right. The values in the tails are less spread out than we would expect for a Normal distribution, so the line is less steep for low and high data values. **(b)** The quantile plot for this data does not clearly suggest heavy tails. (This is consistent with the text's statement: "Average returns. . . over longer periods of time become more Normal.") There are no clear deviations from Normality.



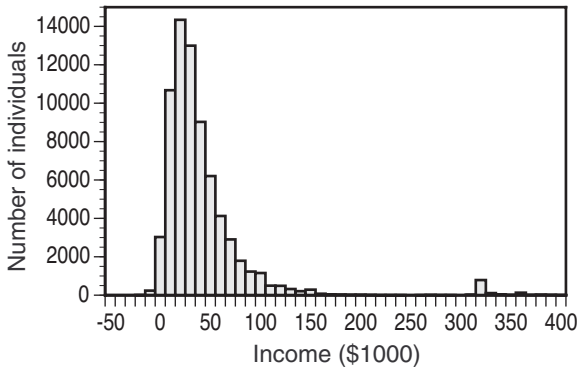
Note: *For an example of a quantile plot of a heavy-tailed distribution, see the tuition-and-fees data from Exercise 1.27; a quantile plot is shown in Figure 1.42(d), which accompanies Exercise 1.146.*

1.171. Results will vary. One set of 20 samples gave the results at the right (Normal quantile plots are not shown).

Means		Standard deviations	
18	589	3	8
19	00124	4	01
19	7789	4	22
20	1333	4	44455
20		4	66
21	223	4	9
21	5	5	000
		5	22
		5	45

Theoretically, $\bar{x}$ will have a $N(20, 1)$ distribution—so that about 99.7% of the time, one should find $\bar{x}$ between 17 and 23. Meanwhile, the theoretical distribution of s is nearly Normal (slightly skewed) with mean $\doteq 4.9482$ and standard deviation $\doteq 0.7178$; about 99.7% of the time, s will be between 2.795 and 7.102. Note that "on the average," s underestimates σ (that is, $4.9482 < 5$). Unlike the mean $\bar{x}$, s is not an unbiased estimator of σ; in fact, for a sample of size n, the mean of s/σ is $\frac{\sqrt{2}\,\Gamma(n/2)}{\sqrt{n-1}\,\Gamma(n/2-1/2)}$. (This factor approaches 1 as n approaches infinity.) The proof of this fact is left as an exercise—for the instructor, not for the average student!

1.172. Shown is a histogram with classes of width \$10,000, which omits the 67 individuals with incomes over \$410,000. A boxplot would also be an appropriate choice, although it would not show the cluster of individuals with incomes between \$300,000 and \$400,000.



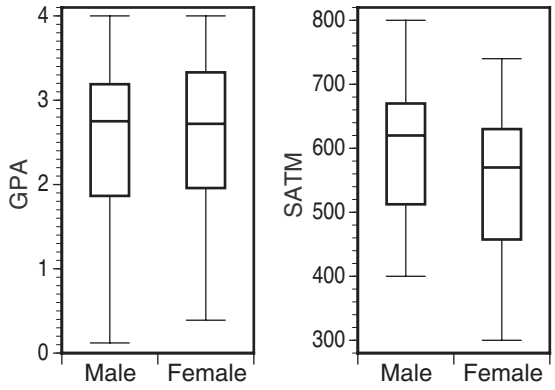
Because this distribution is skewed, the five-number summary is more appropriate than the mean:

$$\text{Min} = -\$23{,}980,\ Q_1 = \$22{,}000,\ M = \$35{,}000,\ Q_3 = \$53{,}000,\ \text{Max} = \$609{,}548$$

For reference, the mean is \$46,050 (larger than the median, as we would expect).

Note: *Processing this data file is no simple task; be sure that your students have adequate software. Some otherwise well-behaved software might choke on a data file as large as this. For example,* Excel *spreadsheets only allow 65,536 rows, so it would need to have this data set broken into at least two pieces.*

1.173. Men seem to have higher SATM scores than women; each number in the five-number summary is 40 to 60 points higher than the corresponding number for women. Women generally have higher GPAs than men, but the difference is less striking; in fact, the men's median is slightly higher.

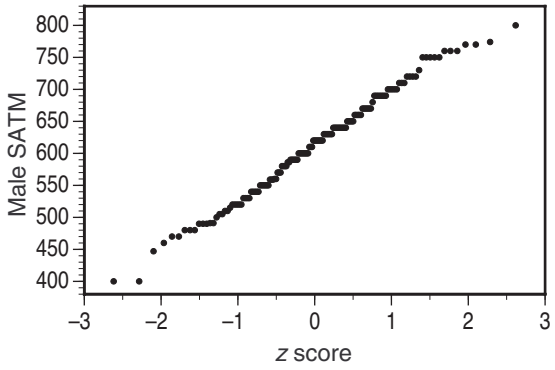


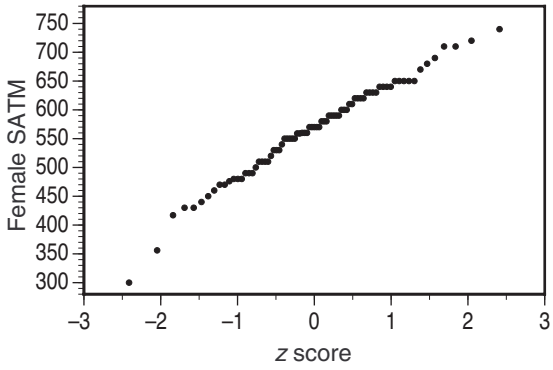
	Min	Q_1	M	Q_3	Max
Male GPA	0.12	2.135	2.75	3.19	4.00
Female GPA	0.39	2.250	2.72	3.33	4.00
Male SATM	400	550	620	670	800
Female SATM	300	510	570	630	740

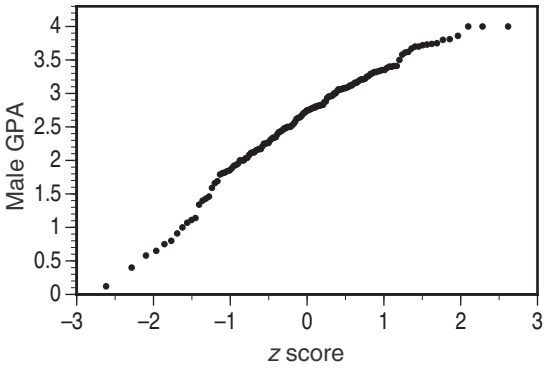
Quantile plots are shown on the next page. Judging from these (and from the $1.5 \times IQR$ criterion), student 183 is an outlier for female SATM (300). For male GPA, outliers are students 127 (GPA 0.12) and 90 (GPA 0.4), and for female GPA, the outlier is student 188 (GPA 0.39). (Judgments of these may vary if the $1.5 \times IQR$ criterion is not used.)

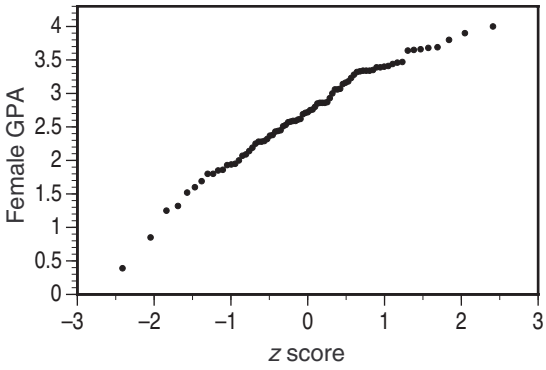
All four Normal quantile plots look fairly linear, so students might judge all four data sets to be Normal. However, both GPA sets—especially the male GPA—are somewhat left-skewed; there is some evidence of this in the long bottom tails of the GPA boxplots, as well as by the flatness in the upper right of their quantile plots.

Note: *In fact, statistical tests indicate that the male GPA numbers would not be likely to come from a Normal distribution, even with the outliers omitted.*


Male SATM
800
750
700
650
600
550
500
450
400
–3
–2
–1
0
1
2
3
z score


Female SATM
750
700
650
600
550
500
450
400
350
300
–3
–2
–1
0
1
2
3
z score


Male GPA
4
3.5
3
2.5
2
1.5
1
0.5
0
–3
–2
–1
0
1
2
3
z score


Female GPA
4
3.5
3
2.5
2
1.5
1
0.5
0
–3
–2
–1
0
1
2
3
z score