Chapter 1 Solutions

1.1. (a) The individuals are vehicles (or "cars"). **(b)** The variables are make/model (categorical), vehicle type (categorical), transmission type (categorical), number of cylinders (quantitative), city MPG (quantitative), and highway MPG (quantitative).

1.2. The individuals are students. The variables are name (categorical), major (categorical), points (quantitative), and grade (categorical).

Note: *One might observe that "name" is more a* label *than a* variable*: For most categorical variables, there is no problem with two individuals having the same value, but for student names, we would like each individual to have a unique name. (Of course, that might not always be the case.)*

1.3. (a) Type of wood is categorical. **(b)** Water repellent is categorical. **(c)** Paint thickness is quantitative. **(d)** Paint color is categorical. **(e)** Weathering time is quantitative.

1.4. Possible categorical variables include gender, year in school, race, and perhaps some classification of what the student watched (PBS? MTV? Sitcom? Documentary?). Quantitative variables might include hours watched (per day or week), hours spent studying (per day or week), hours spent sleeping, age (years), GPA.

1.5. Possible answers include time to run a race (instrument: stopwatch), heart rate after exercising (instrument: watch). Answers will depend on how broadly one defines fitness; most instruments will likely be watches or measuring tapes.

1.6. Student answers may vary; for comparison, recent *U.S. News* rankings have used measures such as 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.

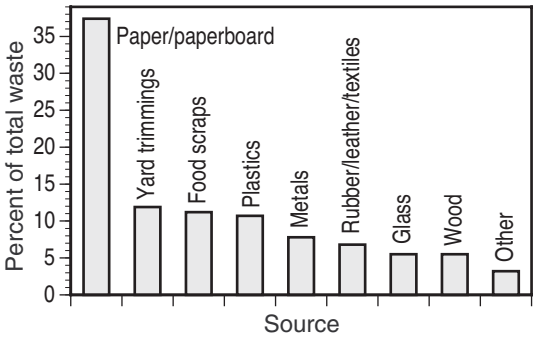
1.7. The rates (in deaths per 100 million miles) are

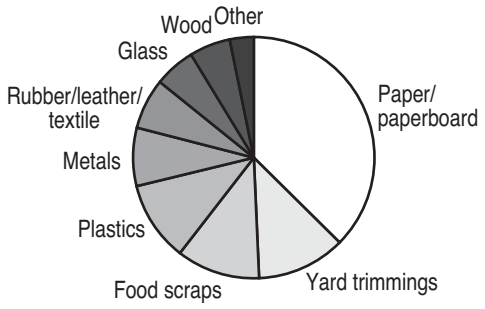
$$\frac{39,250 \text{ deaths}}{22,470 \text{ hundred million miles}} \doteq 1.75 \qquad \text{and} \qquad \frac{42,815 \text{ deaths}}{28,300 \text{ hundred million miles}} \doteq 1.51.$$

These numbers suggest that driving was safer in 2002 than in 1992.

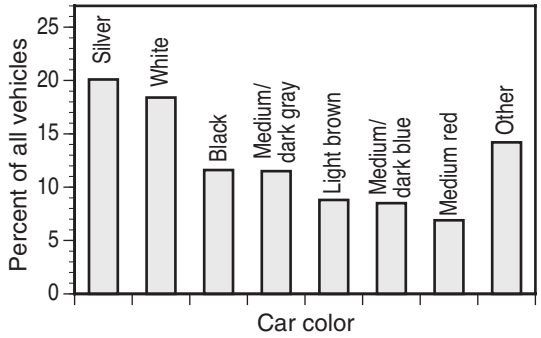
Note: *Most students will have difficulty with one (or both) of two things: expressing distance traveled in units of 100 million miles, and the order of the division (deaths in the numerator, distance in the denominator).*

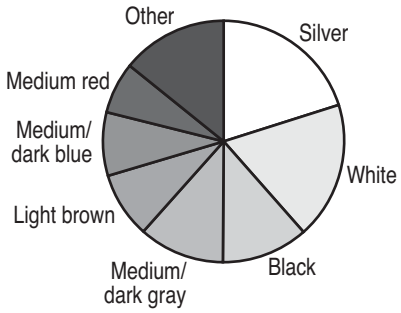
1.8. (a) The weights add to 231.8 million tons. **(b)** & **(c)** The bar and pie graphs are shown below.



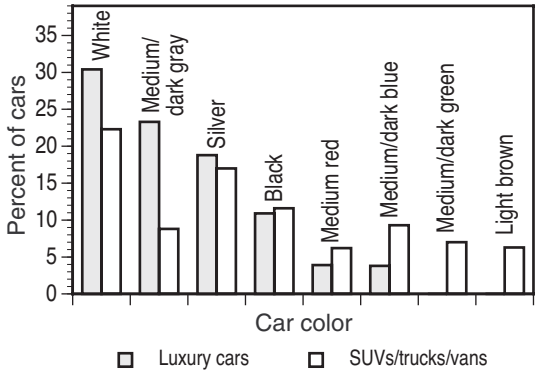


1.9. The given percents add to 85.8%, so the rest (14.2%) are "other colors." Because the numbers represent pieces of a single whole ("all cars"), a pie chart could be used.

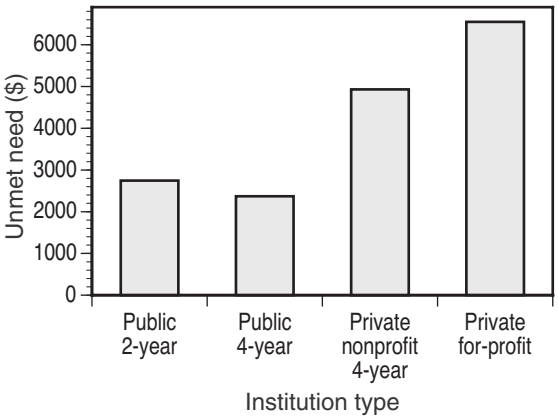




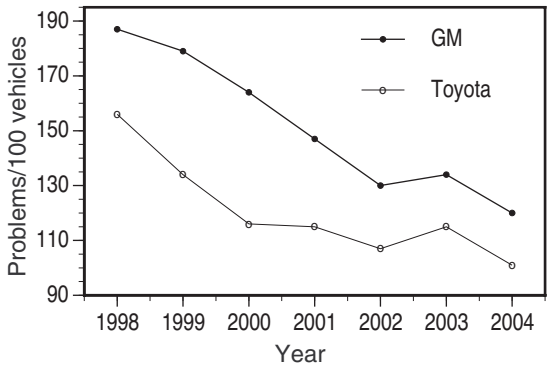
1.10. Shown is one way to create this bar graph, with side-by-side bars, sorted by decreasing luxury-car percentages.



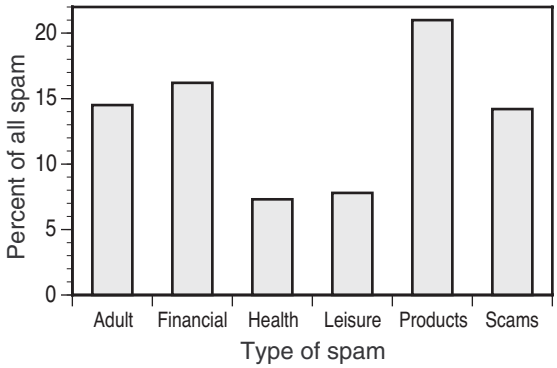
1.11. Unmet need is greatest at private institutions, especially for-profit ones. A pie chart would not be incorrect because these numbers do not represent parts of a single whole. (If the numbers given had been *total* unmet need, rather than *average* unmet need, and if we had information about *all* types of institutions, we would have been able to make a pie chart.)

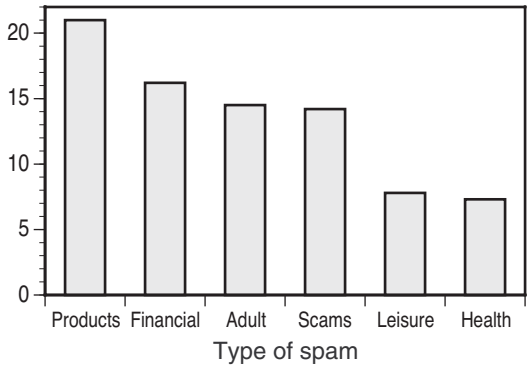


1.12. The time plots show that both manufacturers have generally improved over this period, with one slight jump in problems in 2003. Toyota vehicles typically have fewer problems, but GM has managed to close the gap slightly.

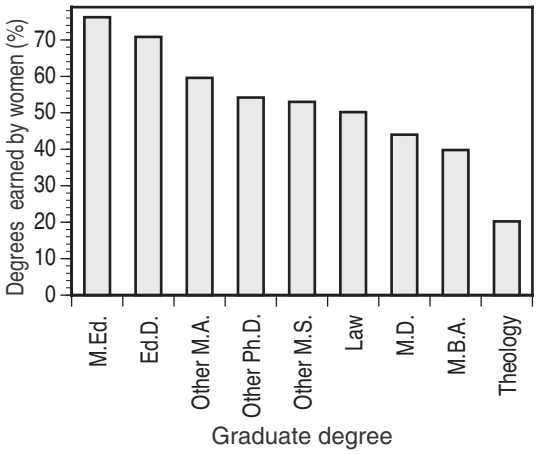


1.13. The two bar graphs are shown below.





1.14. **(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.



1.15. **(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%.

1.16. Shown on the right are the original stemplot (as given in the text for Exercise 1.15, 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

1.17. 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

1.18. 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).

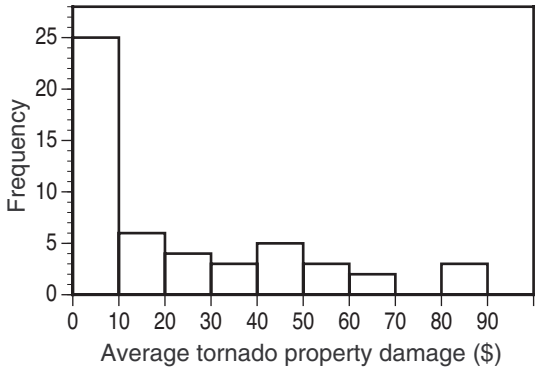
Individual		Class
	0	799
22	1	0134444
99866655	1	5577
22222	2	0
8	2	57
	3	
	3	5

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

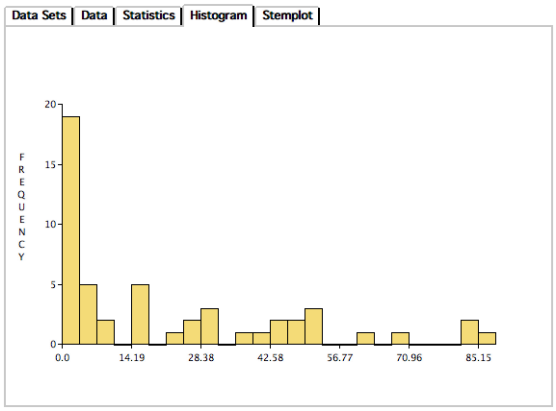
1.20. This distribution is skewed to the right, meaning that Shakespeare's plays contain many short words (up to six letters) and fewer very long words. We would probably expect most authors to have skewed distributions, although the exact shape and spread will vary.

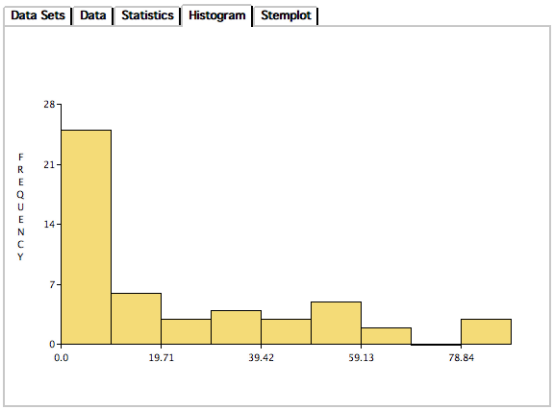
1.21. There are three peaks in the histogram: One at \$4–\$6 thousand, one at \$18–20 thousand, and one at \$28–\$30 thousand. There is a clear break between the least expensive schools and the rest; the line between the middle and most expensive schools is not so clear. Presumably, the lowest group (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).

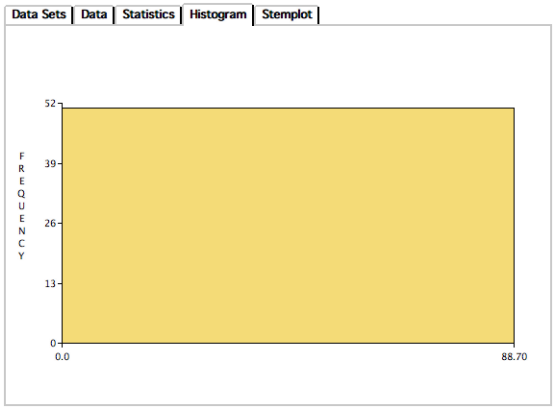
1.22. **(a)** The top five states are Texas, Minnesota, Oklahoma, Missouri, and Illinois. The bottom five are Alaska, Puerto Rico, Rhode Island, Nevada and Vermont. **(b)** The histogram (right) shows a sharp right skew, with a large peak (25 of the 51 numbers) in the "less than 10" category; arguably, 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.

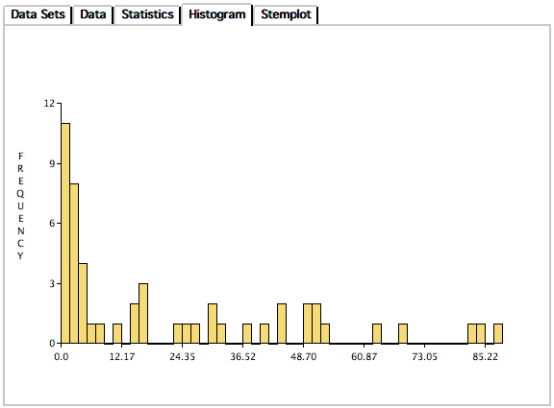


1.23. **(a)** The applet defaulted (for me) to 25 intervals. This histogram is shown below, along with the nine-class histogram. Note that the latter does not *exactly* match the histogram of the previous problem, because the applet's classes are about 9.85 units wide, rather than 10 units wide. **(b)** 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 to 12 seems like a good range.





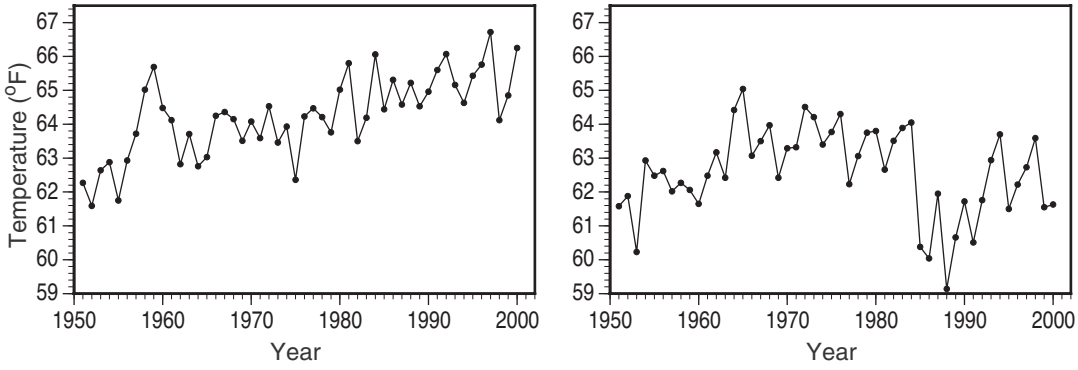




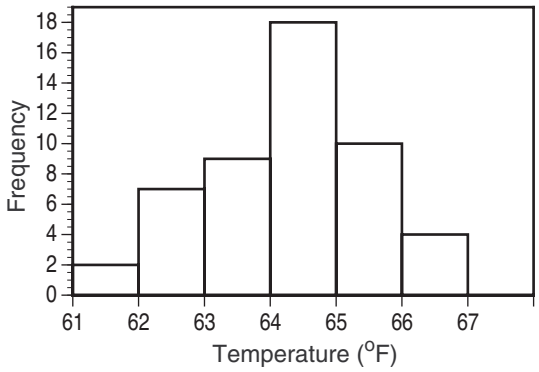
1.24. (a) Totals emissions would almost certainly be higher for very large countries; for example, we would expect that even with great attempts to control emissions, China (with over 1 billion people) would have higher total emissions than the smallest countries in the data set. **(b)** A stemplot is 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 U.S., 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

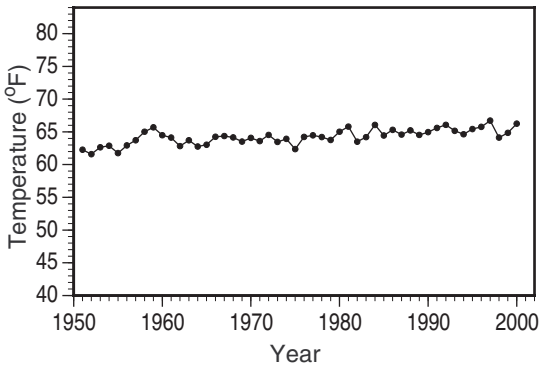
1.25. Shown below are two separate graphs (Pasadena on the left, Redding on the right); students may choose to plot both time series on a single set of axes. If two graphs are created, they should have the same vertical scale for easy comparison. Both plots show random fluctuation. Pasadena temperatures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

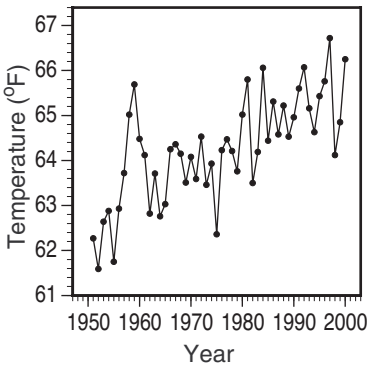


1.26. The distribution is symmetrical and mound-shaped, spread from 61°F to 67°F, with center 64–65°F. The histogram does not show what we see in the time plot from the previous exercise: That mean annual temperature has been rising over time.



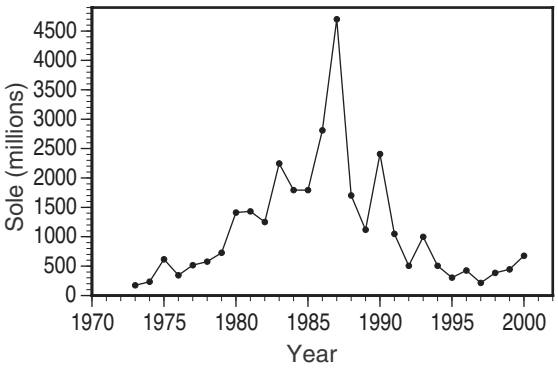
1.27. Shown below are two possible graphs.





1.28. **(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.

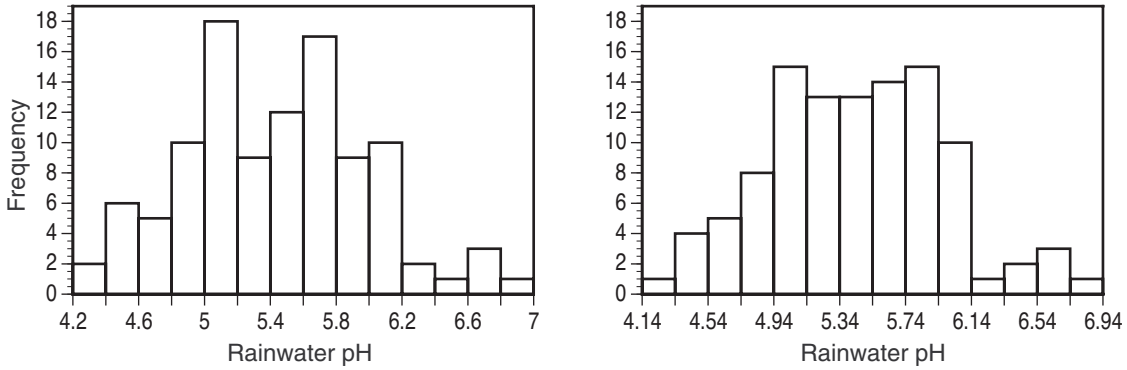
0	12233344
0	55556679
1	01244
1	777
2	24
2	8
3	
3	
4	
4	7



1.29. 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 | 00000000000000000000000000000000000011111111111111111111
0 | 22222222222222222333333333333333333333333
0 | 44444444444444444445555555555555555555
0 | 66666666666666666667777777777777
0 | 88888888888888899999999999999999
1 | 000000000000111111111
1 | 2222222222233333333333
1 | 444444455
1 | 66666777
1 | 8
```

1.30. **(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 would usually be viewed as distinct modes. **(c)** The results will vary with the software used.



1.31. From the top left histogram: 4, 2, 1, 3. The upper-left hand graph is studying time; 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: 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.)

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

1.33. 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 or 40 thousand 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 thousand barrels.

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

1.34. 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

1.35. (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.) **(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	

1.36. 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	55555667778888888888889999999
1	00000000000111222233344 4
1	56777899
2	1144
2	
3	2
3	8
4	0
4	
5	12
5	9

1.37. (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

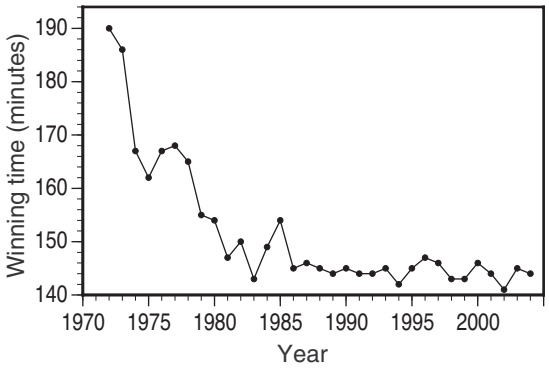
1.38. Stemplot at right, with split stems. The distribution is fairly symmetric—perhaps slightly left-skewed—with center around 110 (clearly above 100). IQs range from the low 70s to the high 130s, with a "gap" in the low 80s.

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

1.39. 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

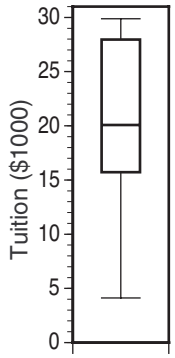
1.40. The time plot on the right shows that women's times decreased quite rapidly from 1972 until the mid-1980s. Since that time, they have been fairly consistent: All times since 1986 are between 141 and 147 minutes.



1.41. 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 histogram. See also the solution to Exercise 1.21.



1.42. **(a)** The five-number summary is

$\text{Min} = 5.7\%,\ Q_1 = 11.7\%,\ M = 12.75\%,\ Q_3 = 13.5\%,\ \text{Max} = 17.6\%.$

(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 whichever state has 8.5%.

1.43. **(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.

1.44. See also the solution to Exercise 1.24. **(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 U.S. 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

1.45. 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.

1.46. **(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 $M - Q_1$.

1.47. The total salary is \$500,000, so the mean is $\bar{x} = \frac{\$500,000}{8} = \$62,500$. Seven of the eight employees (everyone but the owner) earned less than the mean. The median is $M = \$25,000$.

1.48. 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.

1.49. The total salary is now \$700,000, so the new mean is $\bar{x} = \frac{\$700,000}{8} = \$87,500$. The median is unchanged.

1.50. 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

1.51. 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 ten-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.

1.52. **(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

1.53. **(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

1.54. See also the solution to Exercise 1.37. **(a)** The mean of this distribution appears to be higher than 100. (There is no substantial difference between the standard deviations.)

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

(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).

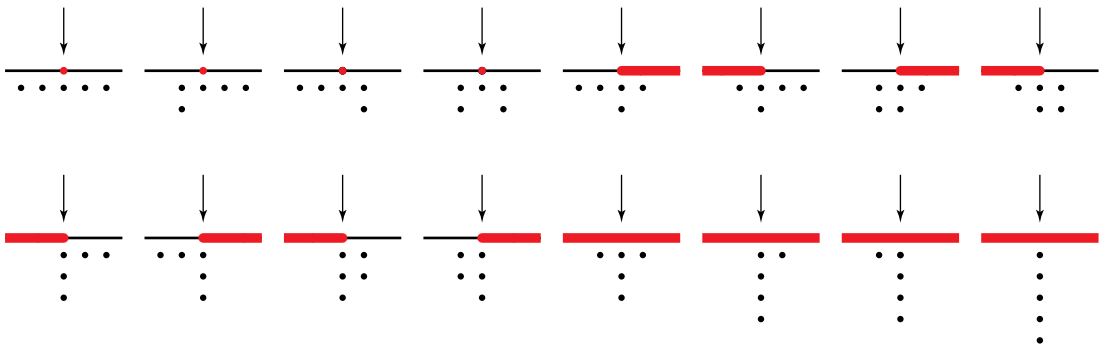
Note: *Students may be somewhat puzzled by the statement in (b) that the median is "close to the mean" (when they differ by 1.1), followed by (c) where they "differ a bit" (when* $M - \bar{x} = 0.382$*). It may be useful to emphasize that we judge the size of such differences relative to the spread of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

1.55. 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.

1.56. **(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.

1.57. **(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 6 must both equal the third point of the set of 5.

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.)

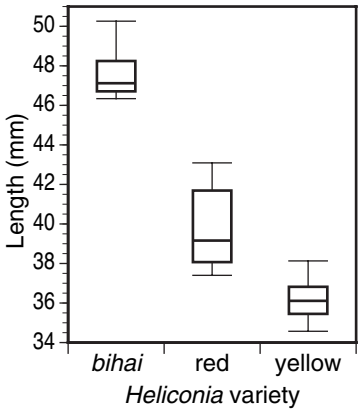


(b) Regardless of the configuration of the first 5 points, if the sixth point is added so as to leave the median unchanged, then in that (sorted) set of 6, the third and fourth points must be equal. One of these 2 points will be the middle (fourth) point of the (sorted) set of 7, no matter where the 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.*

1.58. 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.

1.59. (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	
46	3466789
47	114
48	0133
49	
50	12

red	
37	4789
38	0012278
39	167
40	56
41	4699
42	01
43	0

yellow	
34	56
35	146
36	0015678
37	01
38	1

(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.

1.60. 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	
0	
0	
1	
1	699
2	0124
2	7789
3	3

1 year ago	
0	2
0	9
1	2244
1	57789
2	0
2	
3	

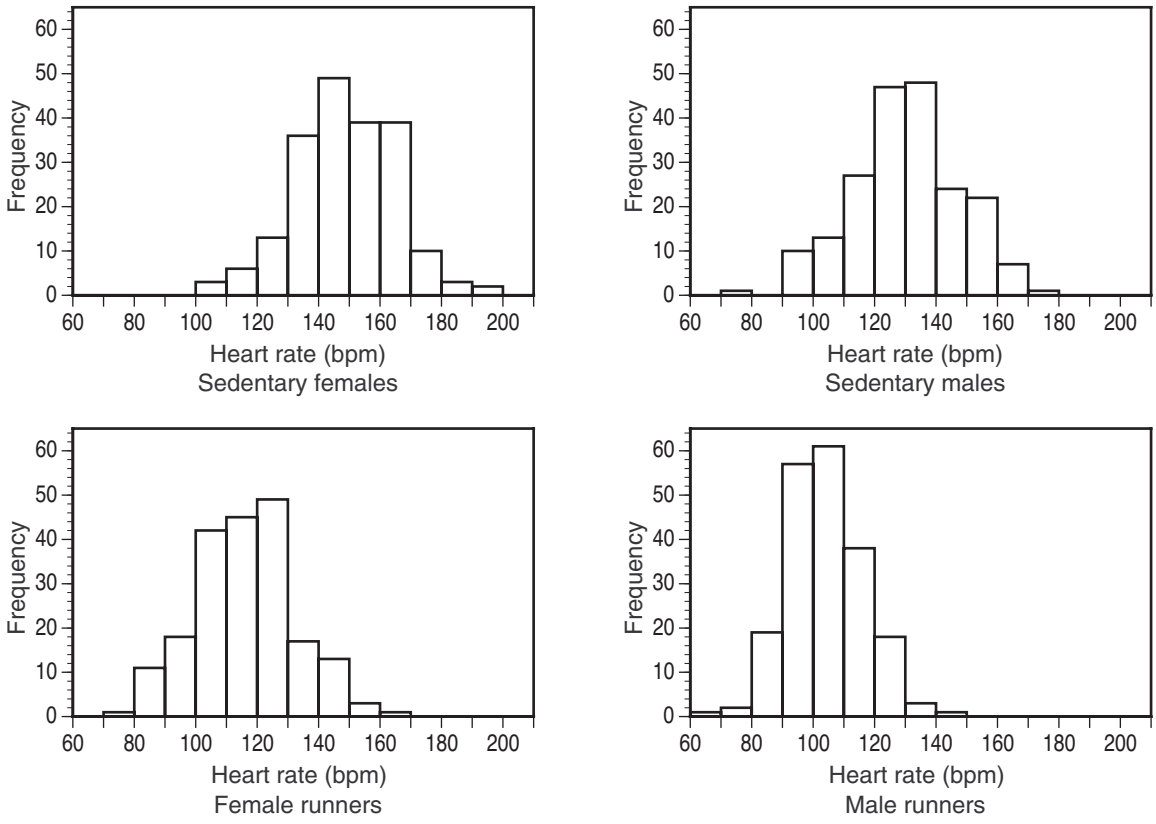
8 years ago	
0	4
0	
1	22
1	5889
2	22
2	
3	

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).

1.61. 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.*



1.62. 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.*

1.63. 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.)

1.64. 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.

1.65. 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

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

1.66. Because the mean is to be 7, the five numbers must add 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

1.67. 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 $\bar{x} = (a+b+c+d+e+f)/6 > (e+f)/6 > (e+5e)/6 = e$.

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

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

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

1.69. **(a)** One possible answer is 1, 1, 1, 1. **(b)** 0, 0, 10, 10. **(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 10 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, 10, 10, 10—or 10, 0, 0, 0—we make the first squared deviation 7.5^2, but the other three are only 2.5^2. Our best choice is two at each extreme, which makes all four squared deviations equal to 5^2.

1.70. Answers will vary. Typical calculators will carry only about 12 to 15 digits; for example, a TI-83 fails (gives $s = 0$) for 13-digit numbers. *Excel* (at least the version I checked) gives $s = 0$ for nine-digit numbers. The version of Minitab used to prepare these answers fails at 100,000,001 (nine digits).

1.71. See Exercise 1.36 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.

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

1.73. The table on the right reproduces the means and standard deviations from the solution to Exercise 1.59, 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, $\bar{x} = (47.5975 \text{ mm})(0.03937 \text{ in/mm}) = (47.5975 \text{ mm}) \div (25.4 \text{ mm/in}) = 1.874$ 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

1.74. (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:*

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

1.75. 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.

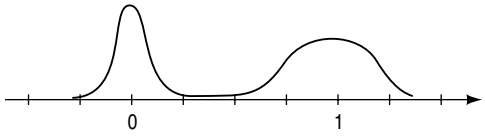
1.76. Variance is changed by a factor of $2.54^2 = 6.4516$; generally, for a transformation $x_{\text{new}} = a + bx$, the new variance is b^2 times the old variance.

1.77. There are 72 survival times, so to find the 10% trimmed mean, remove the highest and lowest 7 values (leaving 58). Remove the highest and lowest 14 values (leaving 44) for the 20% trimmed mean.

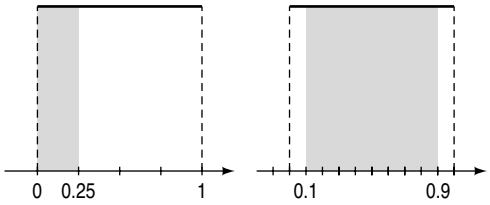
The mean and median for the full data set are $\bar{x} = 141.8$ and $M = 102.5$. The 10% trimmed mean is $\bar{x}^* = 118.16$, and the 20% trimmed mean is $\bar{x}^{**} = 111.68$. Since the distribution is right-skewed, removing the extremes lowers the mean.

1.78. Sketches will vary.

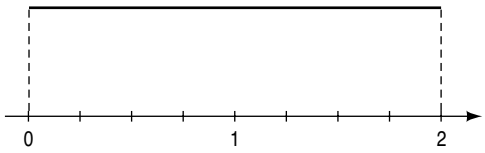
1.79. Sketches will vary, but should be some variation on the one shown here: The peak at 0 should be "tall and skinny," while near 1, the curve should be "short and fat."



1.80. **(a)** The curve forms a 1×1 square, which has area 1.
(b) $P(X > 0.75) = 0.25$.
(c) $P(0.25 < X < 0.75) = 0.5$.



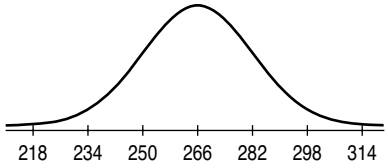
1.81. **(a)** The height should be $\frac{1}{2}$, since the area under the curve must be 1. The density curve is at right. **(b)** $P(X \leq 1) = \frac{1}{2}$. **(c)** $P(0.5 < X < 1.3) = 0.4$.



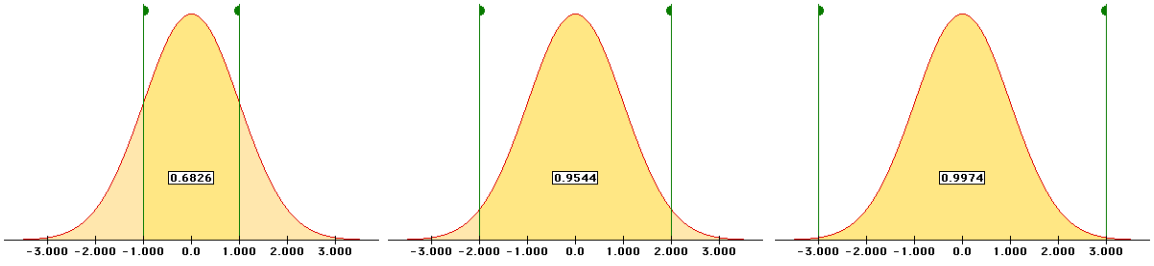
1.82. The mean and median both equal 0.5; the quartiles are $Q_1 = 0.25$ and $Q_3 = 0.75$.

1.83. **(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).

1.84. Hint: 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.

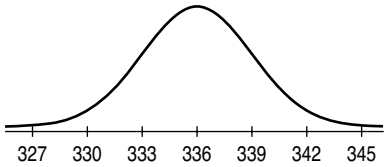


1.85. **(a)** The applet shows an area of 0.6826 between -1.000 and 1.000, while the 68–95–99.7 rule rounds this to 0.68. **(b)** Between -2.000 and 2.000, the applet reports 0.9544 (compared to the rounded 0.95 from the 68–95–99.7 rule). Between -3.000 and 3.000, the applet reports 0.9974 (compared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



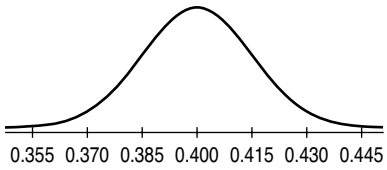
1.86. See the sketch of the curve in the solution to Exercise 1.84. **(a)** The middle 95% fall within two standard deviations of the mean: $266 \pm 2(16)$, or 234 to 298 days. **(b)** The shortest 2.5% of pregnancies are shorter than 234 days (more than two standard deviations below the mean).

1.87. **(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.*

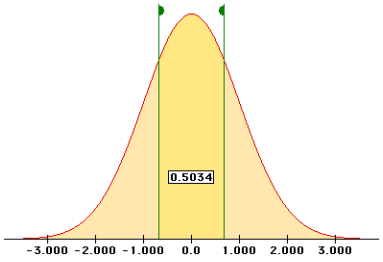
1.88. **(a)** About 50% of samples give values above the mean (0.4). Since 0.43 is two standard deviations above the mean, about 2.5% of sample values are above 0.43. **(b)** 0.37 to 0.43—that is, $0.4 \pm 2(0.015)$.



Note: *As the text models, it is probably best to use decimals for these proportions rather than percentages (0.37 instead of 37%) to lessen the confusion with, for example, 95%.*

1.89. 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 six feet is quite tall for a woman, but not at all extraordinary for a man.

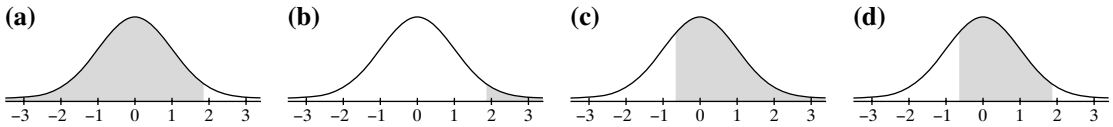
1.90. Because the quartiles of any distribution have 50% of observations between them, we seek to place the flags so that the reported area is 0.5. The closest the applet gets is an area of 0.5034, between -0.680 and 0.680. Thus, the quartiles of any normal distribution are about 0.68 standard deviations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



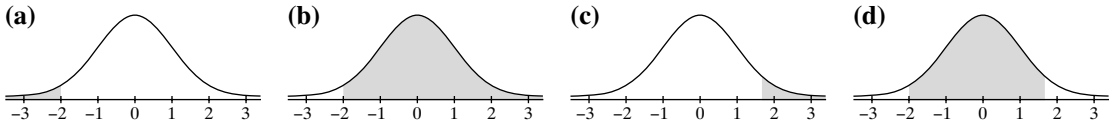
Note: *Table A places the quartiles at about ±0.67; other statistical software gives ±0.6745.*

1.91. 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.

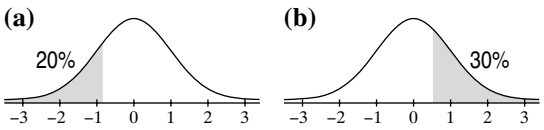
1.92. **(a)** $Z < 1.85$: 0.9678. **(b)** $Z > 1.85$: 0.0322. **(c)** $Z > -0.66$: 0.7454. **(d)** $-0.66 < Z < 1.85$: $0.9678 - 0.2546 = 0.7132$.



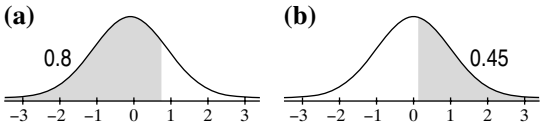
1.93. **(a)** $Z \leq -2$: 0.0228. **(b)** $Z \geq -2$: 0.9772. **(c)** $Z > 1.67$: 0.0475. **(d)** $-2 < Z < 1.67$: $0.9772 - 0.0475 = 0.9297$.



1.94. (a) 20% of the observations fall below −0.8416. (This is the 20th percentile of the standard normal distribution). **(b)** 30% of the observations fall above 0.5244 (the 70th percentile of the standard normal distribution).



1.95. (a) $z = 0.8416$ has cumulative proportion 0.8 (that is, 0.8416 is the 80th percentile of the standard normal distribution). **(b)** If $z = 0.1257$, then $Z > z$ has proportion 0.45 (0.1257 is the 55th percentile).



1.96. 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.

1.97. 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.

1.98. Tonya's score standardizes to $z = \frac{1318-1026}{209} \doteq 1.3971$, while Jermaine's score corresponds to $z = \frac{27-20.8}{4.8} \doteq 1.2917$. Tonya's score is higher.

1.99. Jacob's score standardizes to $z = \frac{16-20.8}{4.8} = -1$, while Emily's score corresponds to $z = \frac{670-1026}{209} \doteq -1.7033$. Jacob's score is higher.

1.100. Jose's score standardizes to $z = \frac{1287-1026}{209} \doteq 1.2488$, so an equivalent ACT score is $20.8 + 1.2488 \times 4.8 \doteq 26.8$. (Of course, ACT scores are reported as whole numbers, so this would presumably be a score of 27.)

1.101. Maria's score standardizes to $z = \frac{28-20.8}{4.8} = 1.5$, so an equivalent SAT score is $1026 + 1.5 \times 209 = 1339.5$ (presumably reported as either 1339 or 1340.)

1.102. Tonya's score standardizes to $z = \frac{1318-1026}{209} \doteq 1.3971$; this is the 92nd percentile.

1.103. Jacob's score standardizes to $z = \frac{16-20.8}{4.8} = -1$; this is the 16th percentile.

1.104. A score of 1600 standardizes to $z = \frac{1600-1026}{209} \doteq 2.7368$. 99.7% of standard scores are below this level, so about 0.3% are above this level (and are therefore reported as 1600).

1.105. A score of 36 standardizes to $z = \frac{36-20.8}{4.8} \doteq 3.1667$. About 99.9% of standard scores are below this level, so about 0.1% are above this level (and are therefore reported as 36).

1.106. 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.

1.107. The top 20% corresponds to a standard score of $z = 0.8416$, which in turn corresponds to a score of $20.8 + 0.8416 \times 4.8 \doteq 24.8$ (or 25) on the ACT.

1.108. 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.

1.109. 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 $1026 + 0.8416 \times 209 = 1202$.

1.110. **(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.

1.111. 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.

1.112. **(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.

1.113. **(a)** About 3.3% of men score 750 or better: Among men, a score of 750 corresponds to standard score $z = \frac{750-537}{116} \doteq 1.8362$, for which the cumulative probability is 0.9668. **(b)** About 1.2% of women score 750 or better: Among women, a score of 750 corresponds to standard score $z = \frac{750-503}{110} \doteq 2.2455$, for which the cumulative probability is 0.9876.

1.114. **(a)** The middle 95% of yearly returns is $8.3\% \pm 2(20.3\%) = -32.3\%$ to 48.9%. **(b)** A return of 0% corresponds to a standard score of $z = \frac{0-8.3}{20.3} \doteq -0.4089$, so the market is down in about 34% of all years. **(c)** A return of 25% corresponds to a standard score of $z = \frac{25-8.3}{20.3} \doteq 0.8227$, so the index gains at least 25% in about 20.5% of all years.

1.115. **(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 $x > 279.47$.)

1.116. **(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.

1.117. (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 $Q_3 = \mu + 0.6745\sigma$.

1.118. 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 $2 \times 0.0035 = 0.70\%$.

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

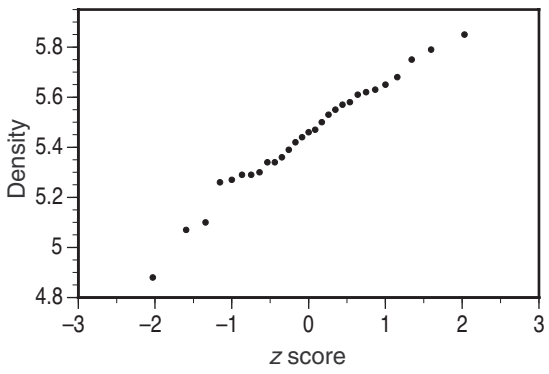
1.120. The shape of the quantile plot suggests that the data are right-skewed (as was observed in Exercises 1.24 and 1.44). This can be seen in the flat section in the lower left—these numbers were less spread out than they should be for normal data—and the three apparent outliers (the U.S., Canada,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

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

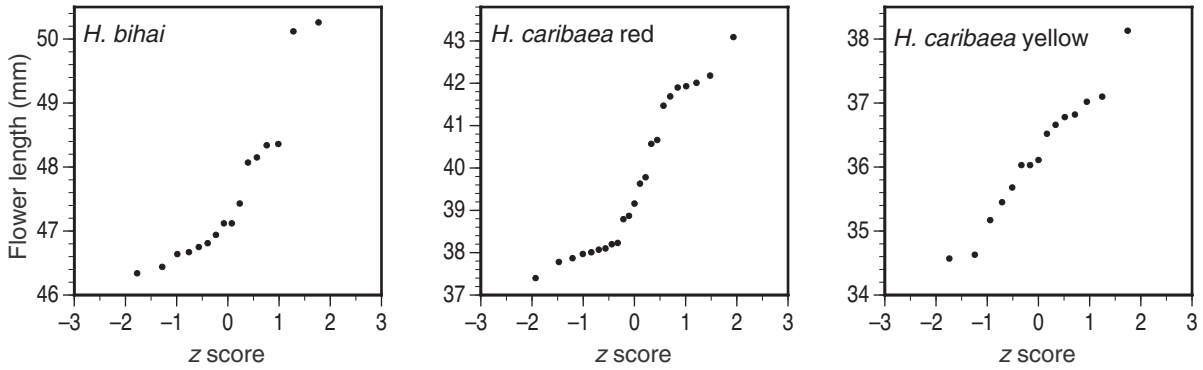
1.122. (a) is the graph of (3) the highway gas mileages: Aside from the Insight, these numbers are reasonably normal, and in this graph, the points fall close to a line aside from one high outlier. **(b)** is the graph of (1) the IQ data: This distribution was the most normal of the four, and this graph 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).*

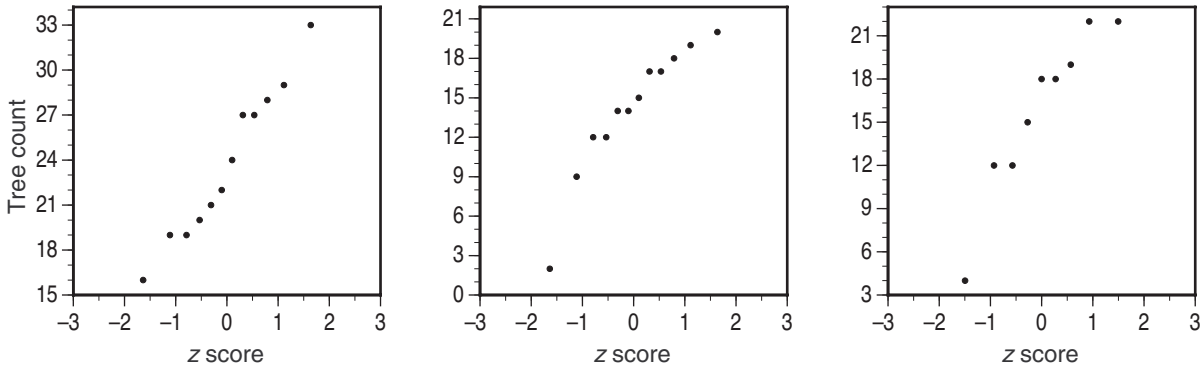
1.123. See also the solution to Exercise 1.34. The plot suggests no major deviations from normality, although the three lowest measurements do not quite fall in line with the other points.



1.124. **(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.

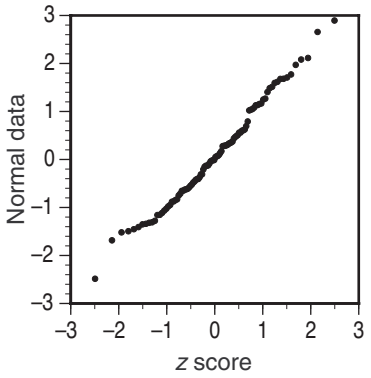


1.125. See also the solution to Exercise 1.60. The first plot (for never-logged areas) is nearly linear. The other two each show a low value, perhaps suggesting a slight skew to the left.



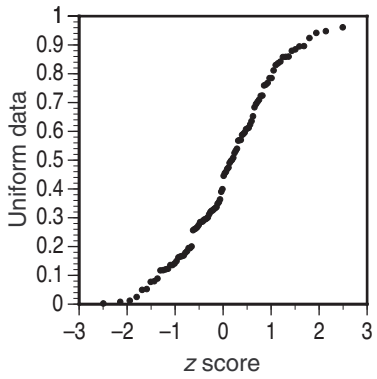
1.126. A stemplot from one sample is shown. Histograms will vary slightly but should suggest a bell curve. The normal probability plot shows something fairly close to a line but illustrates that, even for actual normal data, the tails may deviate slightly from a line.

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



1.127. 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



1.128. (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."

1.129. 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 is preferred:

22 23.735 24.31 24.845 28.55

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

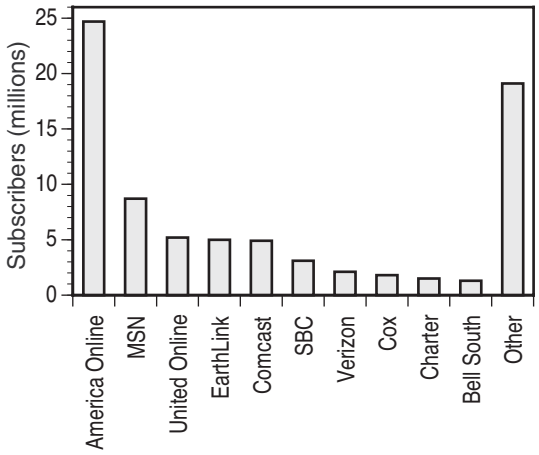
22	013
22	7899
23	000011222233344444
23	55566666667777778888888999
24	000000111111122222222233333333333444444
24	55555566666666677777888888999999
25	00001111233344
25	56666889
26	2
26	56
27	2
27	
28	
28	5

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

1.131. Many—but less 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.*

1.132. 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.

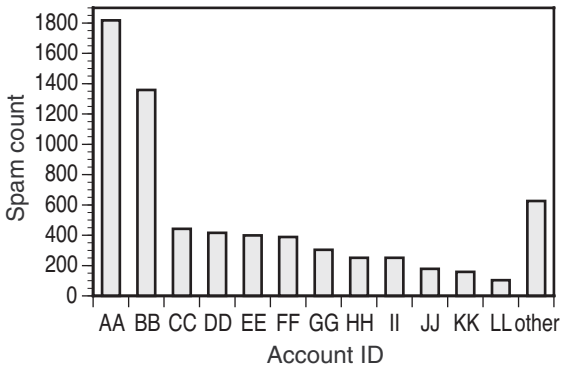


1.133. Women's weights are skewed to the right: This makes the mean higher than the median, and it is also revealed in the differences $M - Q_1 = 14.9$ pounds and $Q_3 - M = 24.1$ pounds.

1.134. **(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.

1.135. **(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.)

1.136. 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.

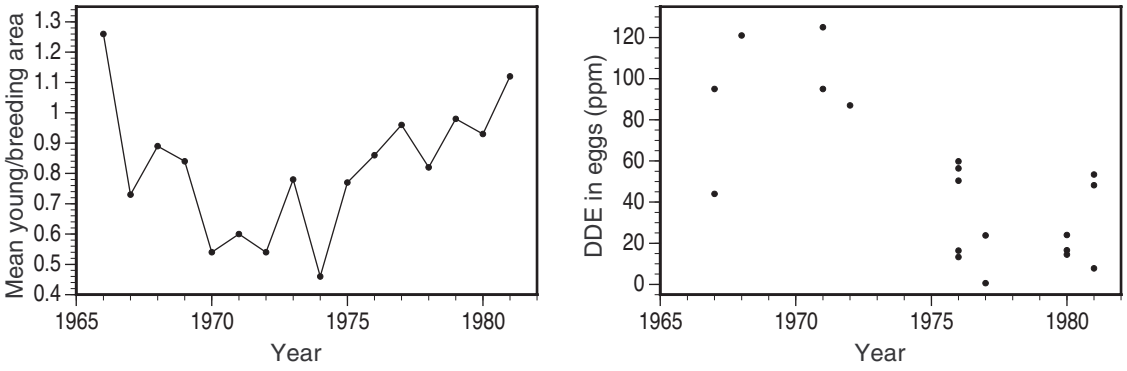


1.137. No, and no: It is easy to imagine examples of many different data sets with mean 0 and standard deviation 1—for example, $\{-1,0,1\}$ and $\{-2,0,0,0,0,0,0,0,2\}$.

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

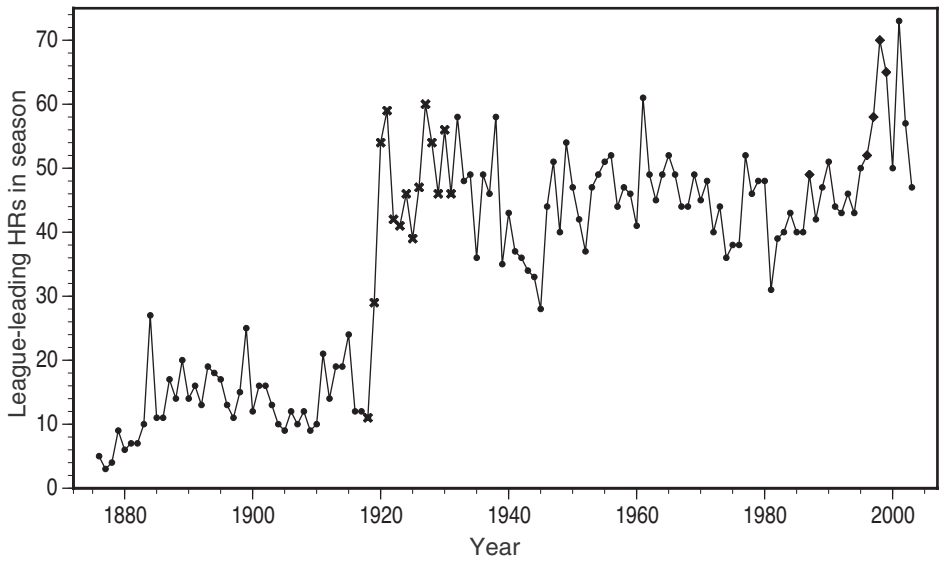
and adding some additional numbers in between them, for example (in increasing order): 10, __, 20, __, __, 30, __, __, 40, __, 50. As long as the number in the first blank is between 10 and 20,and so on, the five-number summary will be 10, 20, 30, 40, 50.

1.138. 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.)



1.139. 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 ten 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.)



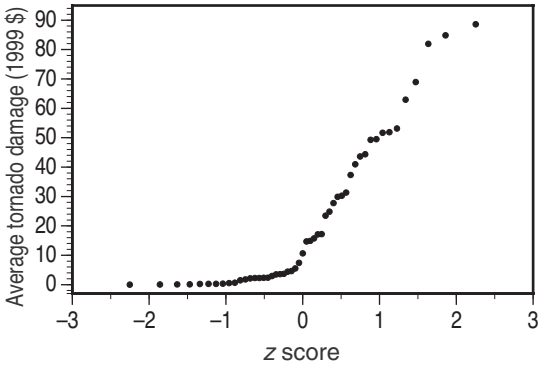
1.140. Bonds's mean changes from 36.56 to 34.41 home runs (a drop of 2.15), while his median changes from 35.5 to 34 home runs (a drop of 1.5). This illustrates that outliers affect the mean more than the median.

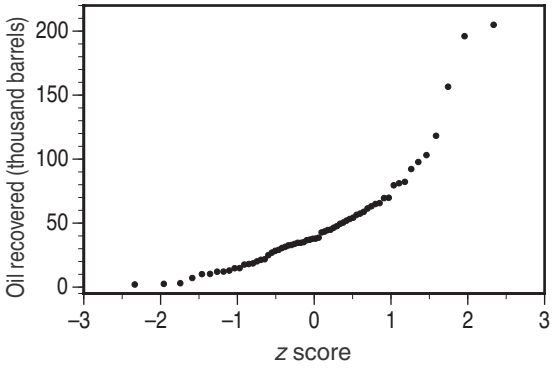
1	69
2	4
2	55
3	3344
3	77
4	02
4	5669
5	
5	
6	
6	
7	3

1.141. 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.

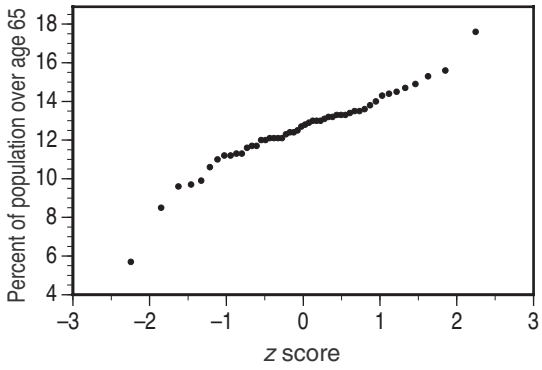
1.142. Shown below are both quantile plots. Skewness shows up in a quantile plot as a flat tail; for right-skewness, that flat portion is at the beginning (the lower left).

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.)

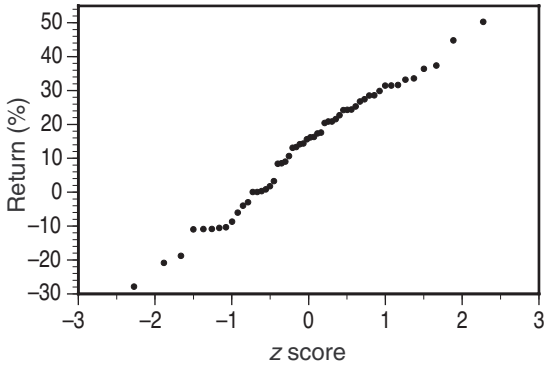




1.143. **(a)** Sketches may vary somewhat, but should be linear in the middle; the outliers would show up as a point in the lower left *below* the line (because low outliers are less than we expect them to be for a normal distribution), and a point in the upper right *above* the line (because high outliers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



1.144. **(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.



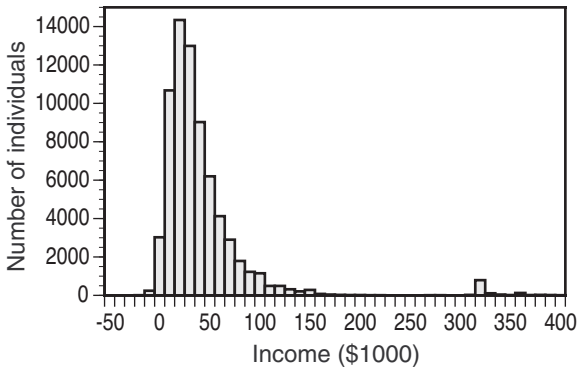
Note: *For an example of a quantile plot of a heavy-tailed distribution, see the tuition-and-fees data from Exercise 1.21; a quantile plot is shown in Figure 1.41(d), which accompanies Exercise 1.122.*

1.145. 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!

1.146. 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:

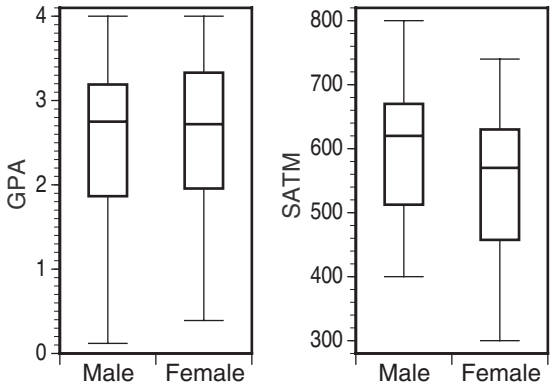
$$\text{Min} = -\$23{,}980,\ Q_1 = \$22{,}000,\ M = \$35{,}000,\ Q_3 = \$53{,}000,\ \text{Max} = \$609{,}548$$

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.*

1.147. 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.

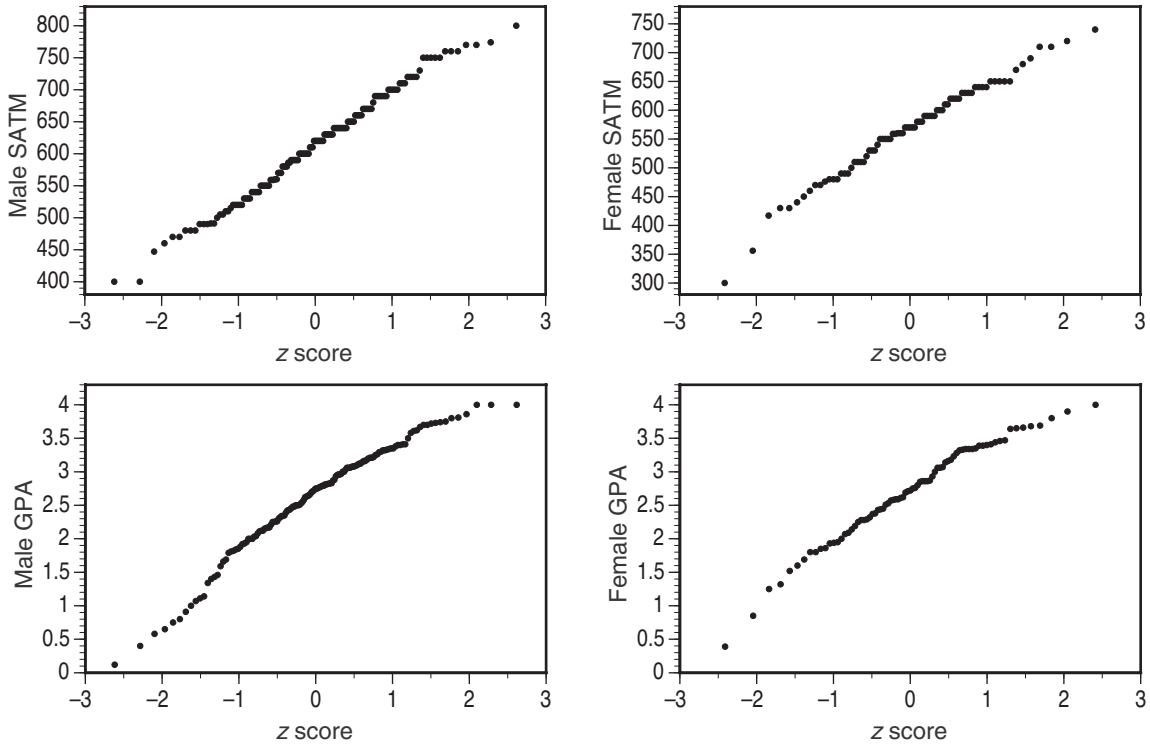
Quantile plots are shown below. Judging from these (and from the $1.5 \times IQR$ criterion), student 183 is an outlier for female SATM (300). For male GPA, outliers are students 127 (GPA 0.12) and 90 (GPA 0.4), and for female GPA, the outlier is student 188 (GPA 0.39). (Judgments of these may vary if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



	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

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.*



For answers to the EESEE Case Studies (exercises 148–150), see the instructor's version of EESEE.

Chapter 2 Solutions

2.1. **(a)** Time spent studying is explanatory; the grade is the response variable. **(b)** Explore the relationship; there is no reason to view one or the other as explanatory. **(c)** Rainfall is explanatory; crop yield is the response variable. **(d)** Explore the relationship. **(e)** Income is explanatory; years of education completed is the response variable.

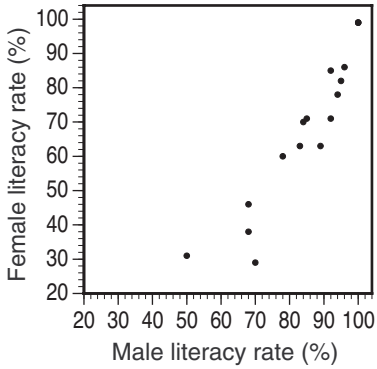
2.2. Parents' income is explanatory, and college debt is the response. Both variables are quantitative. We would expect a negative association: Low income goes with high debt, high income with low debt.

2.3. **(a)** In general, we expect more intelligent children to be better readers, and less intelligent children to be weaker. The plot does show this positive association. **(b)** The four points are for children who have moderate IQs but poor reading scores. **(c)** The rest of the scatterplot is roughly linear, but quite weak (there would be a lot of variation about any line we draw through the scatterplot).

2.4. **(a)** From the scatterplot, we estimate 50% in 1954 and about -28% in 1974. (The data file `ex01-144.dat` gives the values 50.28% and -27.87%.) **(b)** The return on Treasury bills in 1981 was about 14.8%. **(c)** The scatterplot shows no clear pattern. (The statement that "high treasury bill returns tend to go with low returns on stocks" implies a negative association; there may be *some* suggestion of such a pattern, but it is extremely weak.)

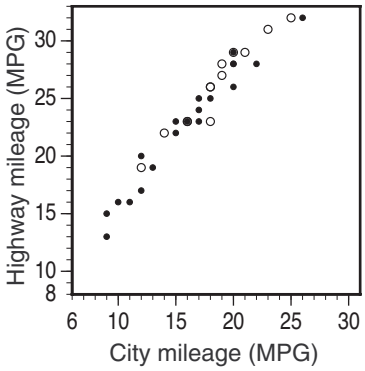
2.5. **(a)** The response variable (estimated level) can only take on the values 1, 2, 3, 4, 5, so the points in the scatterplot must fall on one of those five levels. **(b)** The association is (weakly) positive. **(c)** The estimate is 4, which is an overestimate; that child had the lowest score on the test.

2.6. Ideally, the scales should be the same on both axes. The scatterplot shows a fairly strong, positive, linear association. Three countries (Tajikistan, Kazakhstan, Uzbekistan) reported 100% literacy for men and 99% literacy for women. Yemen (70% for men, 29% for women) might be considered an outlier.

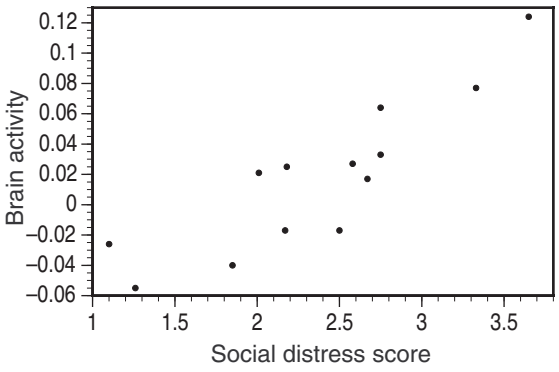


2.7. (a) If we used the number of males returning, then we might not see the relationship, because areas with many breeding pairs would correspondingly have more males that might potentially return. (In the given numbers, the number of breeding pairs varies only from 28 to 38, but considering hypothetical data with 10 and 100 breeding pairs makes more apparent the reason for using percents rather than counts.) **(b)** Scatterplot on the right. Mean responses are shown as crosses; the mean responses with 29 and 38 breeding pairs are (respectively) 71.3333% and 48.5% males returning. **(c)** The scatterplot does show the negative association we would expect if the theory were correct.

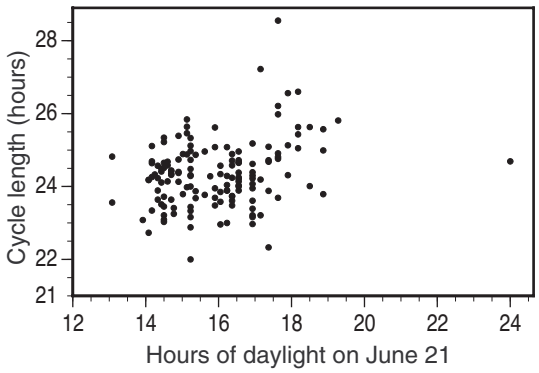
2.8. (a) Two-seater cars are shown as filled circles, minicompact cars as open circles. Ideally, the scales should be the same on both axes. **(b)** The scatterplot shows a strong, positive, linear association. Two-seater cars include several vehicles with poor fuel efficiency (most notably, the Lamborghini and Ferrari models, and perhaps also the Maserati); apart from these cars, the two sets of points show basically the same relationship for both types of cars.



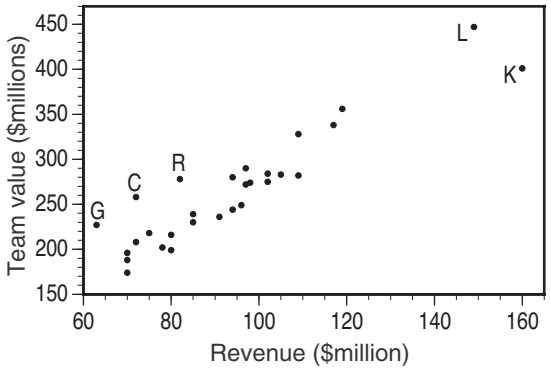
2.9. The scatterplot shows a fairly strong, positive, linear association. There are no particular outliers; each variable has low and high values, but those points do not deviate from the pattern of the rest. Social exclusion does appear to trigger a pain response.

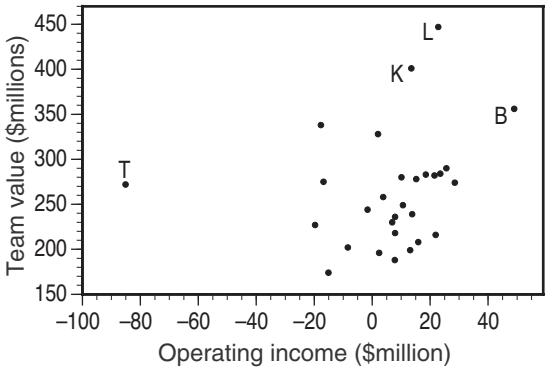


2.10. There appears to be a positive association between cycle length and day length, but it is quite weak: The points of the scatterplot are generally located along a positively-sloped line, but with a lot of spread around that line. (Ideally, both axes should have the same scale.)

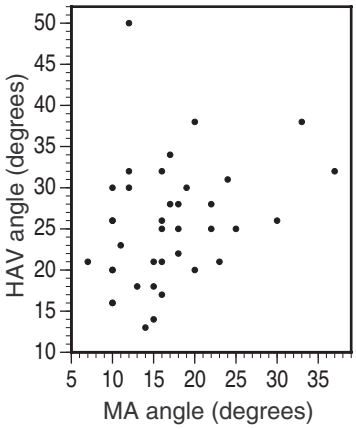


2.11. (Ideally, both graphs should have the same scale on both axes. However, this makes the graph dimensions rather awkward, so the graphs below do not reflect that ideal.) **(a)** The Lakers and the Knicks are high in both variables (but fit the pattern). The Grizzlies, Cavaliers, and Rockets have slightly higher values than their revenues would suggest. The association is positive and linear. **(b)** The Lakers and Knicks still stand out, as do the Bulls and Trailblazers, but the association is quite weak. (It hardly makes sense to speak of outliers when there is little or no pattern.) Revenue is a much better predictor of value.

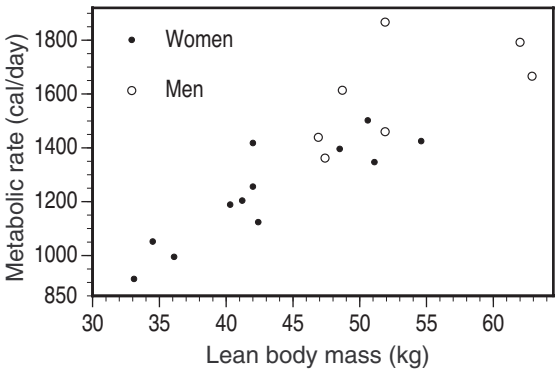




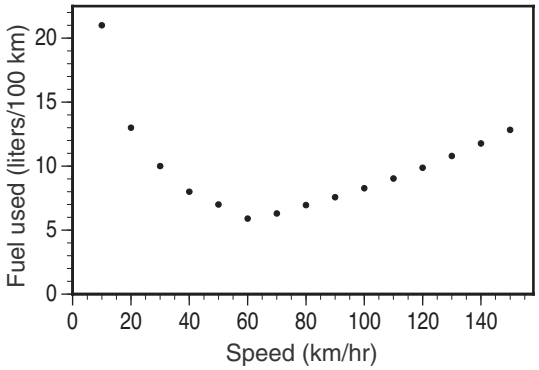
2.12. **(a)** MA angle is the explanatory variable, so it should be on the horizontal axis of the scatterplot. (This scatterplot has the same scale on both axes, because both variables are measured in degrees.) **(b)** The scatterplot shows a moderate-to-weak positive linear association, with one clear outlier (the patient with HAV angle 50°). **(c)** MA angle can be used to give (very rough) estimates of HAV angle, but the spread is so wide that they would not be too reliable.



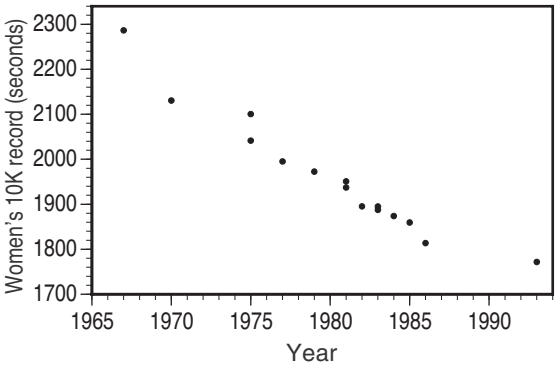
2.13. (a) Women are marked with filled circles, men with open circles. **(b)** The association is linear and positive. The women's points show a stronger association. As a group, males typically have larger values for both variables.



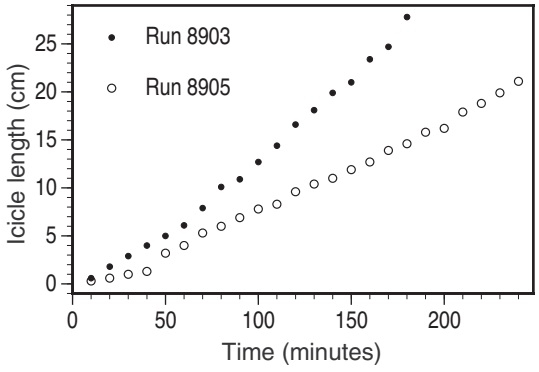
2.14. (a) At right; speed is explanatory, so it belongs on the x-axis. **(b)** The relationship is curved—low in the middle, higher at the extremes. Because low "mileage" is actually *good* (it means that we use less fuel to travel 100 km), this makes sense: moderate speeds yield the best performance. Note that 60 km/hr is about 37 mph. **(c)** Above-average (that is, bad) values of "fuel used" are found with both low and high values of "speed." **(d)** The relationship is very strong—there is little scatter around the curve, and it is very useful for prediction.



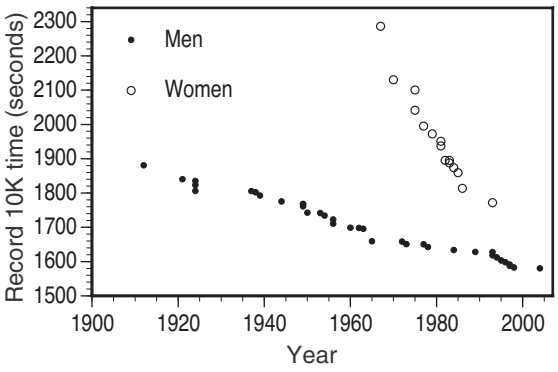
2.15. The plot shows a fairly steady rate of improvement until the mid-1980s, with much slower progress after that (the record has only been broken once since 1986).



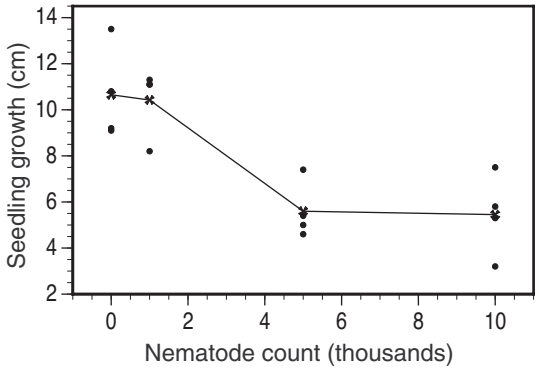
2.16. (a) In the scatterplot on the right, the open circles represent run 8905, the higher flow rate. **(b)** Icicles seem to grow faster when the water runs more slowly. (Note that there is no guarantee that the pattern we observe with these two flow rates applies to rates a lot faster than 29.6 mg/s, or slower than 11.9 mg/s.)



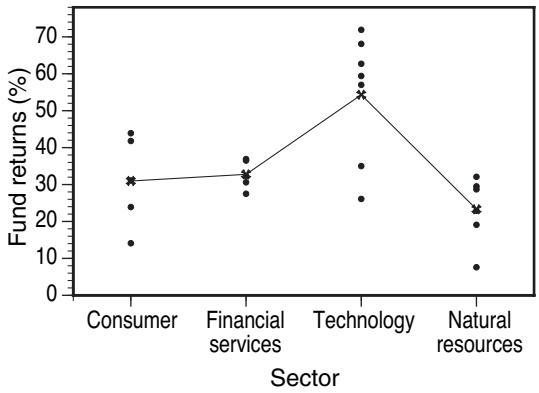
2.17. **(a)** Both men (filled circles) and women (open circles) show fairly steady improvement. Women have made more rapid progress, but their progress seems to have slowed, while men's records may be dropping more rapidly in recent years. **(b)** The data supports the first claim, but does not seem to support the second.



2.18. **(a)** The scatterplot on the right shows both the original data (circles) and the means (crosses). The means are 10.65, 10.43, 5.60, and 5.45 cm. **(b)** There is little difference in the growth when comparing 0 and 1000 nematodes, or 5000 and 10,000 nematodes—but the growth drops substantially between 1000 and 5000 nematodes.



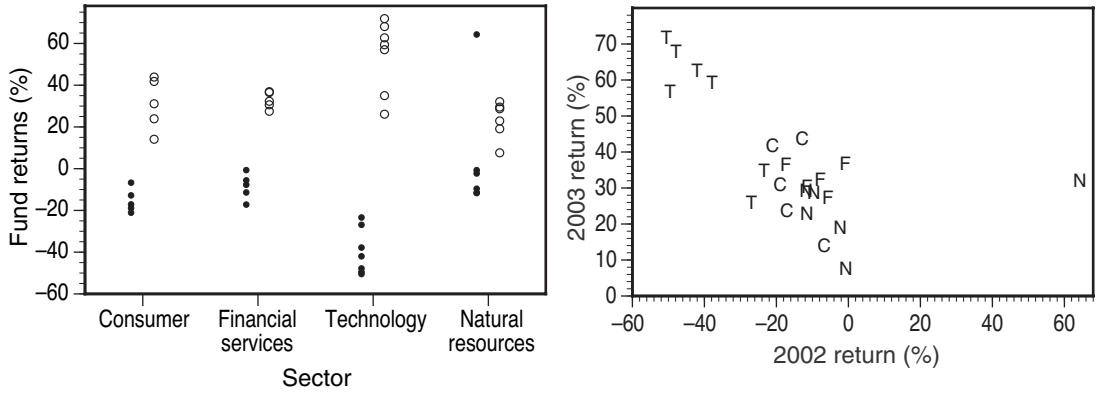
2.19. **(a)** Plot shown on the right. Means (plotted with crosses) are 30.96%, 32.76%, 54.31%, and 23.32%. (Note that the sectors on the horizontal axis are shown there in the order given in the text, but that is completely arbitrary.) **(b)** Technology had the highest average performance. **(c)** Referring to a positive or negative association only makes sense when both variables are quantitative. (There *is* an association here, but it cannot be called positive or negative.)



2.20. Methods of graphical analysis will vary; shown below are two possible approaches. On the left, for each sector, 2002 returns are shown as filled circles and 2003 returns are open circles. On the right is a scatterplot with 2002 return as the explanatory variable; the letters C, F, T, and N indicate the different fund types. The negative association in the second graph makes more clear something that can also be observed in the first graph: Generally, the worse a fund did in 2002, the better it did in 2003 (and vice versa).

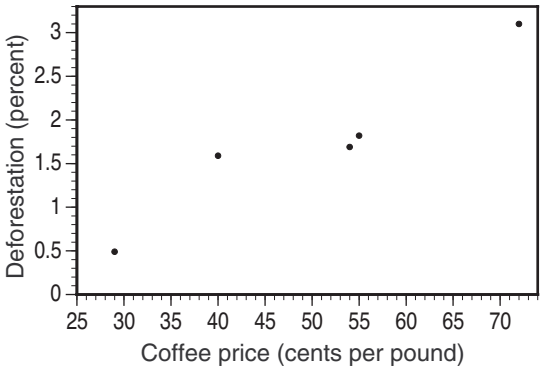
−0	3
−0	
0	0
0	223333
0	444455555
0	6
0	9
1	001
1	2

Also shown (right) is a stemplot of the differences for each fund (that is, each fund's 2003 return minus its 2002 return). Only one fund return decreased; every other fund increased its return by between 8.3% and 122.4%.



2.21. **(a)** Price is explanatory (and so is on the horizontal axis). The plot shows a positive linear association. **(b)** $\bar{x} = 50$ cents/lb and $s_x \doteq 16.3248$ cents/lb; $\bar{y} = 1.738\%$ and $s_y \doteq 0.9278\%$. The standardized values are below; the correlation is $r = 3.8206/4 = 0.955$. **(c)** Obviously, the calculator value should be the same.

z_x	z_y	$z_x\, z_y$
−1.2864	−1.3451	1.7303
−0.6126	−0.1595	0.0977
0.2450	−0.0517	−0.0127
0.3063	0.0884	0.0271
1.3476	1.4679	1.9783
		3.8206



2.22. **(a)** The best guess is $r = 0.6$. There is far too much scatter for $r = 0.9$, and enough of a positive association that r must be more than 0.1. **(b)** The actual correlation is 0.6821.

2.23. The best guess is $r = 0.6$. There is far too much scatter for $r = 0.9$, and enough of a positive association that r must be more than 0.1.

2.24. **(a)** $r = 0.98$ goes with the Dividend Growth fund, which is most similar to the stocks represented by the S&P index. $r = 0.81$ goes with the Small Cap Stock fund; small U.S. companies should be somewhat similar to large U.S. companies. Finally, $r = 0.35$ goes with Emerging Markets, as these stocks would be the most different from those in the S&P index. **(b)** Positive correlations do not indicate that stocks went up. Rather, they indicate that when the S&P index rose, the other funds often did, too—and when the S&P index fell, the other funds were likely to fall.

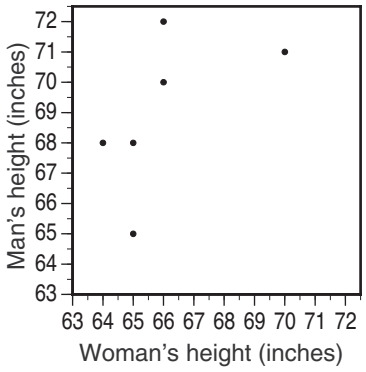
2.25. r would not change; units do not affect correlation.

2.26. **(a)** See the solution to Exercise 2.20 for the scatterplot. (It is the second of the two graphs shown there.) **(b)** For all 23 funds, $r = -0.6230$; with the outlier removed, $r^* = -0.8722$. Removing the Gold fund makes the association stronger, because the remaining points are less scattered about a line drawn through the data points.

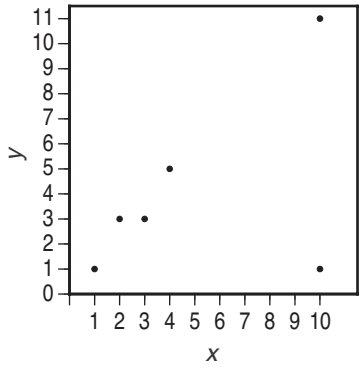
2.27. See also the solution to Exercise 2.11. **(a)** For team value and revenue, $r_1 = 0.9265$; for team value and operating income, $r_2 = 0.2107$. This agrees with conclusions from the scatterplots: revenue is a much better predictor of team value. **(b)** Without Portland (marked with a "T" in the scatterplot), $r_2 = 0.3469$. The removal of this point makes the scatterplot appear (slightly) more linear, so the association is stronger.

2.28. For Exercise 2.10, $r_1 = 0.2797$; for Exercise 2.16, $r_2 = 0.9958$ (run 8903) and $r_3 = 0.9982$ (run 8905).

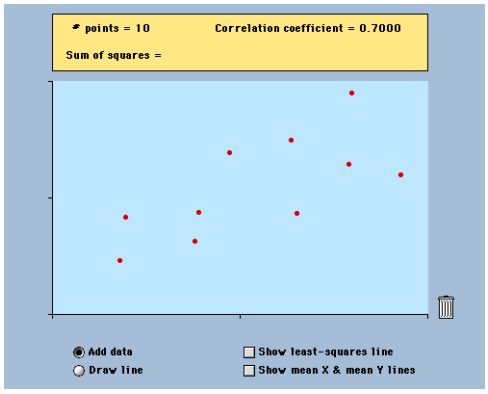
2.29. **(a)** The scatterplot shows a moderate positive association, so r should be positive, but not close to 1. **(b)** The correlation is $r = 0.5653$. **(c)** r would not change if all the men were six inches shorter. A positive correlation does not tell us that the men were generally taller than the women; instead it indicates that women who are taller (shorter) than the average woman tend to date men who are also taller (shorter) than the average man. **(d)** r would not change, because it is unaffected by units. **(e)** r would be 1, as the points of the scatterplot would fall on a positively-sloped line.

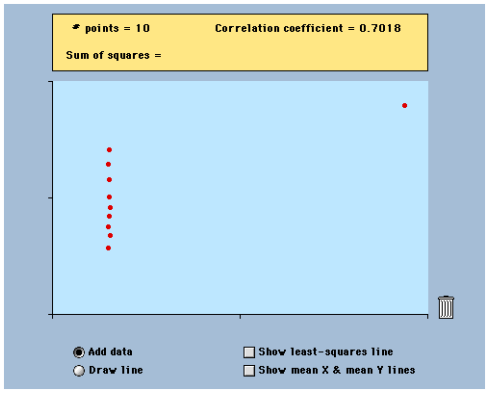


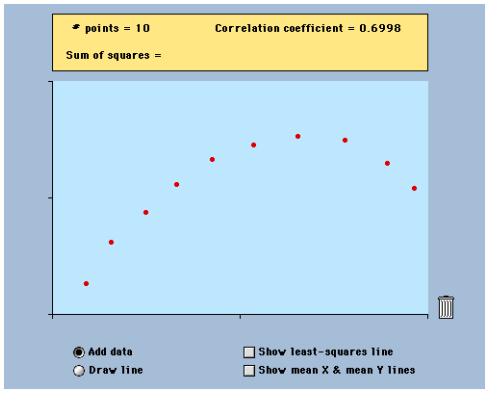
2.30. The correlation is $r \doteq 0.481$. The correlation is greatly lowered by the one outlier. Outliers tend to have fairly strong effects on correlation; it is even stronger here because there are so few observations.



2.31. **(a)** As two points determine a line, the correlation is always either -1 or 1. **(b)** Sketches will vary; an example is shown on the right. Note that the scatterplot must be positively sloped, but r is affected only by the scatter about a line drawn through the data points, not by the steepness of the slope. **(c)** The first nine points cannot be spread from the top to the bottom of the graph because in such a case the correlation cannot exceed about 0.66 (based on empirical evidence—that is, from a reasonable amount of playing around with the applet). One possibility is shown below, left. **(d)** To have $r \doteq 0.7$, the curve must be higher at the right than at the left. One possibility is shown below, right.

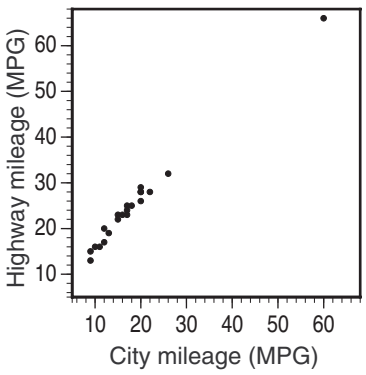






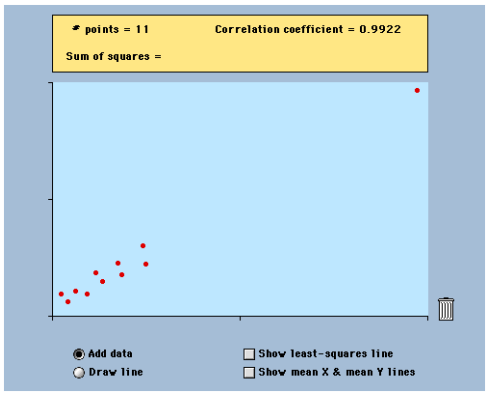
2.32. See the solution to Exercise 2.14 for the scatterplot. $r = -0.172$—it is close to zero, because the relationship is a curve rather than a line; correlation measures the degree of *linear* association.

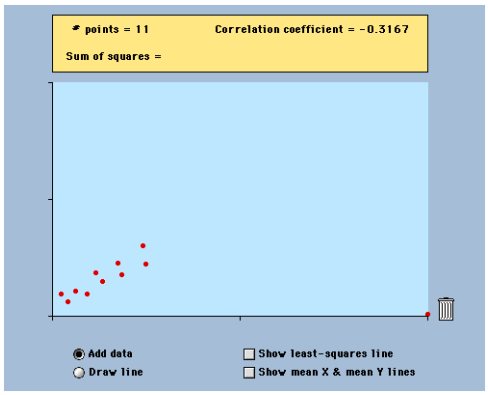
2.33. **(a)** The Insight seems to fit the line suggested by the other points. **(b)** Without the Insight, $r = 0.9757$; with it, $r^* = 0.9934$. The Insight increases the strength of the association (the line is the same, but the scatter about that line is *relatively* less when the Insight is included).



2.34. **(a)** The correlation will be closer to 1. One possible answer is shown below, left. **(b)** Answers will vary, but the correlation will decrease, and can be made negative by

dragging the point down far enough (see below, right).





2.35. (Scatterplot not shown.) If the husband's age is y and the wife's x, the linear relationship $y = x + 2$ would hold, and hence $r = 1$ (because the slope is positive).

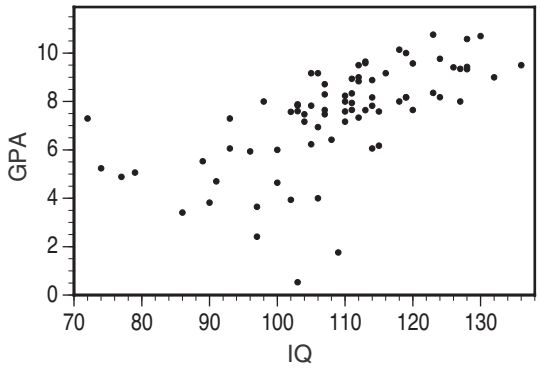
2.36. Explanations and sketches will vary, but should note that correlation measures the strength of the association, not the slope of the line. The hypothetical Funds A and B mentioned in the report, for example, might be related by a linear formula with slope 2 (or $1/2$).

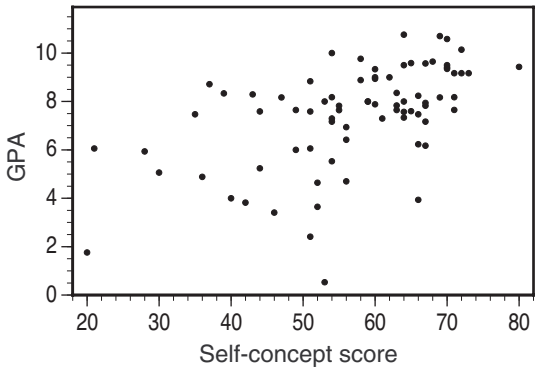
2.37. The person who wrote the article interpreted a correlation close to 0 as if it were a correlation close to -1 (implying a negative association between teaching ability and research productivity). Professor McDaniel's findings mean there is little linear association between research and teaching—for example, knowing that a professor is a good researcher gives little information about whether she is a good or bad teacher.

2.38. **(a)** Because gender has a categorical (nominal) scale, we cannot compute the correlation between sex and anything. (There is a strong *association* between gender and income. Some writers and speakers use "correlation" as a synonym for "association." It is much better to retain the more specific meaning.) **(b)** A correlation $r = 1.09$ is impossible because $-1 \le r \le 1$ always. **(c)** Correlation has no units, so $r = 0.23$ *bushel* is incorrect.

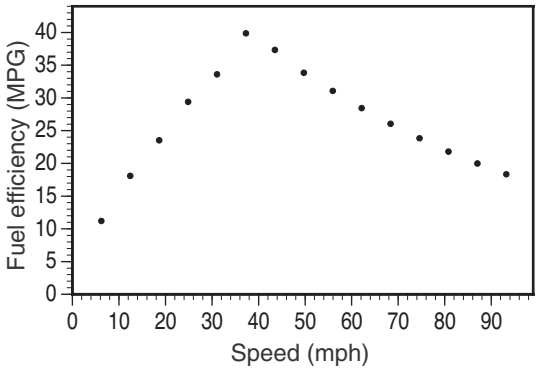
2.39. Both relationships (scatterplots below) are somewhat linear. The GPA/IQ scatterplot ($r = 0.6337$) shows a stronger association than GPA/self-concept ($r = 0.5418$). The two students with the lowest GPAs stand out in both plots; a few others stand out in at least one plot. Generally speaking, removing these points raises r (because the remaining points look

more linear). An exception: removing the lower-left point in the self-concept plot decreases r, because the relative scatter of the remaining points is greater.





2.40. **(a)** The new speed and fuel consumption (respectively) values are $x^* = x \div 1.609$ and $y^* = y \times 1.609 \div 100 \div 3.785 \doteq 0.004251y$. (The factor of $1/100$ is needed since we were measuring fuel consumption in liters/100 km.) The transformed data has the same correlation as the original—$r = -0.172$ (computed in the solution to Exercise 2.32)—since a linear transformation does not alter the correlation. The scatterplot of the transformed data is not shown here; it resembles (except for scale) the plot shown in the solution to Exercise 2.14. **(b)** The new correlation is $r^* = -0.043$; the new plot is even less linear than the first.



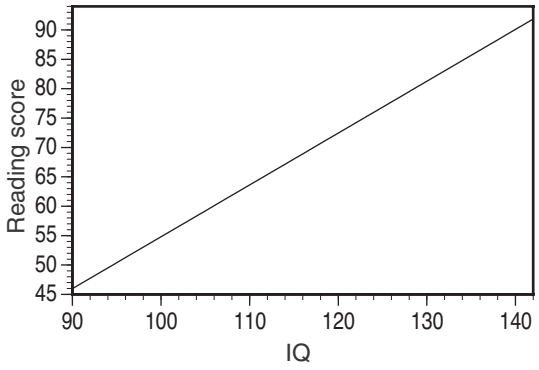
2.41. **(a)** The predicted scores for $x = 90$ and $x = 140$ are

$-33.4 + 0.882 \times 90 = 45.98$ and

$-33.4 + 0.882 \times 140 = 90.08$.

Plot the points $(90, 46)$ and $(140, 90)$ and draw the line connecting them. **(b)** If the reading score increases by 1 for each IQ point, the professor's line has slope 1. In order for an IQ of 100 to correspond to a reading score of 50, the equation must be

reading score = IQ − 50.

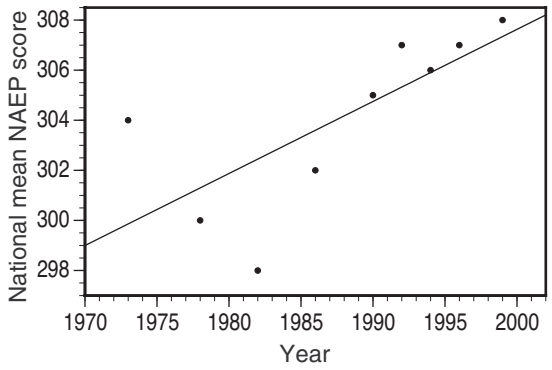


2.42. **(a)** The slope is 2.59, meaning that (on the average) team value rises 2.59 units (dollars, $million, or whatever) from each one-unit increase in revenue. (Most students may make this statement in terms of millions of dollars, as the table gives values with those units, but the ratio holds regardless of the unit, provided the same unit is used for both variables.) **(b)** The predicted value is $21.4 + 2.59 \times 149 = 407.31$ million dollars; the error is -39.69 million dollars. **(c)** The high correlation means that the line does a fairly good job of predicting value; specifically, the regression line explains about $r^2 \doteq 86\%$ of the variability in team value.

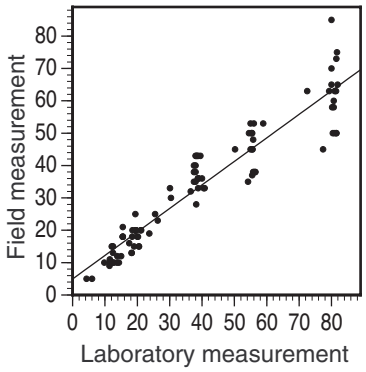
2.43. **(a)** The slope tells us that volume increases at an average rate of 4.2255 km^3/year. **(b)** The estimate for 1780 is -271 km^3; a negative number makes no sense in this context. **(c)** The estimate for 1990 is 617 km^3. Based on the time plot, it appears that the actual discharge in 1990 was around 680 km^3 (this is the value given in Table 1.4), so the prediction error is about 63 km^3. **(d)** There are high spikes in the time plot in the two flood years.

2.44. **(a)** Because the slope is 0.0086 (in units of "proportion of perch eaten per perch count"), an increase of 10 in the perch count increases the proportion eaten by 0.086 (on the average). **(b)** When the perch count is 0, the equation tells us that 12% (0.12) of those perch will be eaten. Of course, 12% of 0 is 0, so one could argue that this is in some sense correct, but computing the proportion eaten would require dividing by zero.

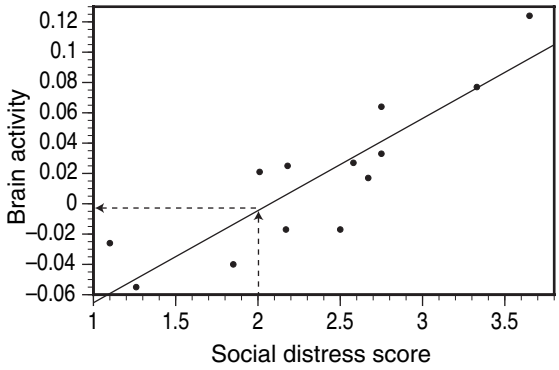
2.45. **(a)** Time plot shown on the right, along with the regression line. **(b)** The means and standard deviations are $\bar{x} = 1987.7778$, $\bar{y} = 304.1111$, $s_x = 8.7003$, and $s_y = 3.4440$. With the correlation $r = 0.7268$, the slope and intercept are $b = r\, s_y/s_x = 0.2877$ and $a = \bar{y} - b\bar{x} = -267.78$. The equation is therefore $\hat{y} = 0.2877x - 267.78$; this line explains about $r^2 \doteq 53\%$ of the variation in score.



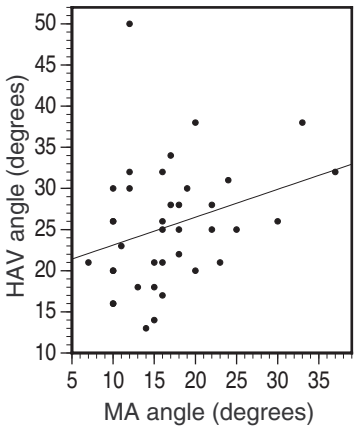
2.46. **(a)** The least-squares line is $\hat{y} = 0.7267x + 4.9433$. This is less steep than the line $y = x$, reflecting the observation that field measurements tend to be lower for greater depths. **(b)** The line $y = x$ has slope 1; the regression line has slope 0.7267. A slope of 1 would mean that for every additional unit of depth as measured in the laboratory, the field measurement would also increase by one unit. The slope of 0.7267 means that on the average, the field measurement increases by only 0.7267 units for every one unit in the lab.



2.47. See also the solution to Exercise 2.9. **(a)** The regression equation is $\hat{y} = 0.06078x - 0.1261$. **(b)** Based on the "up-and-over" method, most students will probably estimate that $\hat{y} \doteq 0$; the regression formula gives $\hat{y} = -0.0045$. **(c)** The correlation is $r \doteq 0.8782$, so the line explains $r^2 = 77\%$ of the variation in brain activity.

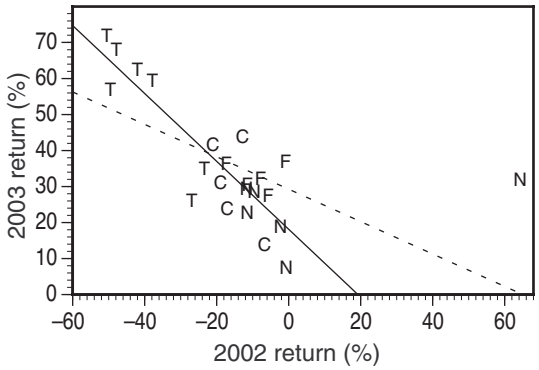


2.48. See also the solution to Exercise 2.12. **(a)** The regression line is $\hat{y} = 19.7 + 0.339x$. **(b)** For $x = 25^\circ$, we predict $\hat{y} = 28.2^\circ$. **(c)** The scatterplot shows a lot of spread, so predictions based on this line will not be very reliable. This is confirmed by the value of $r^2 = 9.1\%$; the straight-line relationship explains less than 10% of the variation in HAV angle.



2.49. The regression equations are $\hat{y} = -2.39 + 0.158x$ (Run 8903, 11.9 mg/s) and $\hat{y} = -1.45 + 0.0911x$ (Run 8905, 29.6 mg/s). Therefore, the growth rates are (respectively) 0.158 cm/minute and 0.0911 cm/minute; this suggests that the faster the water flows, the more slowly the icicles grow.

2.50. (a) For all the funds, $\hat{y} = 29.2512 - 0.4501x$ (the dashed line in the plot); with the outlier omitted, the equation is $\hat{y} = 18.1106 - 0.9429x$ (the solid line). As in the solution to Exercise 2.20, the scatterplot uses the letters C, F, T, and N to indicate the fund type. **(b)** Because the least-squares criterion attempts to minimize the total squared distances from points to the line, the point for Fidelity Gold Fund pulls the line toward it.



2.51. No, we could not predict stock returns accurately from Treasury bill returns: The scatterplot shows little or no association, and regression only explains 1.3% of the variation in stock return.

2.52. The means and standard deviations are $\bar{x} = 95$ min, $\bar{y} = 12.6611$ cm, $s_x = 53.3854$ min, and $s_y = 8.4967$ cm; the correlation is $r = 0.9958$.

For predicting length from time, the slope and intercept are $b_1 = r\,s_y/s_x \doteq 0.158$ cm/min and $a_1 = \bar{y} - b_1\bar{x} \doteq -2.39$ cm, giving the equation $\hat{y} = -2.39 + 0.158x$ (as in Exercise 2.48).

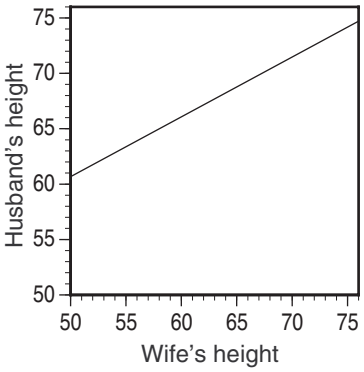
For predicting time from length, the slope and intercept are $b_2 = r\,s_x/s_y \doteq 6.26$ min/cm and $a_2 = \bar{x} - b_2\bar{y} \doteq 15.79$ min, giving the equation $\hat{x} = 15.79 + 6.26y$.

2.53. The means and standard deviations are: For lean body mass, $\bar{m} = 46.74$ and $s_m = 8.28$ kg, and for metabolic rate, $\bar{r} = 1369.5$ and $s_r = 257.5$ cal/day. The correlation is $r = 0.8647$. For predicting metabolic rate from body mass, the slope is $b_1 = r \cdot s_r/s_m \doteq 26.9$ cal/day per kg. For predicting body mass from metabolic rate, the slope is $b_2 = r \cdot s_m/s_r \doteq 0.0278$ kg per cal/day.

2.54. The correlation of IQ with GPA is $r_1 = 0.634$; for self-concept and GPA, $r_2 = 0.542$. IQ does a slightly better job; it explains about $r_1^2 = 40.2\%$ of the variation in GPA, while self-concept explains about $r_2^2 = 29.4\%$ of the variation.

2.55. Women's heights are the x values; men's are the y values. The slope is $b = (0.5)(2.7)/2.5 = 0.54$ and the intercept is $a = 68.5 - (0.54)(64.5) = 33.67$.

The regression equation is $\hat{y} = 33.67 + 0.54x$. Ideally, the scales should be the same on both axes. For a 67-inch tall wife, we predict the husband's height will be about 69.85 inches.



2.56. We have slope $b = r\,s_y/s_x$ and intercept $a = \bar{y} - b\bar{x}$, and $\hat{y} = a + bx$, so when $x = \bar{x}$,

$$\hat{y} = a + b\bar{x} = (\bar{y} - b\bar{x}) + b\bar{x} = \bar{y}.$$

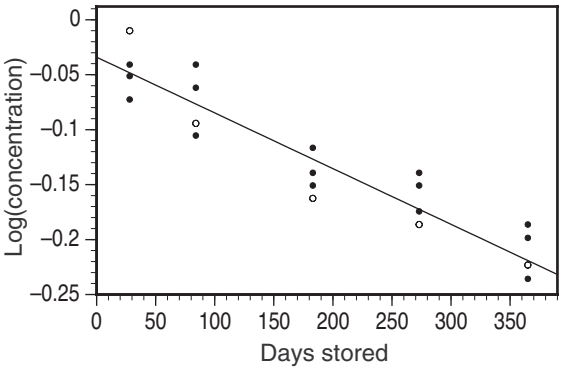
(Note that the value of the slope does not actually matter.)

2.57. **(a)** $\bar{x} = 95$ min, $s_x = 53.3854$ min, $\bar{y} = 12.6611$ cm, and $s_y = 8.4967$ cm. The correlation $r \doteq 0.9958$ has no units. **(b)** Multiply the old values of $\bar{y}$ and s_y by 2.54: $\bar{y} = 32.1591$ and $s_y = 21.5816$ inches. The correlation r is unchanged. **(c)** The slope is $r\,s_y/s_x$; with s_y from part (b), this gives 0.4025 in/min. (Or, multiply by 2.54 the appropriate slope from the solution to Exercise 2.49.)

2.58. **(a)** The slope is $b = r\,s_y/s_x = (0.6)(8)/(30) = 0.16$, and the intercept is $a = \bar{y} - b\bar{x} = 30.2$. **(b)** Julie's predicted score is $\hat{y} = 78.2$. **(c)** $r^2 = 0.36$; only 36% of the variability in y is accounted for by the regression, so the estimate $\hat{y} = 78.2$ could be quite different from the real score.

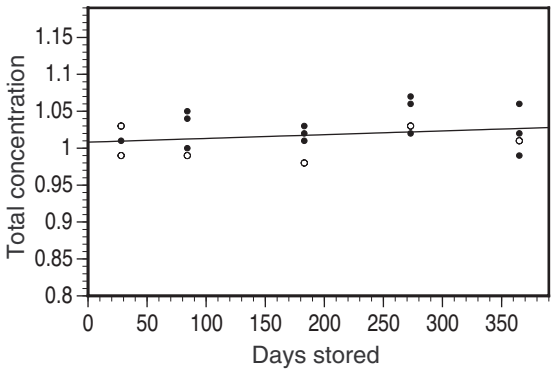
2.59. $r = \sqrt{0.16} = 0.40$ (high attendance goes with high grades, so the correlation must be positive).

2.60. **(a)** In the scatterplot (shown on the right), open circles represent two observations. This plot does suggest a linear association between days stored and the logarithm of the concentration, which supports the simple exponential decay model. **(b)** The regression equation is $\log C = -0.0341 - 0.0005068t$; we therefore estimate k to be 0.0005068.

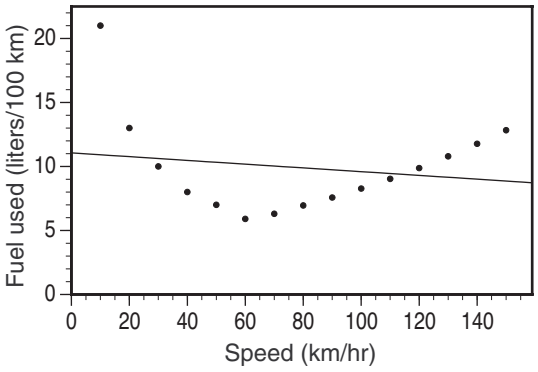


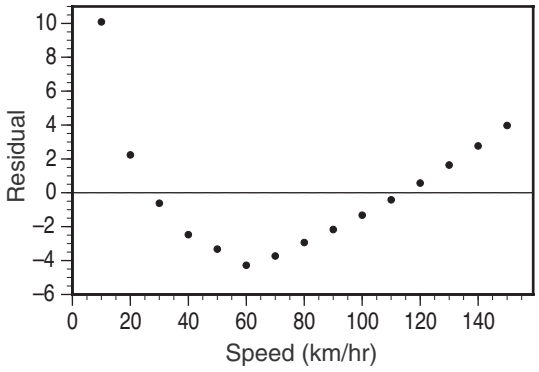
Note: *Students may need some help in performing this computation, especially in making sure that they compute the natural rather than the common logarithm. With most calculators and software, the correct function is "ln."*

2.61. **(a)** In the scatterplot on the right, open circles represent two observations. **(b)** The regression line slope is about 0.000051; the scatterplot suggests a nearly horizontal line (which would have slope 0). **(c)** Storing the oil doesn't help, as the total toxin level does not change over time; all that happens is the fenthion gradually changes to fenthion sulfoxide.

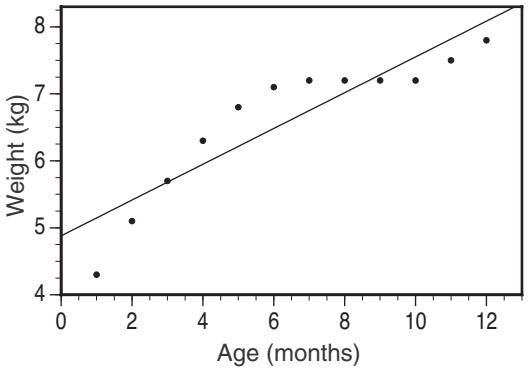


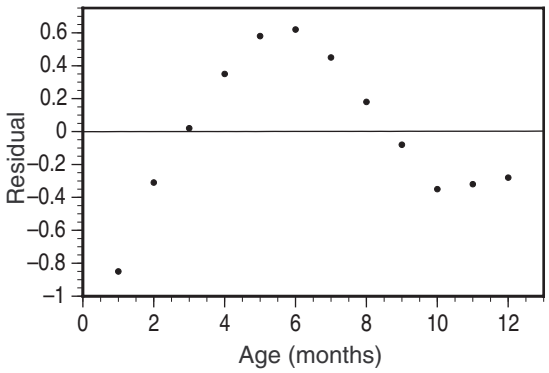
2.62. See also the solutions to Exercises 2.14 and 2.40. **(a)** Below, left. **(b)** The sum is −0.01. **(c)** The first two and last four residuals are positive, and those in the middle are negative. Plot below, right.



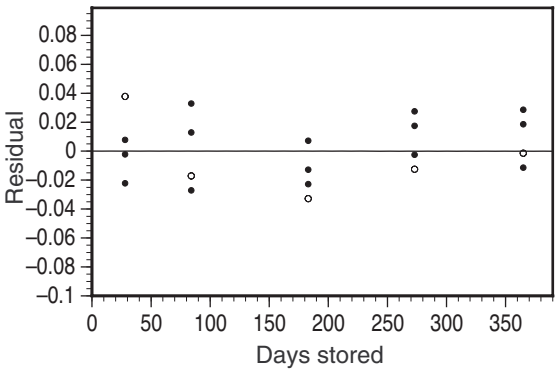


2.63. **(a)** Below, left. **(b)** This line is not a good summary of the pattern, because the pattern is curved rather than linear. **(c)** The sum is 0.01. The first two and last four residuals are negative, and those in the middle are positive. Plot below, right.





2.64. **(a)** The predicted concentration is $\hat{y} = 0.9524$, so the residual is $0.99 - \hat{y} = 0.0376$. **(b)** Rounding in the regression coefficients (slope and intercept) accounts for the difference between our residual (0.0376) and the value 0.0378 given in this list. The residuals do sum to 0. **(c)** In the residual plot, open circles represent two observations. There is a very slight curved pattern—high on the left and right, and low in the middle.



2.65. With individual children, the correlation would be smaller (closer to 0), because the additional variation of data from individuals would increase the "scatter" on the scatterplot, thus decreasing the strength of the relationship.

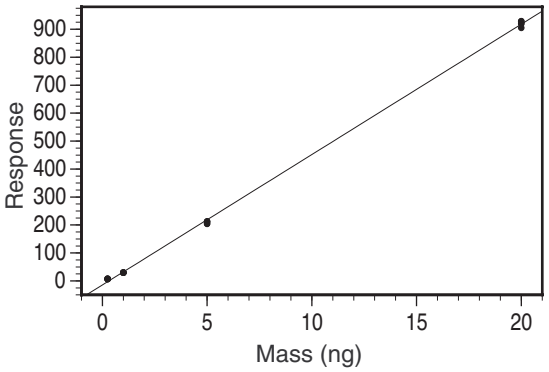
2.66. Presumably, those applicants who were hired would generally have been those who scored well on the test. As a result, we have little or no information on the job performance of those who scored poorly (and were therefore not hired). Those with higher test scores (who were hired) will likely have a range of performance ratings, so we will only see the various ratings for those with high scores, which will almost certainly show a weaker relationship than if we had performance ratings for all applicants.

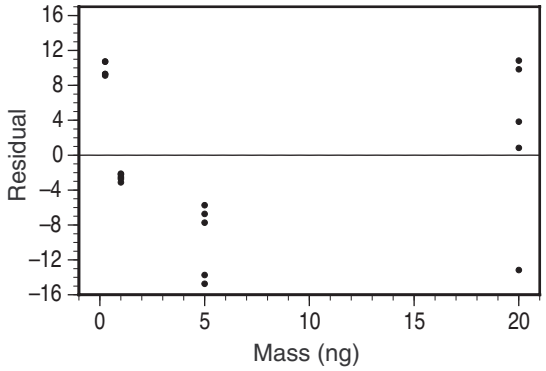
2.67. For example, a student who in the past might have received a grade of B (and a lower SAT score) now receives an A (but has a lower SAT score than an A student in the past). While this is a bit of an oversimplification, this means that today's A students are yesterday's A and B students, today's B students are yesterday's C students, and so on. Because of the grade inflation, we are not comparing students with equal abilities in the past and today.

2.68. A simple example illustrates this nicely: Suppose that everyone's current salary is their age (in thousands of dollars); for example, a 52-year-old worker makes \$52,000 per year. Everyone receives a \$500 raise each year. That means that in two years, every worker's income has increased by \$1000, but their age has increased by 2, so each worker's salary is not their age minus 1 (thousand dollars).

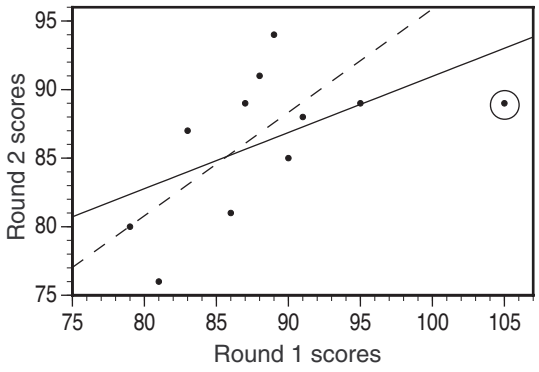
2.69. The correlation between BMR and fat gain is $r = 0.08795$; the slope of the regression line is $b = 0.000811$ kg/cal. These both show that BMR is less useful for predicting fat gain. The small correlation suggests a very weak linear relationship (explaining less than 1% of the variation in fat gain). The small slope means that changes in BMR have very little impact on fat gain; for example, increasing BMR by 100 calories changes fat gain by only 0.08 kg.

2.70. (a) The scatterplot of the data is below on the left. **(b)** The regression equation is $\hat{y} = -14.4 + 46.6x$. **(c)** Residual plot below, right. The residuals for the extreme x-values ($x = 0.25$ and $x = 20.0$) are almost all positive; all those for the middle two x values are negative.

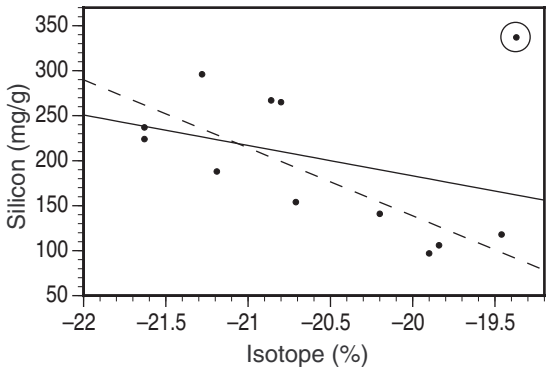




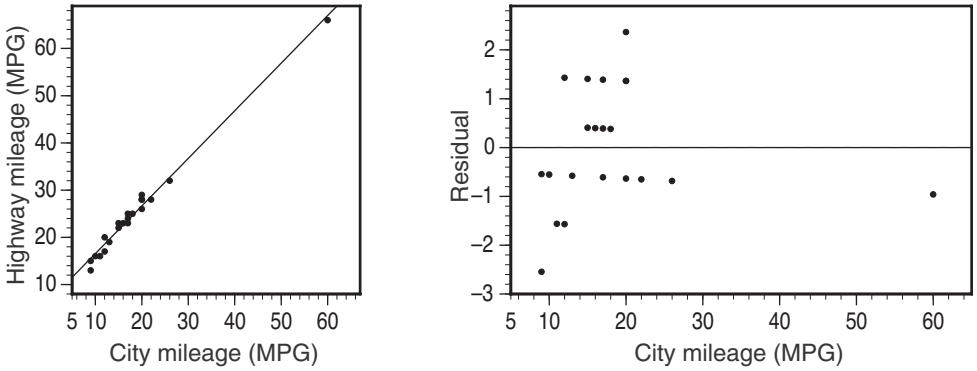
2.71. (a) There is a moderate positive relationship; player 7's point is an outlier. Ideally, both scales should be equal. **(b)** The first equation is the dashed line in the plot. It omits the influential observation; the other (solid) line is pulled toward the outlier.



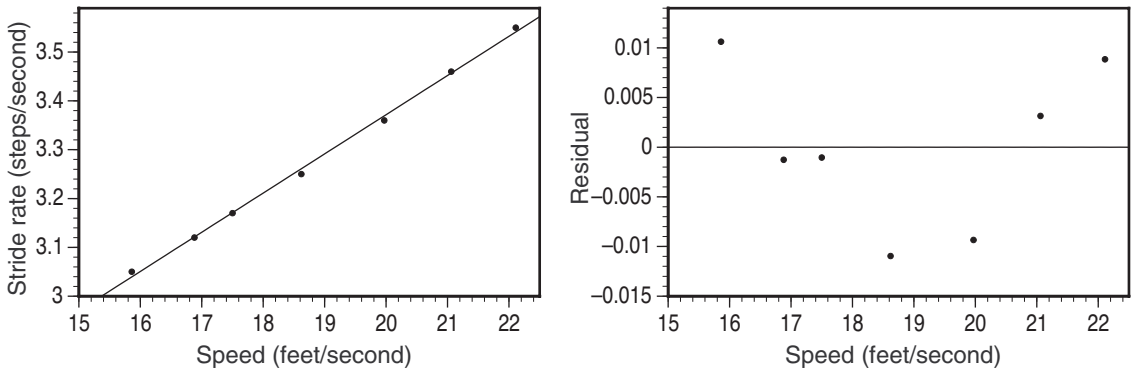
2.72. (a) Apart from the outlier—circled for part (b)—the scatterplot shows a moderate linear negative association. **(b)** With the outlier, $r = -0.3387$; without it, $r^* = -0.7866$. **(c)** The two regression formulas are $\hat{y} = -492.6 - 33.79x$ (the solid line, with all points), and $\hat{y} = -1371.6 - 75.52x$ (the dashed line, with the outlier omitted). The omitted point is also influential, as it has a noticeable impact on the line.



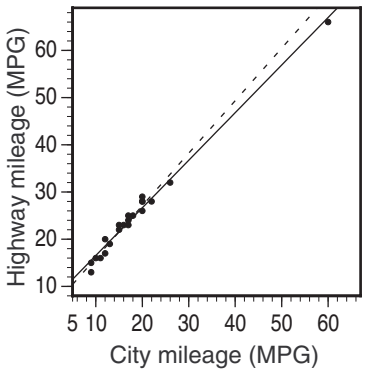
2.73. **(a)** Scatterplot below on the left. **(b)** The regression line is $\hat{y} = 6.47 + 1.01x$. The residual plot is below on the right. **(c)** The largest residuals are the Porsche Boxster (2.365) and Lamborghini Murcielago (-2.545). **(d)** The Insight is influential; it pulls the line toward its point, so that it not far from the regression line.



2.74. **(a)** Scatterplot below, left. The relationship seems linear. **(b)** The regression line is $\hat{y} = 1.77 + 0.0803x$ (y is stride rate, x is speed). **(c)** The residuals (reported by Minitab, then rounded to 3 decimal places) are 0.011, -0.001, -0.001, -0.011, -0.009, 0.003, 0.009. These add to 0.001. Results will vary with rounding, and also with the number of decimal places used in the regression equation. **(d)** Residuals are positive for low and high speeds, negative for moderate speeds; this suggests that a curve (like a parabola) may be a better fit. No observations are particularly influential; the line would change very little if we omitted any point.

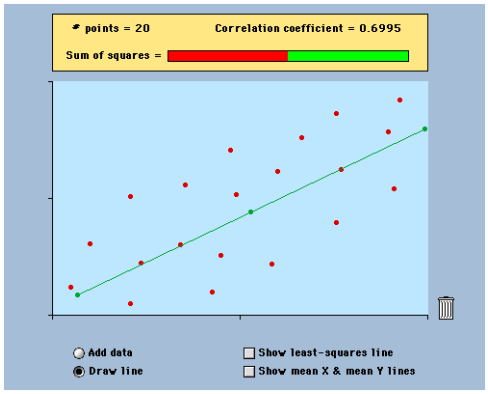


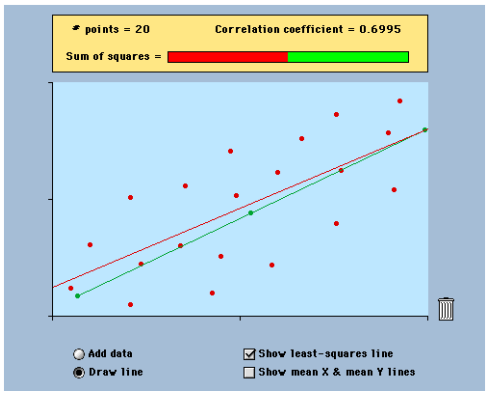
2.75. Without the Insight, $\hat{y} = 4.87 + 1.11x$ (the dashed line in the plot). For city mileages between 10 and 30 MPG, the difference in predicted highway mileage (with or without the Insight) is no more than 1.4 MPG, so the Insight is not very influential; it falls near the line suggested by the other points.



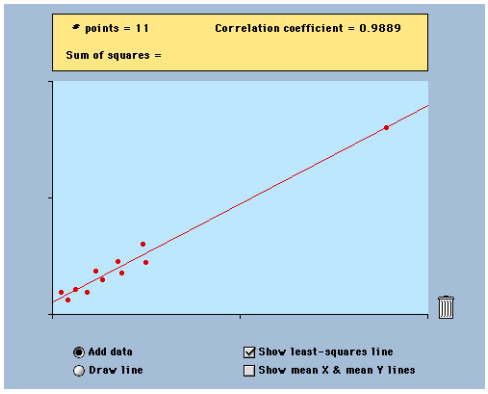
2.76. The correlation is $r = 0.999$. With individual runners, the correlation would be smaller (closer to 0), since using data from individual runners would increase the "scatter" on the scatterplot, thus decreasing the strength of the relationship.

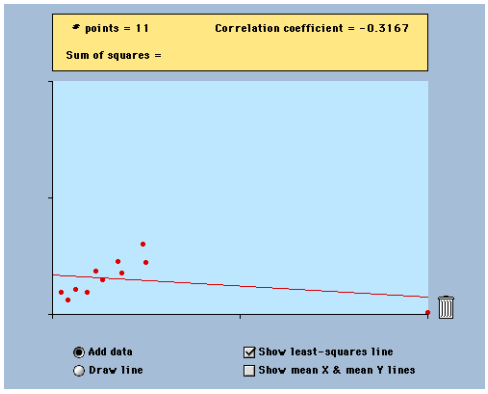
2.77. **(a)** Drawing the "best line" by eye is a very inaccurate process; few people choose the best line (although you can get better at it with practice). **(b)** Most people tend to overestimate the slope for a scatterplot with $r \doteq 0.7$; that is, most students will find that the least-squares line is less steep than the one they draw.



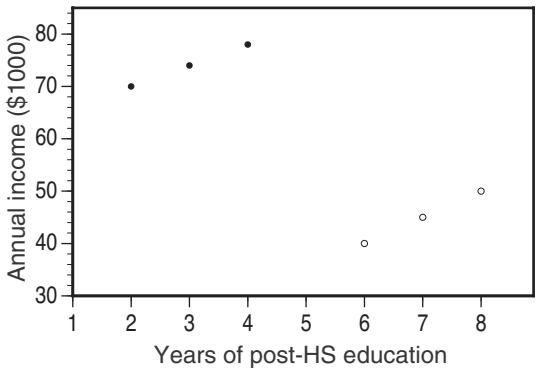


2.78. **(a)** Any point that falls exactly on the regression line will not increase the sum of squared vertical distances (which the regression line minimizes). Any other line—even if it passes through this new point—will necessarily have a higher total sum of squares. Thus, the regression line does not change. Possible output below, left. **(b)** Influential points are those whose x coordinates are outliers; this point is on the right side, while all others are on the left. Possible output below, right.

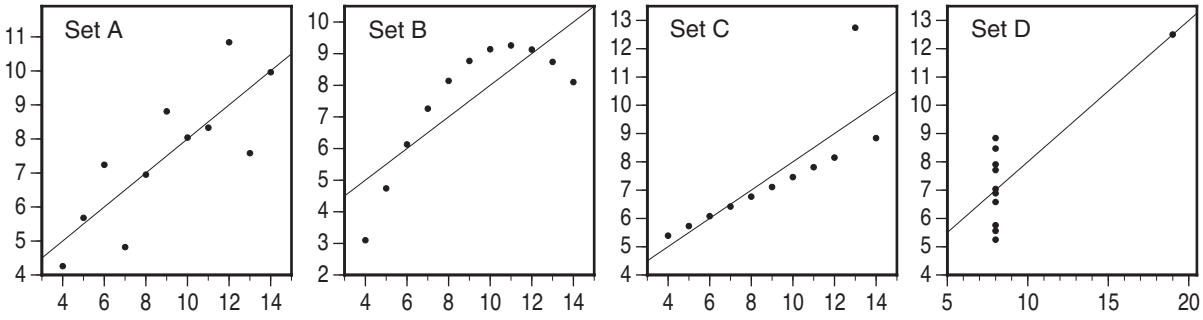




2.79. The plot shown is a very simplified (and not very realistic) example. Filled circles are economists in business; open circles are teaching economists. The plot should show positive association when either set of circles is viewed separately, and should show a large number of bachelor's degree economists in business and graduate degree economists in academia.

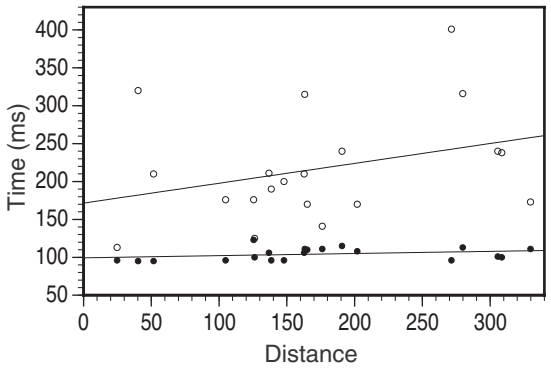


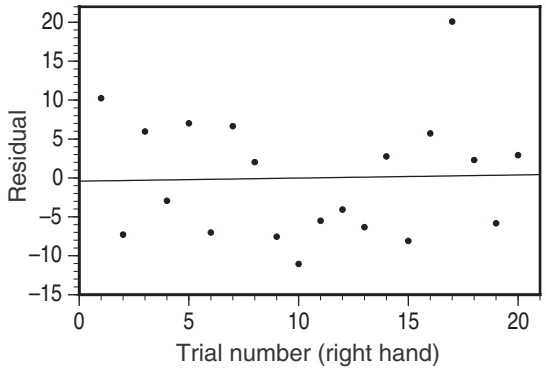
2.80. **(a)** To three decimal places, the correlations are all approximately 0.816 (for set D, r actually rounds to 0.817), and the regression lines are all approximately $\hat{y} = 3.000 + 0.500x$. For all four sets, we predict $\hat{y} \doteq 8$ when $x = 10$. **(b)** Below. **(c)** For Set A, the use of the regression line seems to be reasonable—the data do seem to have a moderate linear association (albeit with a fair amount of scatter). For Set B, there is an obvious *non*linear relationship; we should fit a parabola or other curve. For Set C, the point (13, 12.74) deviates from the (highly linear) pattern of the other points; if we can exclude it, the (new) regression formula would be very useful for prediction. For Set D, the data point with $x = 19$ is a very influential point—the other points alone give no indication of slope for the line. Seeing how widely scattered the y-coordinates of the other points are, we cannot place too much faith in the y-coordinate of the influential point; thus, we cannot depend on the slope of the line, and so we cannot depend on the estimate when $x = 10$. (We also have no evidence as to whether or not a line is an appropriate model for this relationship.)

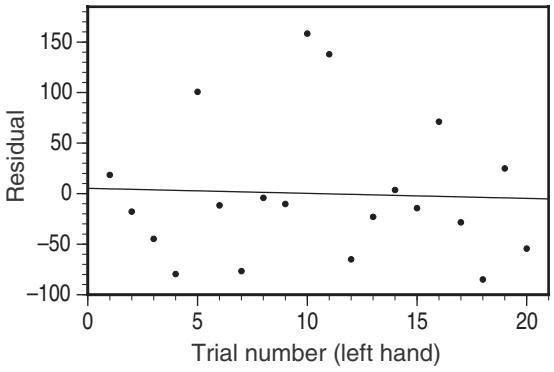


2.81. **(a)** As the slope of the line is negative, there is some support for this idea, but the relationship is quite weak. **(b)** There seems to be some suggestion of curvature, and there is considerably more scatter on the left side. This weakens the conclusion.

2.82. **(a)** Right-hand points are filled circles; left-hand points are open circles. **(b)** The right-hand points lie below the left-hand points. (This means the right-hand times are shorter, so the subject is right-handed.) There is no striking pattern for the left-hand points; the pattern for right-hand points is obscured because they are squeezed at the bottom of the plot. **(c)** Right hand: $\hat{y} = 99.4 + 0.0283x$ ($r = 0.305$, $r^2 = 9.3\%$). Left hand: $\hat{y} = 172 + 0.262x$ ($r = 0.318$, $r^2 = 10.1\%$). The left-hand regression is slightly better, but neither is very good: distance accounts for only 9.3% (right) and 10.1% (left) of the variation in time. **(d)** The two residual plots are shown below; neither shows a systematic pattern.

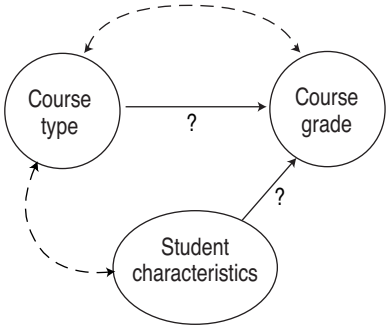




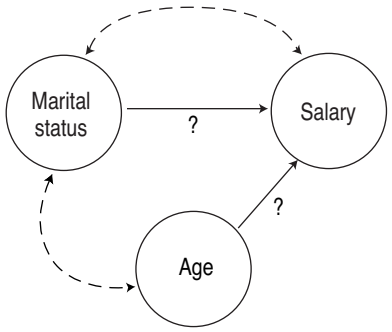


2.83. **(a)** There is clearly higher scatter for higher predicted values; the regression more accurately predicts low salaries than high salaries. **(b)** The residual plot is curved. Salaries are typically overestimated for players who are new to the majors, and for those who have been in the majors for 15 or more years (these residuals are mostly negative, meaning the predicted value is greater than the observed value). Those in for eight years will generally have their salaries underestimated; these residuals are mostly positive.

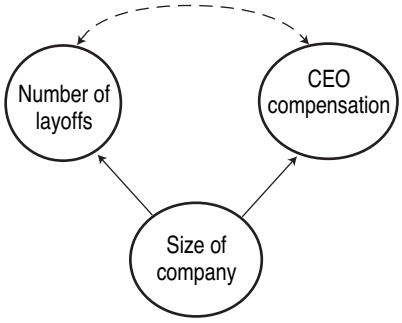
2.84. Responses will vary. For example, students who choose the online course might have more self-motivation, or have better computer skills. A diagram is shown on the right; the generic "Student characteristics" might be replaced with something more specific.



2.85. Age is one lurking variable: Married men would generally be older than single men, so they would have been in the work force longer, and therefore had more time to advance in their careers. The diagram shown on the right shows this lurking variable; other variables could also be shown in place of "age."



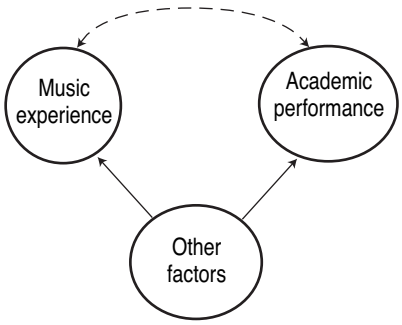
2.86. A large company has more workers who might be laid off, and often pays more to its CEO (because, presumably, there is more work involved in running a large company than a small one). Smaller companies typically pay less, and have fewer workers to lay off.



2.87. No: Self-confidence and improving fitness could be a common response to some other personality trait, or high self-confidence could make a person more likely to join the exercise program.

2.88. If a nation's population has high income, they have more money to spend on things that can help to keep them healthy: health care, medicine, better food, better sanitation, and so on. On the other hand, if a nation's population is healthy, they can spend less on health care and instead put their money to more productive uses. Additionally, they miss fewer work days, so they would typically earn more money.

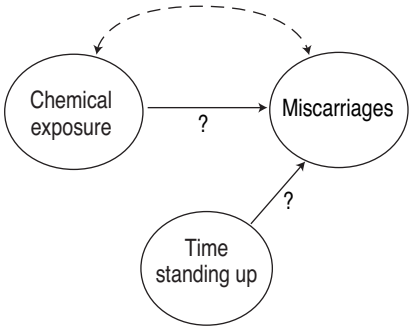
2.89. Students with music experience may have other advantages (wealthier parents, better school systems, and so forth.). That is, experience with music may have been a "symptom" (common response) of some other factor that also tends to cause high grades.



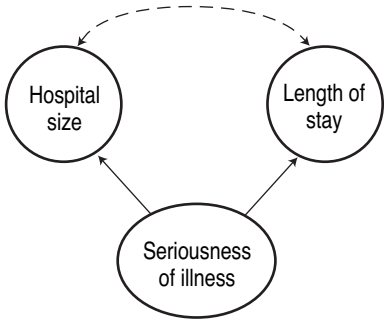
2.90. Two possibilities are that they might perform better simply because this is their second attempt, or because they feel better prepared as a result of taking the course (whether or not they really *are* better prepared).

2.91. The diagram below illustrates the confounding between exposure to chemicals and standing up.

For 2.91.



For 2.92.



2.92. Patients suffering from more serious illnesses are more likely to go to larger hospitals (which may have more or better facilities) for treatment. They are also likely to require more time to recuperate afterwards.

2.93. Spending more time watching TV means that *less* time is spent on other activities; this may suggest lurking variables. For example, perhaps the parents of heavy TV watchers do not spend as much time at home as other parents. Also, heavy TV watchers would typically not get as much exercise.

2.94. In this case, there may be a causative effect, but in the direction opposite to the one suggested: People who are overweight are more likely to be on diets, and so choose artificial sweeteners over sugar. (Also, heavier people are at a higher risk to develop diabetes; if they do, they are likely to switch to artificial sweeteners.)

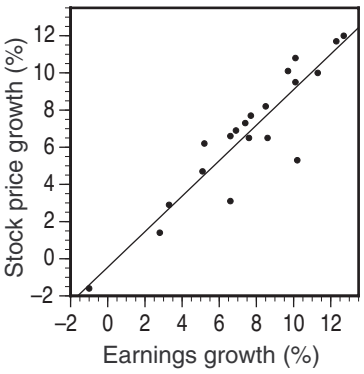
2.95. (a) Statements such as this typically mean that the risk of dying at a given age is half as great; that is, given two groups of the same age, where one group walks and the other does not, the walkers are half as likely to die in (say) the next year. **(b)** Men who choose to walk might also choose (or have chosen, earlier in life) other habits and behaviors that reduce mortality.

2.96. A reasonable explanation is that the cause-and-effect relationship goes in the other direction: Doing well makes students or workers feel good about themselves, rather than vice versa.

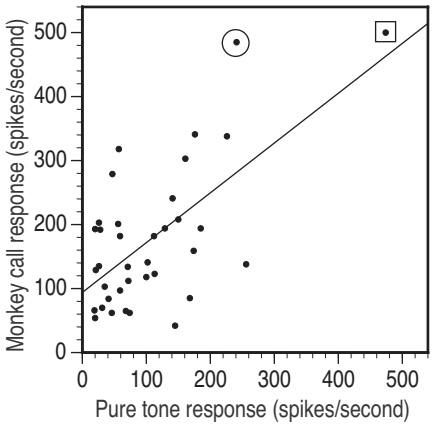
2.97. These results support the idea (the slope is negative, so variation decreases with increasing diversity), but the relationship is only moderately strong ($r^2 = 0.34$, so diversity only explains 34% of the variation in population variation).

Note: *That last parenthetical comment is awkward and perhaps confusing, but is consistent with similar statements interpreting* r^2.

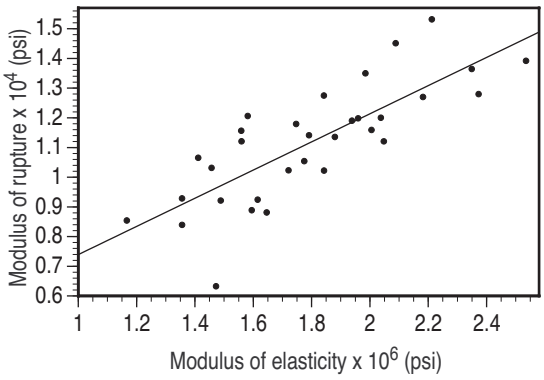
2.98. **(a)** A scatterplot of stock price growth against earnings growth shows a positive association, which supports the idea. Additionally, each y value is fairly similar to its x value, which indicates that stock price growth is roughly predicted by earnings growth (that is, $\hat{y} \approx x$)—this is a stronger statement than simply saying that the two variables have a positive association. **(b)** The regression explains $r^2 = 0.846 = 84.6\%$ of the variation in stock price growth. **(c)** The slope would be 1 (and the equation would be $\hat{y} = x$) because "stock prices exactly follow[ing] earnings" means that stock prices would change (increase or decrease) in exactly the same way that earnings change. The actual slope is 0.9552 (the full regression equation is $\hat{y} = 0.9552x - 0.4551$). **(d)** The correlation is $r = 0.9198$. With data from individual companies, the correlation would be much lower, because the additional variation of data from individuals would increase the "scatter" on the scatterplot, thus decreasing the strength of the relationship.



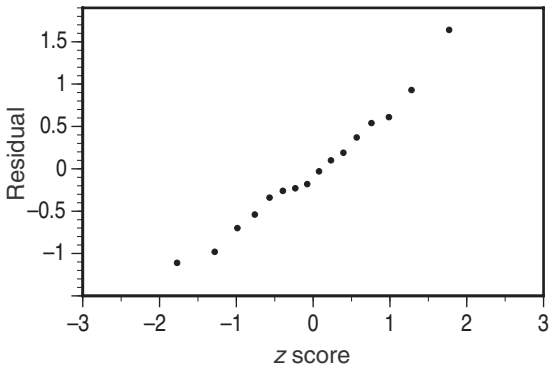
2.99. **(a)** One possible measure of the difference is the mean response: 106.2 spikes/second for pure tones and 176.6 spikes/second for monkey calls—an average of an additional 70.4 spikes/second. **(b)** The regression equation is $\hat{y} = 93.9 + 0.778x$. The third point (pure tone 241, call 485 spikes/second) has the largest residual; it is circled. The first point (474 and 500 spikes/second) is an outlier in the x direction; it is marked with a square. **(c)** The correlation drops only slightly (from 0.6386 to 0.6101) when the third point is removed; it drops more drastically (to 0.4793) without the first point. **(d)** Without the first point, the line is $\hat{y} = 101 + 0.693x$; without the third point, it is $\hat{y} = 98.4 + 0.679x$.



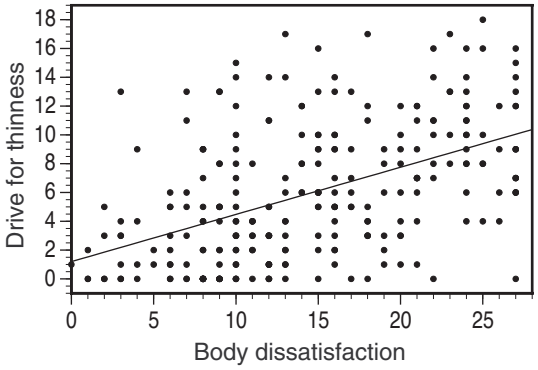
2.100. On the right is a scatterplot of MOR against MOE, showing a moderate, linear, positive association. The regression equation is $\hat{y} = 2653 + 0.004742x$; this regression explains $r^2 = 0.6217 \doteq 62\%$ of the variation in MOR. So, we can use MOE to get fairly good (though not perfect) predictions of MOR.



2.101. The quantile plot (right) is reasonably close to a straight line, so we have little reason to doubt that they come from a normal distribution.



2.102. **(a)** The scatterplot is on the right. **(b)** The regression equation is $\hat{y} = 1.2027 + 0.3275x$. As we see from the scatterplot, the relationship is not too strong; the correlation ($r = 0.4916$, $r^2 = 0.2417$) confirms this.

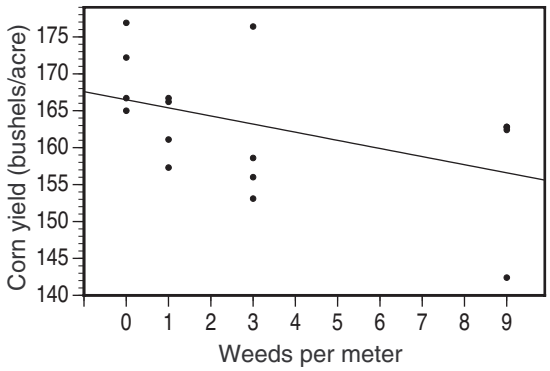


2.103. **(a)** Yes: The two lines appear to fit the data well. There do not appear to be any outliers or influential points. **(b)** Compare the slopes: before—0.189; after—0.157. (The units for these slopes are 100 ft^3/day per degree-day/day; for students who are comfortable with units, 18.9 ft^3 vs. 15.7 ft^3 would be a better answer.) **(c)** Before: $\hat{y} = 1.089 + 0.189(35) = 7.704 = 770.4$ ft^3. After: $\hat{y} = 0.853 + 0.157(35) = 6.348 = 634.8$ ft^3. **(d)** This amounts to an additional $(\$1.20)(7.704 - 6.348) = \$1.63$ per day, or \$50.44 for the month.

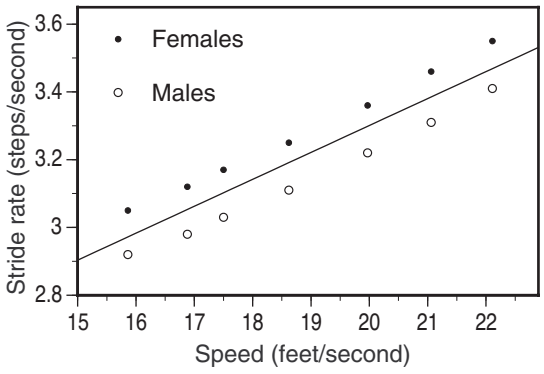
2.104. **(a)** $b = r \cdot s_y/s_x \doteq 1.1694$; $a = \bar{y} - b\bar{x} \doteq 0.3531$. The regression equation is $\hat{y} = 0.3531 + 1.1694x$; it explains $r^2 \doteq 27.6\%$ of the volatility in Philip Morris stock. **(b)** On the average, for every percentage-point rise in the S&P monthly return, Philip Morris stock returns rise about 1.17 percentage points. (And similarly, Philip Morris returns fall 1.17% for each 1% drop in the S&P index return.) **(c)** When the market is rising, the investor would like to earn money faster than the prevailing rate, and so prefers beta > 1. When the market falls, returns on stocks with beta < 1 will drop more slowly than the prevailing rate.

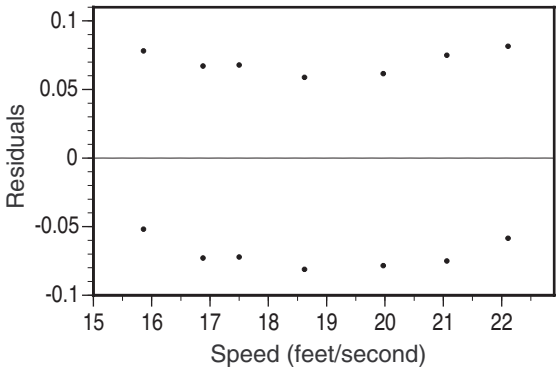
2.105. **(a)** Explanatory: weeds per meter (wpm). Response: corn yield. **(b)** The stemplots (below) give some evidence that yield decreases when there are more lamb's-quarter plants. **(c)** Scatterplot below. The regression equation is $\hat{y} = 166 - 1.10x$. Each additional lamb's-quarter per meter decreases yield by about 1.1 bushels/acre. **(d)** $\hat{y} = 166 - 1.10(6) = 159.4$ bushels/acre.

0 wpm		1 wpm		3 wpm		9 wpm	
14		14		14		14	2
14		14		14		14	
15		15		15	3	15	
15		15	7	15	69	15	
16		16	1	16		16	233
16	57	16	67	16		16	
17	2	17		17		17	
17	7	17		17	6	17	

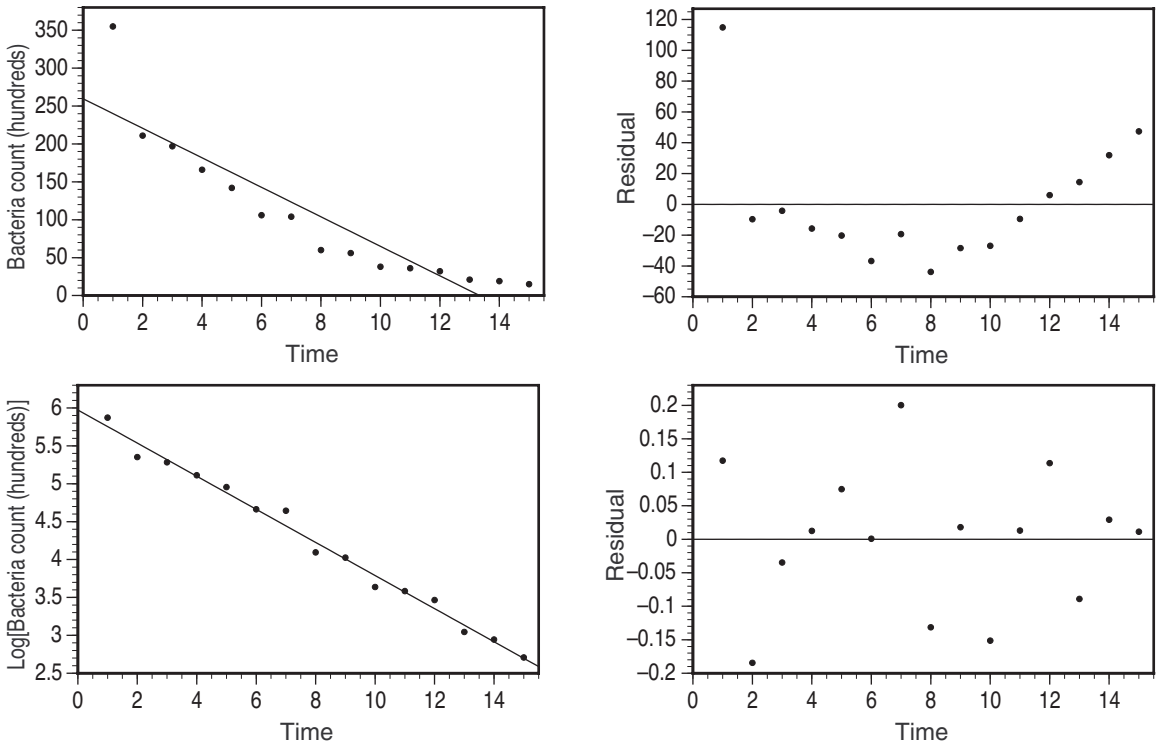


2.106. **(a)** Below, left. **(b)** The regression equation is $\hat{y} = 1.71 + 0.0795x$. **(c)** Below, right. The points for the residuals, like those of the original data, are split with women above the line (zero), and men below. (Men are taller on the average, so they have longer legs, and therefore longer strides. Thus, they need fewer steps per second to run at a given speed.)

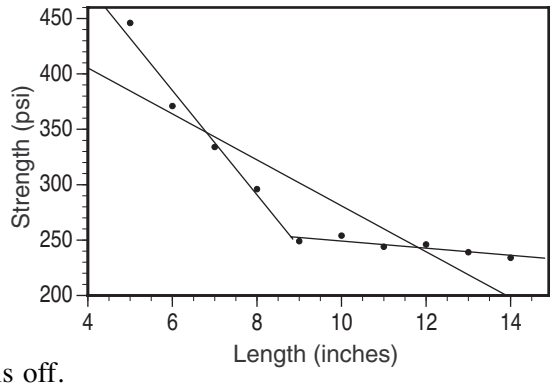




2.107. (a) Shown below are plots of count against time, and residuals against time for the regression, which gives the formula $\hat{y} = 259.58 - 19.464x$. Both plots suggest a curved relationship rather than a linear one. **(b)** With natural logarithms, the regression equation is $\hat{y} = 5.9732 - 0.2184x$; with common logarithms, $\hat{y} = 2.5941 - 0.09486x$. The second pair of plots below show the (natural) logarithm of the counts against time, suggesting a fairly linear relationship, and the residuals against time, which shows no systematic pattern. (If common logarithms are used instead of natural logs, the plots will look the same, except the vertical scales will be different.) The correlations confirm the increased linearity of the log plot: $r^2 = 0.8234$ for the original data, $r^2 = 0.9884$ for the log-data.



2.108. (a) At right. **(b)** The plot shows a negative association (longer beams are less strong), with no outliers. **(c)** The regression equation is $\hat{y} = 488 - 20.7x$; it is not a good match because the scatterplot does not suggest a straight line. **(d)** Length 5 to 9 inches: $\hat{y} = 668 - 46.9x$. Length 9 to 14 inches: $\hat{y} = 283 - 3.37x$. These two lines together describe the data fairly well. One might ask why strength at first decreases so rapidly with increasing length, then almost levels off.



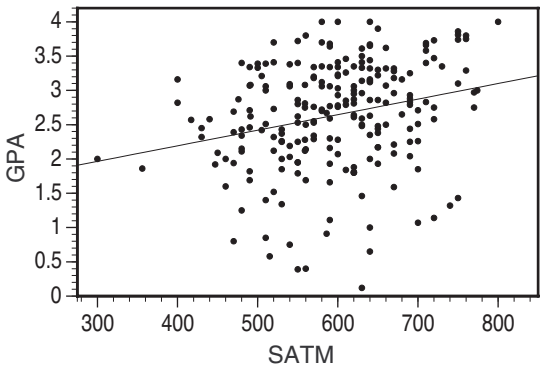
2.109. In the mid-1990s, European and American stocks were only weakly linked, but now it is more common for them to rise and fall together. Thus investing in both types of stocks is not that much different from investing in either type alone.

2.110. The article is incorrect; a correlation of 0.8 means that a straight-line relationship explains about $r^2 = 64\%$ of the variation of European stock prices.

2.111. Number of firefighters and amount of damage both increase with the seriousness of the fire (that is, they are common responses to the fire's seriousness.)

2.112. Note that $\bar{y} = 46.6 + 0.41\bar{x}$. We predict that Octavio will score 4.1 points above the mean on the final exam: $\hat{y} = 46.6 + 0.41(\bar{x} + 10) = 46.6 + 0.41\bar{x} + 4.1 = \bar{y} + 4.1$. (Alternatively, because the slope is 0.41, we can observe that an increase of 10 points on the midterm yields an increase of 4.1 on the predicted final exam score.)

2.113. The scatterplot is not very promising. The regression equation is $\hat{y} = 1.28 + 0.00227x$; the correlation is $r = 0.252$, and the regression explains $r^2 = 6.3\%$ of the variation in GPA. By itself, SATM does not give reliable predictions of GPA.



For answers to the EESEE Case Studies (exercises 114–117), see the instructor's version of EESEE.