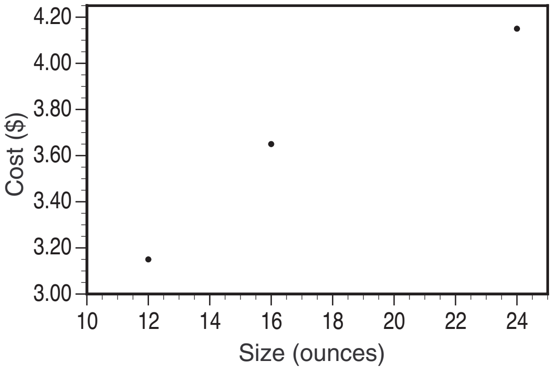
Chapter 2 Solutions

2.1. The individuals are students.

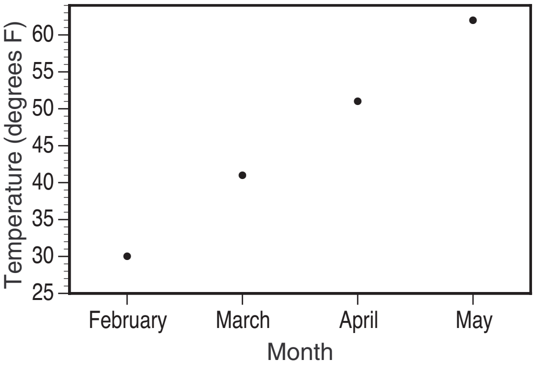
2.2. With this change, the cases are dog breeds; the variables (both quantitative) are breed size and average life span.

2.3. With this change, the cases are cups of Mocha Frappuccino (as before). The variables (both quantitative) are size and price.

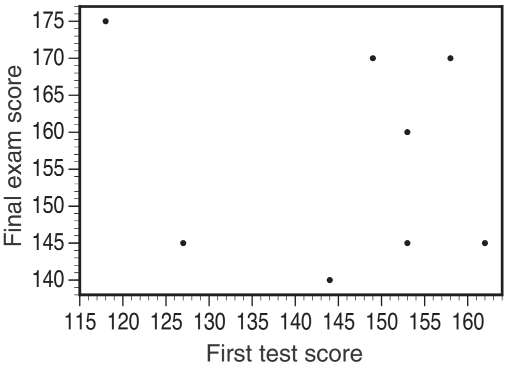
2.4. Size seems to be the most reasonable choice for explanatory variable because it seems nearly certain that Starbucks first decided which sizes to offer, then determined the appropriate price for each size (rather than vice versa). The scatterplot shows a positive association between size and price.



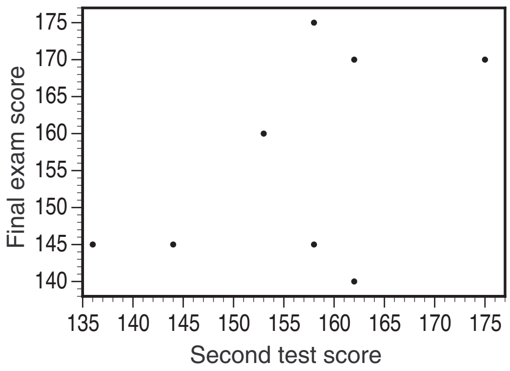
2.5. **(a)** "Month" (the passage of time) explains changes in temperature (not vice versa). **(b)** Temperature increases linearly with time (about 10 degrees per month); the relationship is strong.



2.6. **(a)** First test score should be explanatory since it comes first chronologically. **(b)** The scatterplot shows no clear association; however, the removal of one point (the sixth student, in the upper left corner of the scatterplot) leaves a weak-to-moderate positive association. **(c)** A few students can disrupt the pattern quite a bit; for example, perhaps the sixth student studied very hard after scoring so low on the first test, while some of those who did extremely well on the first exam became overconfident and did not study hard enough for the final (the points in the lower right corner of the scatterplot).



2.7. **(a)** The second test happens before the final exam, so that score should be viewed as explanatory. **(b)** The scatterplot shows a weak positive association. **(c)** Students' study habits are more established by the middle of the term.



2.8. To be considered an outlier, the point for the ninth student should be in either the upper left or lower right portion of the scatterplot. The former would correspond to a student who had a below-average second-test score but an above-average final-exam score. The latter would be a student who did well on the second test but poorly on the final.

2.9. **(a)** Age is explanatory; weight is the response variable. **(b)** Explore the relationship; there is no reason to view one or the other as explanatory. **(c)** Number of bedrooms is explanatory; price is the response variable. **(d)** Amount of sugar is explanatory; sweetness is the response variable. **(e)** Explore the relationship.

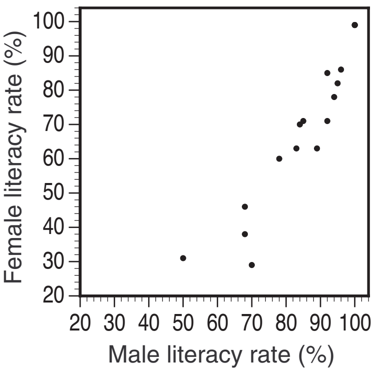
2.10. Parents' income is explanatory, and college debt is the response. Both variables are quantitative. We would expect a negative association: Low income goes with high debt, high income with low debt.

2.11. **(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).

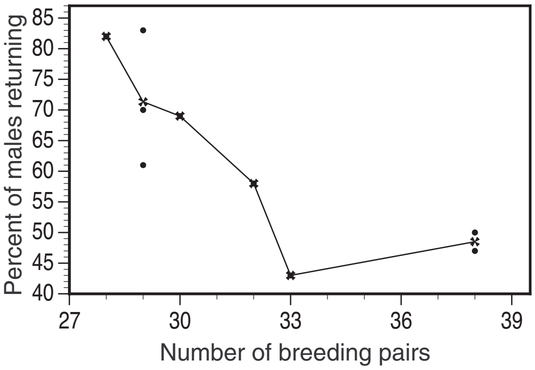
2.12. **(a)** From the scatterplot, we estimate 50% in 1954 and about −28% in 1974. (The data file `ex01-144.dat` gives the values 50.28% and −27.87%.) **(b)** The return on Treasury bills in 1981 was about 14.8%. **(c)** The scatterplot shows no clear pattern. (The statement that "high treasury bill 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.)

2.13. **(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.

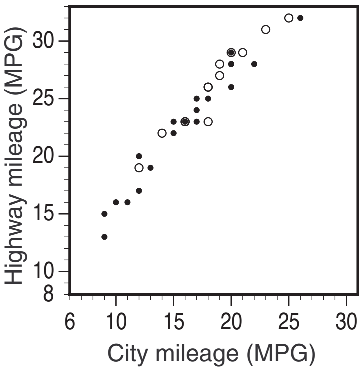
2.14. Ideally, the scales should be the same on both axes. The scatterplot shows a fairly strong, positive, linear association. Three countries (Tajikistan, Kazakhstan, and Uzbekistan) reported 100% literacy for men and 99% literacy for women. Yemen (70% for men, 29% for women) might be considered an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



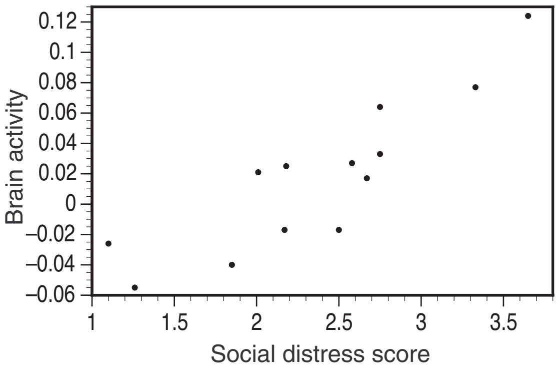
2.15. **(a)** If we used the number of males returning, then we might not see the relationship because areas with many breeding pairs would correspondingly have more males that might potentially return. (In the given numbers, the number of breeding pairs varies only from 28 to 38, but considering hypothetical 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.



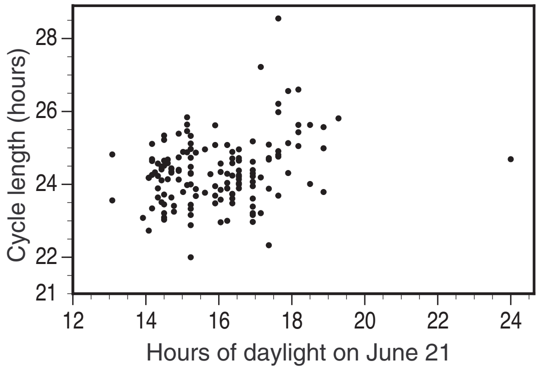
2.16. **(a)** Two-seater cars are shown as filled circles, minicompact cars as open circles. Ideally, the scales should be the same on both axes. **(b)** The scatterplot shows a strong, positive, linear association. Two-seater cars include several vehicles with poor fuel efficiency (most notably, the 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.



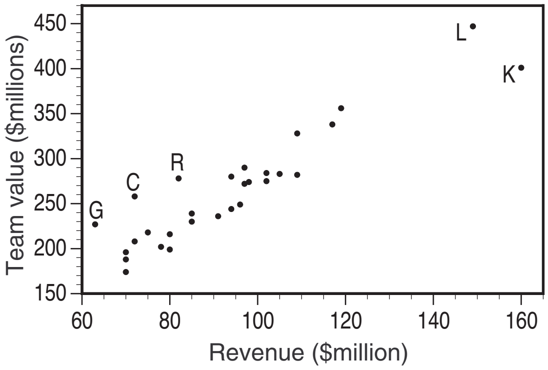
2.17. The scatterplot shows a fairly strong, positive, linear association. There are no particular outliers; each variable has low and high values, but those points do not deviate from the pattern of the rest. Social exclusion does appear to trigger a pain response.

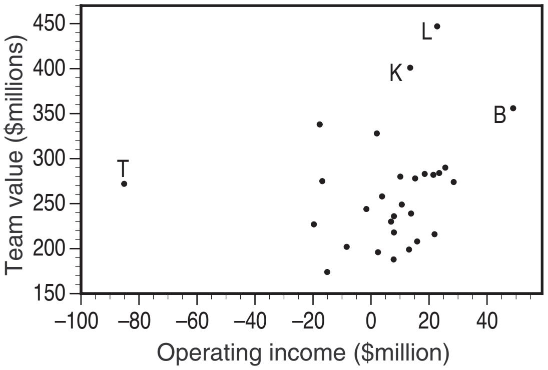


2.18. There appears to be a positive association between cycle length and day length, but it is quite weak: The points of the scatterplot are generally located along a positively-sloped line but with a lot of spread around that line. (Ideally, both axes should have the same scale.)

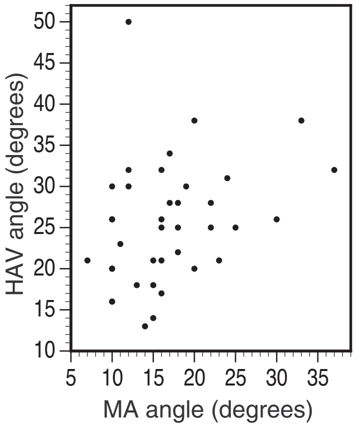


2.19. (Ideally, both graphs should have the same scale on both axes. However, this makes the graph dimensions rather awkward, so the graphs below do not reflect that ideal.) **(a)** The Lakers and the Knicks are high in both variables (but fit the pattern). The Grizzlies, Cavaliers, and Rockets have 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.

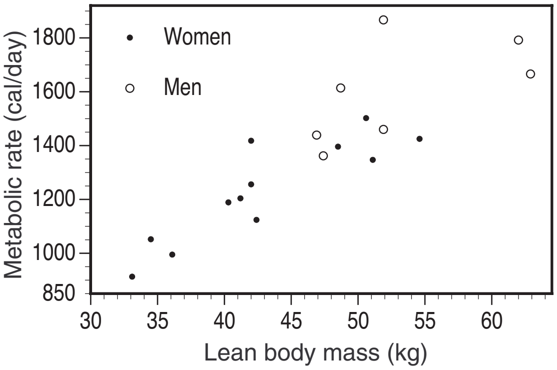




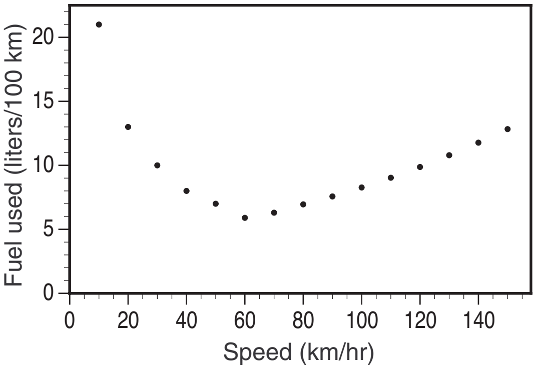
2.20. **(a)** MA angle is the explanatory variable, so it should be on the horizontal axis of the scatterplot. (This scatterplot has the same scale on both axes, because both variables are measured in degrees.) **(b)** The scatterplot shows a moderate-to-weak positive linear association, with one clear 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.



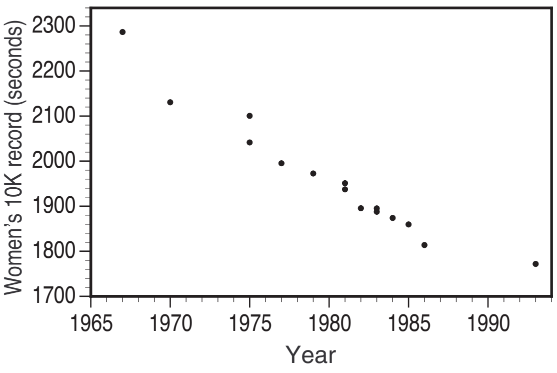
2.21. **(a)** Women are marked with filled circles, men with open circles. **(b)** The association is linear and positive. The women's points show a stronger association. As a group, males typically have larger values for both variables.



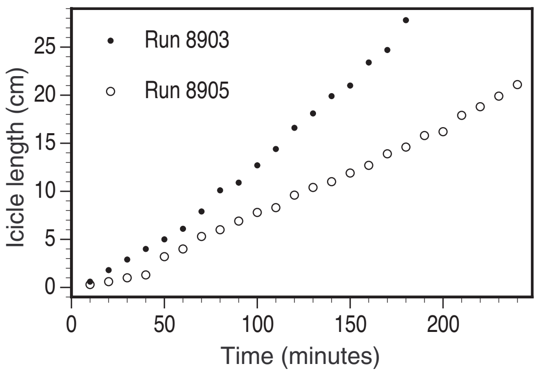
2.22. **(a)** At right; speed is explanatory, so it belongs on the x-axis. **(b)** The relationship is curved—low in the middle, higher at the extremes. Because low "mileage" is actually *good* (it means that we use less fuel to travel 100 km), this makes sense: Moderate speeds yield the best performance. 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.



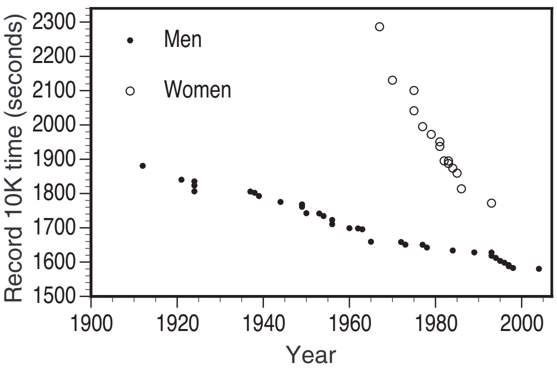
2.23. The plot shows a fairly steady rate of improvement until the mid-1980s, with much slower progress after that (the record has only been broken once since 1986).



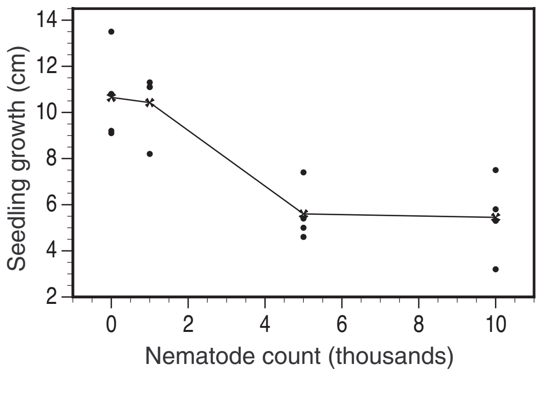
2.24. (a) In the scatterplot on the right, the open circles represent run 8905, the higher flow rate. **(b)** Icicles seem to grow faster when the water runs more slowly. (Note that there is no guarantee that the pattern we observe with these two flow rates applies to rates a lot faster than 29.6 mg/s,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



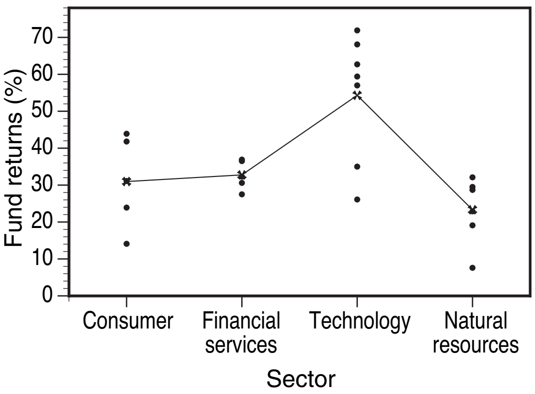
2.25. (a) Both men (filled circles) and women (open circles) show fairly steady improvement. Women have made more rapid progress, but their progress seems to have slowed, while men's records may be dropping more rapidly in recent years. **(b)** The data support the first claim but do not seem to support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



2.26. (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.



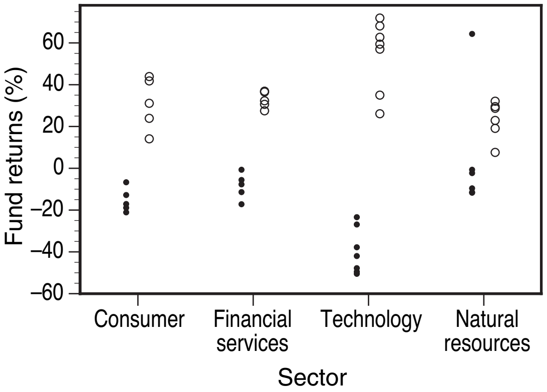
2.27. **(a)** Plot shown on the right. Means (plotted with crosses) are 30.96%, 32.76%, 54.31%, and 23.32%. (Note that the sectors on the horizontal axis are shown there in the order given in the text, but that is completely arbitrary.) **(b)** Technology had the highest average performance. **(c)** 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.)

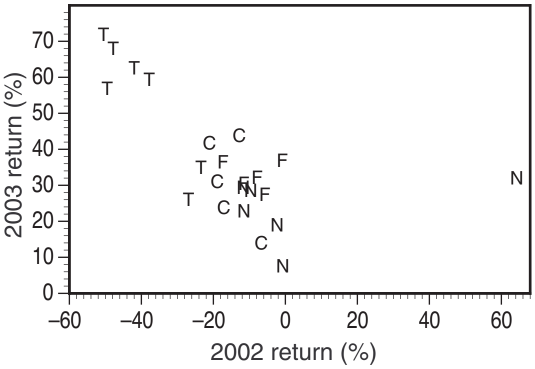


2.28. Methods of graphical analysis will vary; shown below are two possible approaches. On the left, for each sector, 2002 returns are shown as filled circles and 2003 returns are open circles. On the right is a scatterplot with 2002 return as the explanatory variable; the letters C, F, T, and N indicate 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

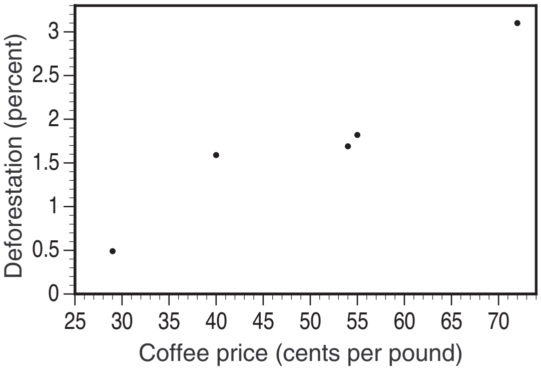
Also shown (right) is a stemplot of the differences for each fund (that is, each fund's 2003 return minus its 2002 return). Only one fund return decreased; every other fund increased its return by between 8.3% and 122.4%.





2.29. **(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



2.30. **(a)** $r = -0.2013$. **(b)** The small correlation is consistent with a weak association.

2.31. **(a)** $r = 0.5194$. **(b)** This correlation is much larger (farther from 0) than the first, which is consistent with the stronger association.

2.32. See also the solution to Exercise 2.8, where the location of this outlier point is discussed. Any outlier should make r closer to 0, because it weakens the relationship.

Note: *In this case, because $r > 0$, this means r gets* smaller. *If r had been negative, getting closer to 0 would mean that r gets* larger *(but gets smaller in absolute value).*

2.33. Such a point should be at the lower left part of the scatterplot. Because it tends to strengthen the relationship, the correlation increases.

Note: *It may be more descriptive to say that r gets further from 0; see the note in the solution to the previous exercise.*

2.34. **(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.

2.35. 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.

2.36. **(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.

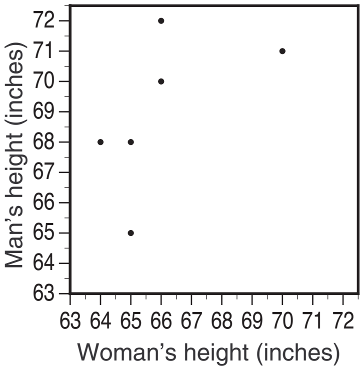
2.37. r would not change; units do not affect correlation.

2.38. **(a)** See the solution to Exercise 2.28 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.

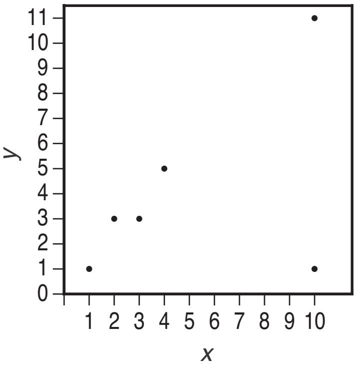
2.39. See also the solution to Exercise 2.19. **(a)** For team value and revenue, $r_1 = 0.9265$; for team value and operating income, $r_2 = 0.2107$. This agrees with conclusions from the scatterplots: Revenue is a much better predictor of team value. **(b)** Without Portland (marked with a "T" in the scatterplot), $r_2 = 0.3469$. The removal of this point makes the scatterplot appear (slightly) more linear, so the association is stronger.

2.40. For Exercise 2.18, $r_1 = 0.2797$; for Exercise 2.24, $r_2 = 0.9958$ (run 8903) and $r_3 = 0.9982$ (run 8905).

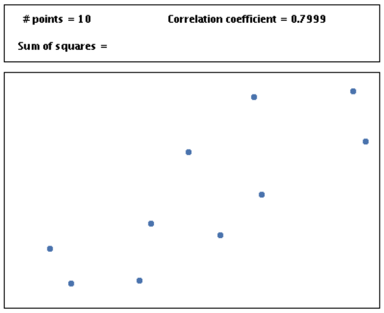
2.41. **(a)** The scatterplot shows a moderate positive association, so r should be positive, but not close to 1. **(b)** The correlation is $r = 0.5653$. **(c)** r would not change if all the men were six inches shorter. A positive correlation does not tell us that the men were generally taller than 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 because the points of the scatterplot would fall exactly on a positively-sloped line (with no scatter).



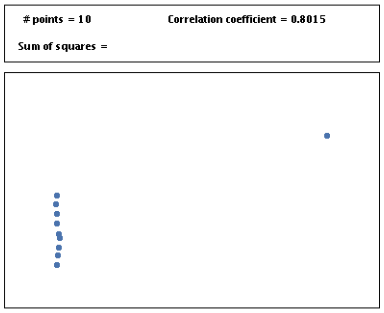
2.42. The correlation is $r \doteq 0.481$. The correlation is greatly lowered by the one outlier. Outliers tend to have fairly strong effects on correlation; it is even stronger here because there are so few observations.



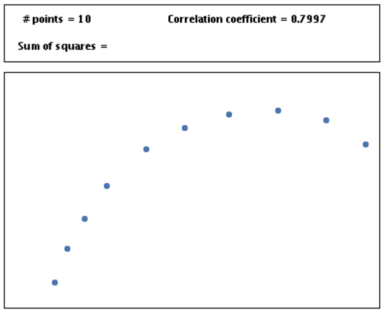
2.43. **(a)** As two points determine a line, the correlation is always either -1 or 1. **(b)** Sketches will vary; an example is shown as the first graph on the next page. Note that the scatterplot must be positively sloped, but r is affected only by the scatter about a line drawn through the data points, 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 as the second graph on the next page. **(d)** To have $r \doteq 0.8$, the curve must be higher at the right than at the left. One possibility is shown as the third graph on the next page.


points = 10 Correlation coefficient = 0.7999

Sum of squares =


points = 10 Correlation coefficient = 0.8015

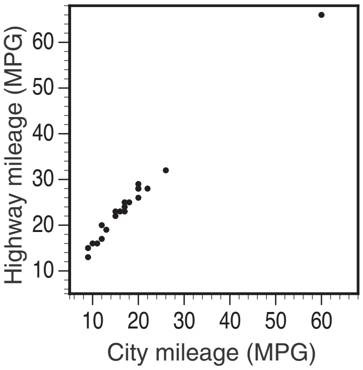
Sum of squares =


points = 10 Correlation coefficient = 0.7997

Sum of squares =

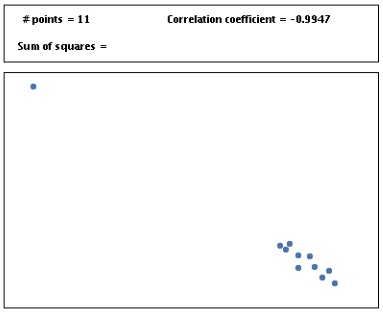
2.44. See the solution to Exercise 2.22 for the scatterplot. $r = -0.172$—it is close to zero, because the relationship is a curve rather than a line; correlation measures *linear* association.

2.45. (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).

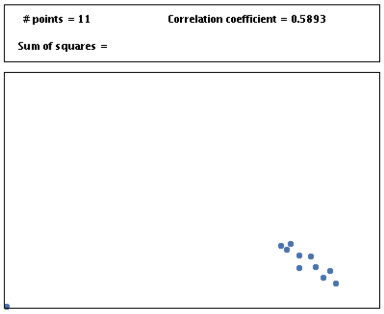


2.46. (a) The correlation will be closer to −1. One possible answer is shown below, left. **(b)** Answers will vary, but the correlation will increase and can be made positive by dragging the point down far enough (see below, right).

Note: *The first printing of the text mistakenly said to place the initial set of 10 points in the lower* left *instead of the lower right.*


points = 11 Correlation coefficient = -0.9947

Sum of squares =


points = 11 Correlation coefficient = 0.5893

Sum of squares =

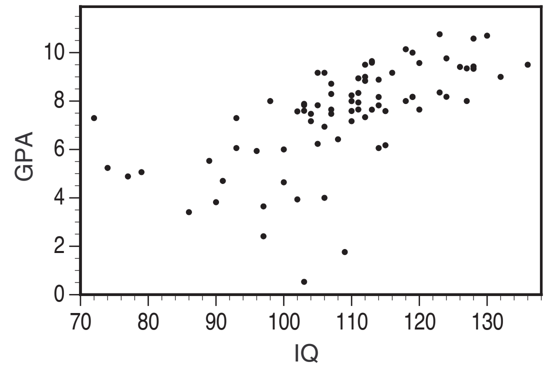
2.47. (Scatterplot not shown.) If the husband's age is y and the wife's x, the linear relationship $y = x + 2$ would hold, and hence $r = 1$ (because the slope is positive).

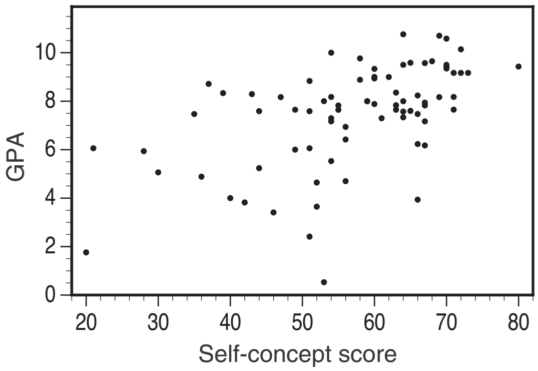
2.48. Explanations and sketches will vary, but should note that correlation measures the strength of the association, not the slope of the line (except for the sign of the slope—positive or negative). The hypothetical Funds A and B mentioned in the report, for example, might be related by a linear formula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

2.49. 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.

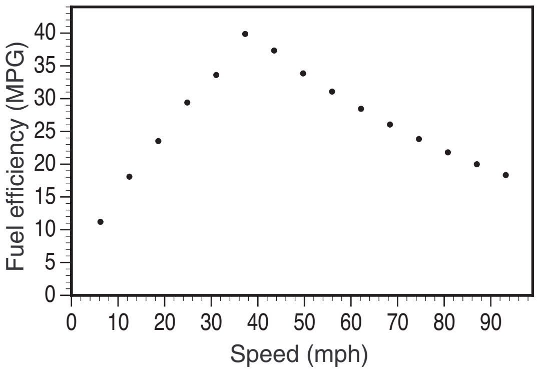
2.50. **(a)** Because gender has a categorical (nominal) scale, we cannot compute the correlation between sex and anything. (There is a strong *association* between gender and income. Some writers and speakers use "correlation" as a synonym for "association." It is much better to retain the more specific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

2.51. 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.

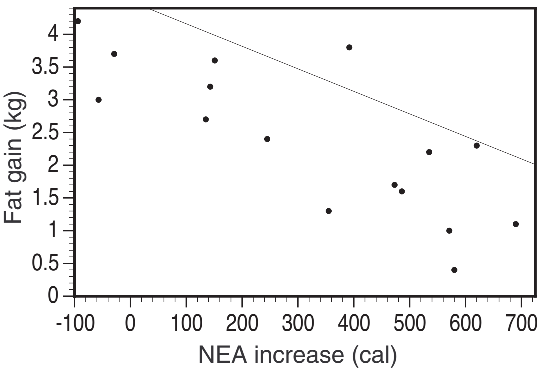




2.52. **(a)** The new speed and fuel consumption (respectively) values are $x^* = x \div 1.609$ and $y^* = y \times 1.609 \div 100 \div 3.785 \doteq 0.004251y$. (The factor of 1/100 is needed since we were measuring fuel consumption in liters/100 km.) The transformed data have the same correlation as 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.



2.53. The line lies almost entirely above the points in the scatterplot. (The slope -0.00344 of this line is the same as the regression line given in Example 2.13, but the intercept 4.505 is one more than the regression intercept.)



2.54. The estimated fat gain is $3.505 - 0.00344 \times 600 \doteq 1.441$ kg.

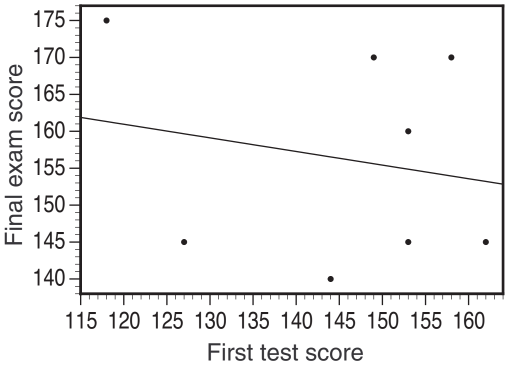
2.55. The data used to determine the regression line had NEA increase values ranging from -94 to 690 calories, so estimates for values inside that range (like 200 and 500) should be relatively safe. For values far outside this range (like -400 and 1000), the predictions would not be trustworthy.

2.56. The table on the right shows the values of r^2 (expressed as a percentage). From this we can observe that (i) the fraction of variation explained depends only on the magnitude (absolute value) of r, not its sign, and (ii) the fraction of explained variation drops off drastically as $|r|$ moves away from 1.

r	-0.9	-0.5	-0.3	0	0.3	0.5	0.9
r^2	81%	25%	9%	0%	9%	25%	81%

2.57. **(a)** When $x = 5$, $y = 10 + 5 \times 5 = 35$. **(b)** y increases by 5. (The change in y corresponding to a unit increase in x is the slope of this line.) **(c)** The intercept of this equation is 10.

2.58. **(a)** The plot (right) is the same as in Exercise 2.6, but with the regression line added. **(b)** The regression line is Final $=$ $183 - 0.184 \times$ First. Minitab output below.

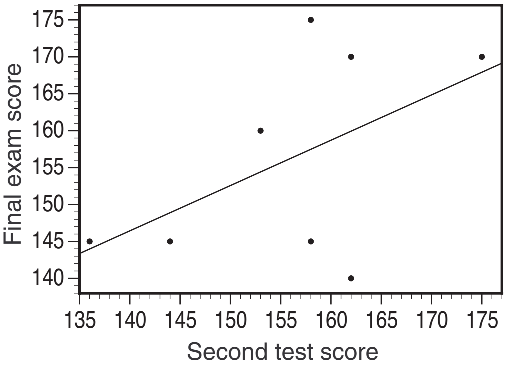


Minitab output

```
Predictor        Coef       Stdev    t-ratio        p
Constant       183.08       53.55       3.42    0.014
first         -0.1844      0.3663      -0.50    0.633

s = 14.90       R-sq = 4.1%      R-sq(adj) = 0.0%
```

2.59. **(a)** The plot (right) is the same as in Exercise 2.7, but with the regression line added. **(b)** The regression line is Final $=$ $60.5 + 0.614 \times$ Second. Minitab output below.



Minitab output

```
Predictor        Coef       Stdev    t-ratio        p
Constant        60.52       64.46       0.94    0.384
second         0.6137      0.4122       1.49    0.187

s = 12.99        R-sq = 27.0%     R-sq(adj) = 14.8%
```

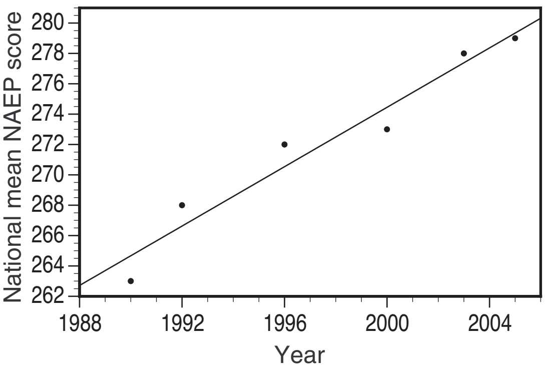
2.60. See also the solutions to Exercises 2.8 and 2.32. To be considered an outlier, the point for the ninth student should be in either the upper left or lower right portion of the scatterplot—either a student who had a below-average second-test score but an above-average final-exam score, or a student who did well on the second test but poorly on the final. In either case, the outlier should alter the equation slightly (the slope will decrease because the line is pulled toward the outlier), and the value of r^2 will decrease (because the relationship has been weakened).

2.61. See also the solution to Exercise 2.33. Because this new point fits the pattern of the other points, the regression equation should change very little. r^2 will increase because the relationship is stronger (i.e., the relative scatter about the regression line is less).

2.62. **(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.

2.63. **(a)** Based on the slope, 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.

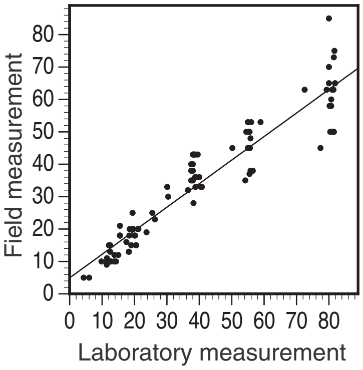
2.64. **(a)** Because the slope is 0.0086 (in units of "proportion of perch eaten per perch count"), an increase of 10 in the perch count increases the proportion eaten by 0.086 (on the average). **(b)** When the perch count is 0, the equation tells us that 12% (0.12) of those perch will be eaten. Of 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.



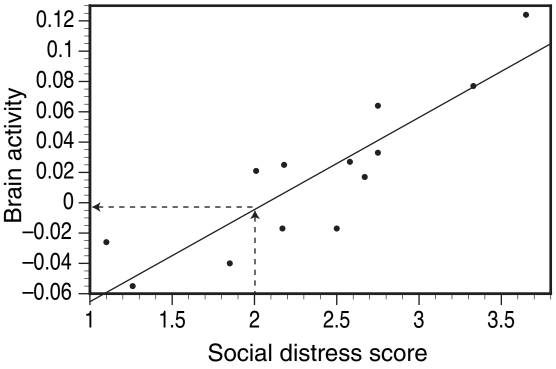
2.65. **(a)** Time plot shown on the right, along with the regression line. **(b)** The means and standard deviations are $\bar{x} = 1997.\overline{6}$, $\bar{y} = 272.1\overline{6}$, $s_x \doteq 6.0222$, and $s_y \doteq 6.0470$. With the correlation $r \doteq 0.9739$, the slope and intercept are

$$b = r\, s_y/s_x \doteq 0.9779 \text{ and}$$
$$a = \bar{y} - b\bar{x} \doteq -1681$$

The equation is therefore $\hat{y} = -1681 + 0.9779x$; this line explains about $r^2 \doteq 95\%$ of the variation in score.

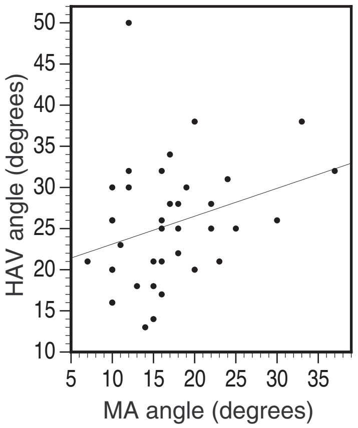


2.66. **(a)** The least-squares line is $\hat{y} = 0.7267x + 4.9433$. This is less steep than the line $y = x$, reflecting the observation that field measurements tend to be lower for greater depths. **(b)** The line $y = x$ has slope 1; the regression line has slope 0.7267. A slope of 1 would mean that, 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.



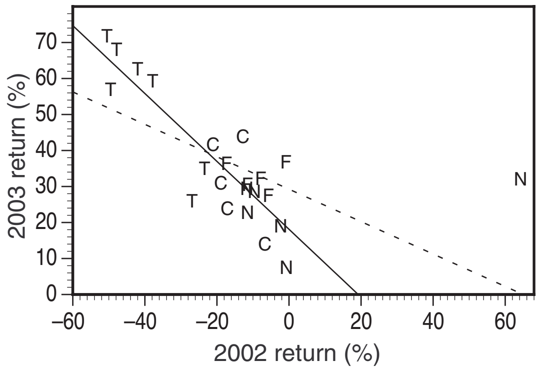
2.67. See also the solution to Exercise 2.17. **(a)** The regression equation is $\hat{y} = 0.06078x - 0.1261$. **(b)** Based on the "up-and-over" method, most students will probably estimate that $\hat{y} \doteq 0$; the regression formula gives $\hat{y} = -0.0045$. **(c)** The correlation is $r \doteq 0.8782$, so the line explains $r^2 = 77\%$ of the variation in brain activity.

2.68. See also the solution to Exercise 2.20. **(a)** The regression line is $\hat{y} = 19.7 + 0.339x$. **(b)** For $x = 25°$, we predict $\hat{y} = 28.2°$. **(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.



2.69. 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.

2.70. **(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.



2.71. 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.

2.72. 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 2.69).

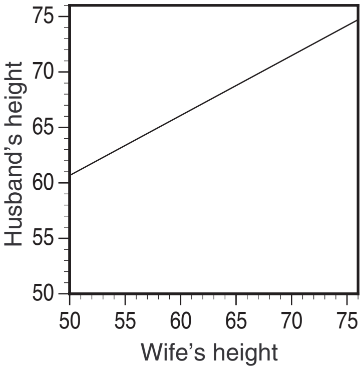
For predicting time from length, the slope and intercept are $b_2 = r\, s_x/s_y \doteq 6.26$ min/cm and $a_2 = \bar{x} - b_2\bar{y} \doteq 15.79$ min, giving the equation $\hat{x} = 15.79 + 6.26y$.

2.73. 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.

2.74. 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.

2.75. 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.



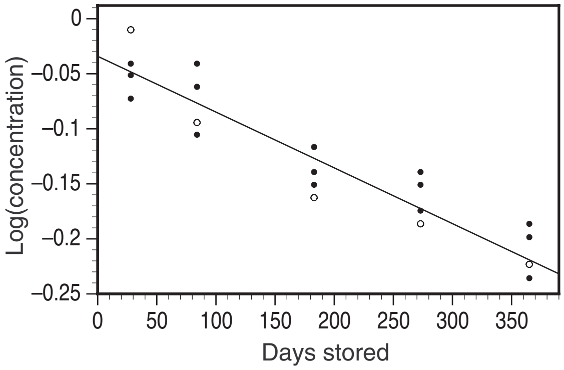
2.76. We have slope $b = r\, s_y/s_x$ and intercept $a = \bar{y} - b\bar{x}$, and $\hat{y} = a + bx$, so when $x = \bar{x}$, $\hat{y} = a + b\bar{x} = (\bar{y} - b\bar{x}) + b\bar{x} = \bar{y}$. (Note that the value of the slope does not actually matter.)

2.77. (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 2.69.)

2.78. (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.

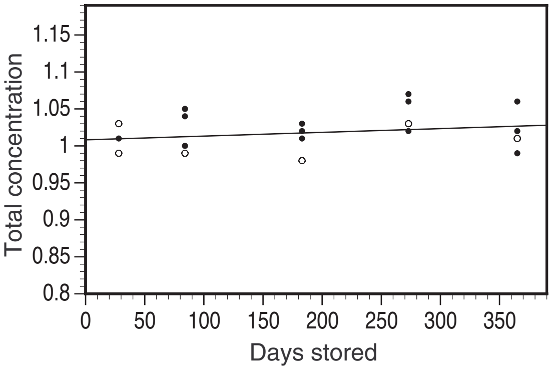
2.79. $r = \sqrt{0.16} = 0.40$ (high attendance goes with high grades, so r must be positive).

2.80. (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."*

2.81. **(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.

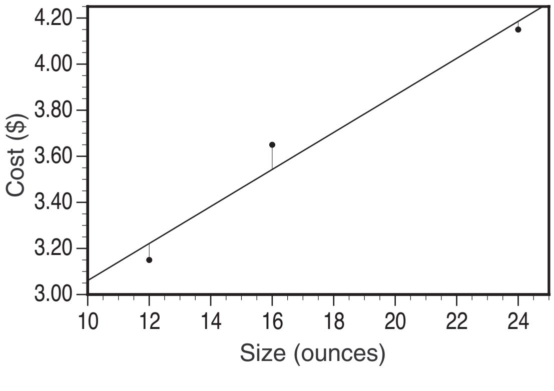


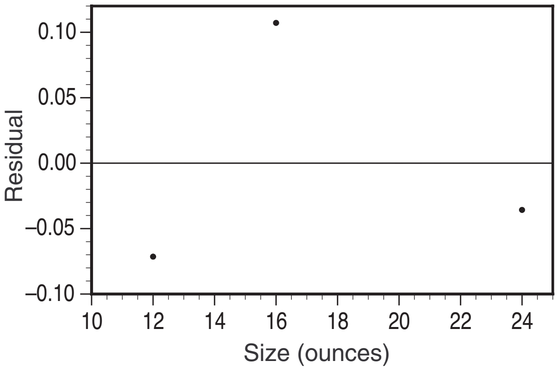
2.82. For an NEA increase of 143 calories, the predicted fat gain is $\hat{y} = 3.505 - 0.00344 \times 143 \doteq 3.013$, so the residual is $y - \hat{y} \doteq 0.187$. This residual is positive because the actual fat gain was greater than the prediction.

Note: *The first printing of the text mistakenly asked why this residual is* negative *instead of positive*.

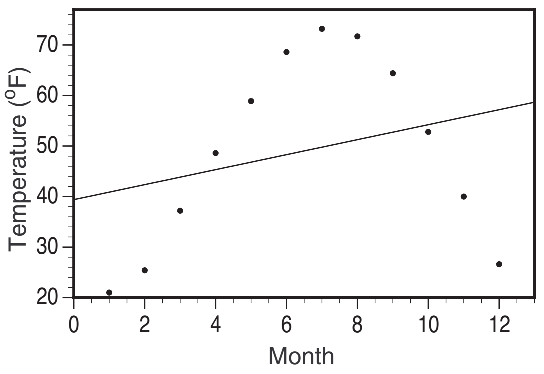
2.83. The sum of the residuals is 0.01.

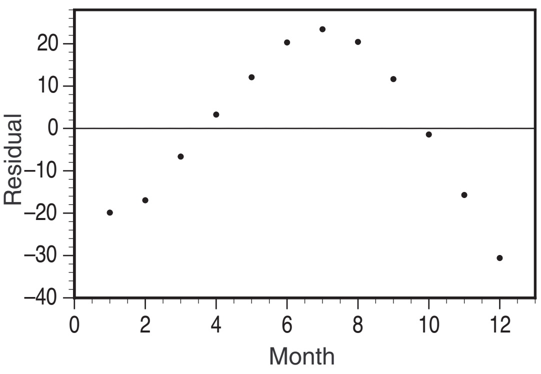
2.84. See also the solution to Exercise 2.4. **(a)** Size is the explanatory variable. **(b)** The regression line is $\hat{y} = 2.2571 + 0.0804x$. **(c)** See the plot (below, left). **(d)** Rounded to four decimal places, the residuals (as computed by software) are -0.0714, 0.1071, and -0.0357. It turns out that these three residuals add up to 0, no matter how much they are rounded. However, if they are computed by hand using the regression equation given in part (b)—which has rounded values for the slope and intercept—there is some roundoff error; in that case, the residuals are -0.0719, 0.1065, and -0.0367, which add up to -0.0021. **(e)** The middle residual is positive and the other two are negative, meaning that the 16-ounce drink costs more than the predicted value and the other two sizes cost less than predicted. Note that the residuals show the same pattern (relative to a horizontal line at 0) as the original points around the regression line.



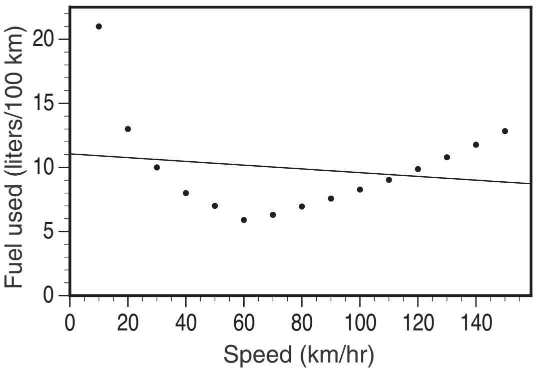


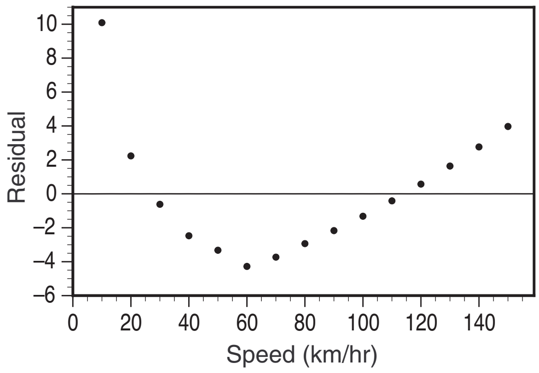
2.85. **(a)** The plot (below, left) is curved (low at the beginning and end of the year, high in the middle). **(b)** The regression line is $\hat{y} = 39.392 + 1.4832x$. It does not fit well because a line is poor summary of this relationship. **(c)** Residuals are negative for January through March and October through December (when actual temperature is less than predicted temperature), and positive from April to September (when it is warmer than predicted). **(d)** A similar pattern would be expected in any city that is subject to seasonal temperature variation. **(e)** Seasons in the Southern Hemisphere are reversed, so temperature would be cooler in the middle of the year.



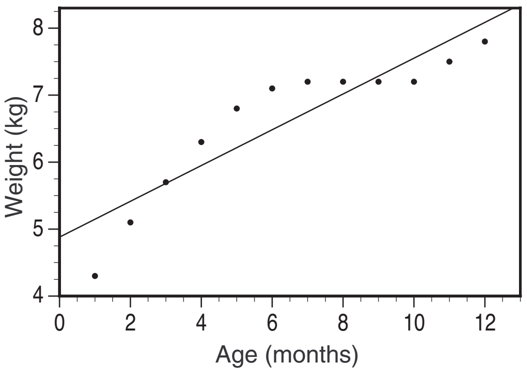


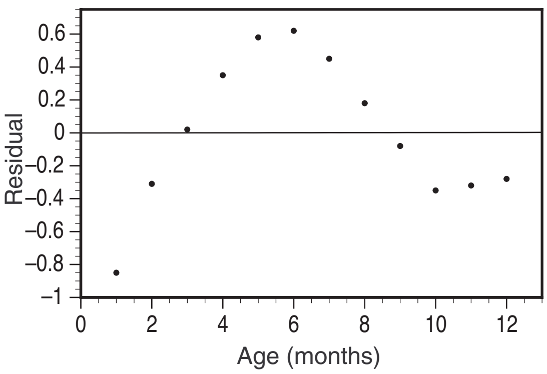
2.86. See also the solutions to Exercises 2.22 and 2.52. **(a)** Below, left. **(b)** The sum is −0.01. **(c)** The first two and last four residuals are positive, and those in the middle are negative. Plot below, right.



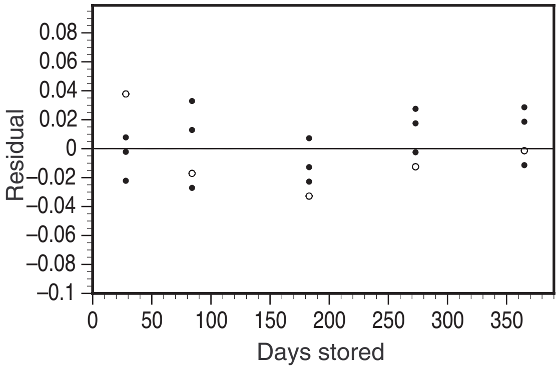


2.87. **(a)** Below, left. **(b)** This line is not a good summary of the pattern because the pattern is curved rather than linear. **(c)** The sum is 0.01. The first two and last four residuals are negative, and those in the middle are positive. Plot below, right.





2.88. **(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.



2.89. 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.

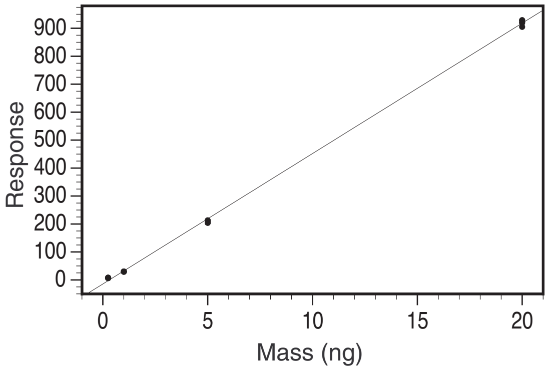
2.90. Presumably, those applicants who were hired would generally have been those who scored well on the test. As a result, we have little or no information on the job performance of those who scored poorly (and were therefore not hired). Those with higher test scores (who were hired) will likely have 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.

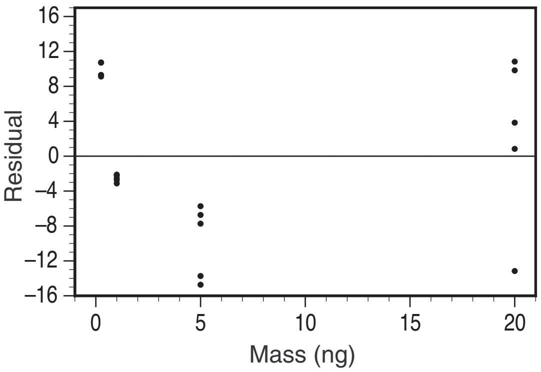
2.91. For example, a student who in the past might have received a grade of B (and a lower SAT score) now receives an A (but has a lower SAT score than an A student in the past). While this is a bit of an oversimplification, this means that today's A students are yesterday's A and B students, today's 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.

2.92. 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).

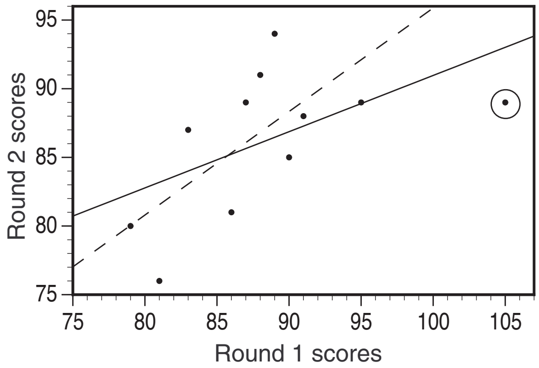
2.93. The correlation between BMR and fat gain is $r = 0.08795$; the slope of the regression line is $b = 0.000811$ kg/cal. These both show that BMR is less useful for predicting fat gain. The small correlation suggests a very weak linear relationship (explaining less than 1% of the variation in fat 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.

2.94. **(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.

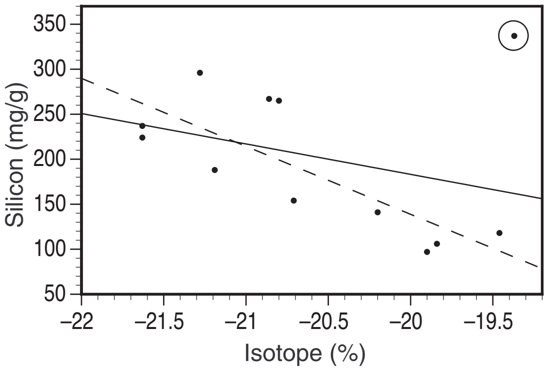




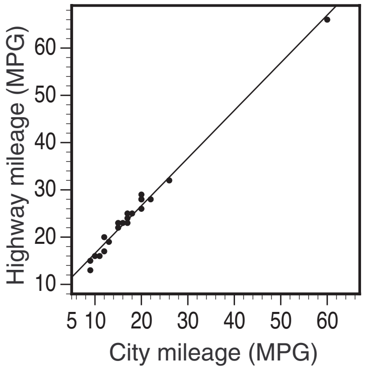
2.95. **(a)** There is a moderate positive relationship; player 7's point is an outlier. Ideally, both scales should be equal. **(b)** The first equation is the dashed line in the plot. It omits the influential observation; the other (solid) line is pulled toward the outlier.

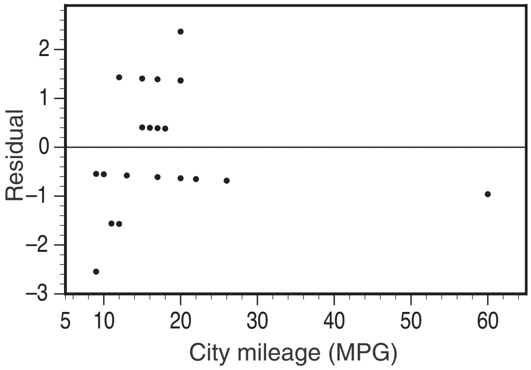


2.96. **(a)** Apart from the outlier—circled for part (b)—the scatterplot shows a moderate linear negative association. **(b)** With the outlier, $r = -0.3387$; without it, $r^* = -0.7866$. **(c)** The two regression formulas are $\hat{y} = -492.6 - 33.79x$ (the solid line, with all points) and $\hat{y}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

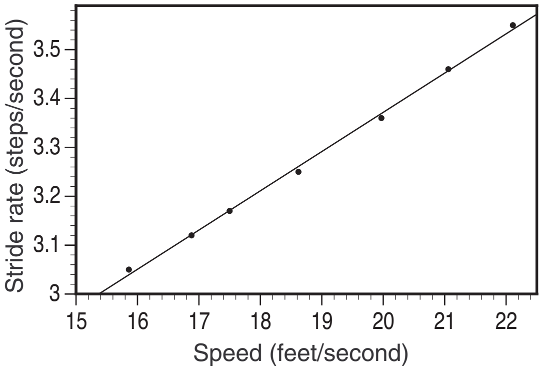


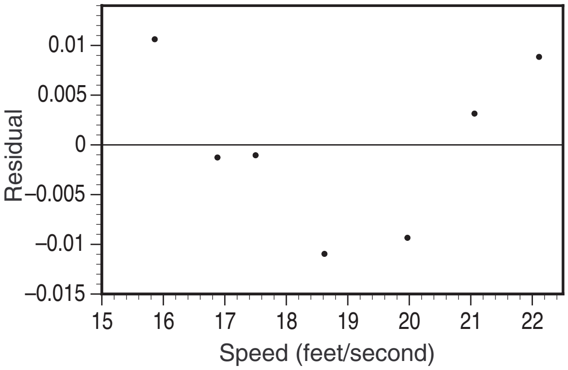
2.97. **(a)** Scatterplot below on the left. **(b)** The regression line is $\hat{y} = 6.47 + 1.01x$. The residual plot is below on the right. **(c)** The largest residuals are the Porsche Boxster (2.365) and Lamborghini Murcielago (−2.545). **(d)** The Insight is influential; it pulls the line toward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



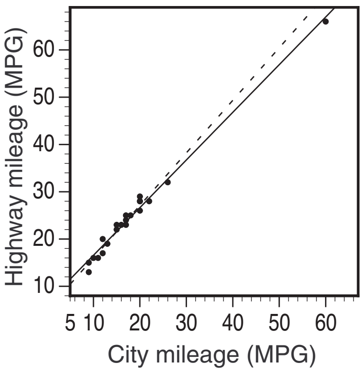


2.98. **(a)** Scatterplot on top of the next page, left. The relationship seems linear. **(b)** The regression line is $\hat{y} = 1.77 + 0.0803x$ (y is stride rate, x is speed). **(c)** The residuals (reported by Minitab, then rounded to 3 decimal places) are 0.011, −0.001, −0.001, −0.011, −0.009, 0.003, and 0.009. These add up 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.



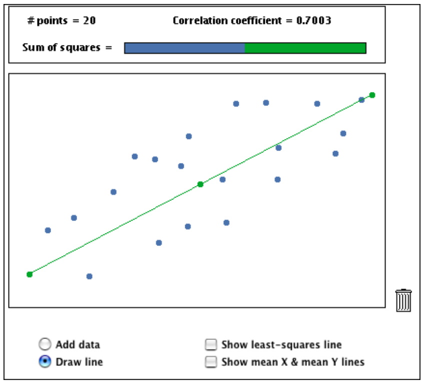


2.99. Without the Insight, $\hat{y} = 4.87 + 1.11x$ (the dashed line in the plot). For city mileages between 10 and 30 MPG, the difference in predicted highway mileage (with or without the Insight) is no more than 1.4 MPG, so the Insight is not very influential; it falls near the line suggested by the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



2.100. 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.

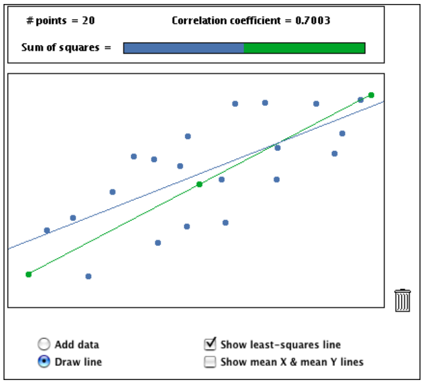
2.101. (a) Drawing the "best line" by eye is a very inaccurate process; few people choose the best line (although you can get better at it with practice). **(b)** Most people tend to overestimate the slope for a scatterplot with $r \doteq 0.7$; that is, most students will find that the least-squares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



points = 20 Correlation coefficient = 0.7003

Sum of squares =

○ Add data
● Draw line
□ Show least-squares line
□ Show mean X & mean Y lines



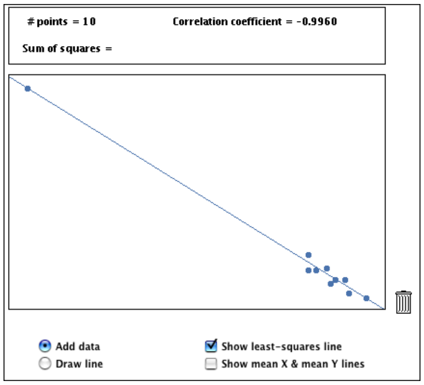
points = 20 Correlation coefficient = 0.7003

Sum of squares =

○ Add data
● Draw line
☑ Show least-squares line
□ Show mean X & mean Y lines

2.102. (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.

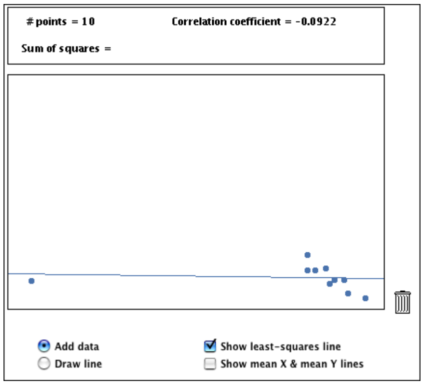
Note: *The first printing of the text mistakenly said to place the initial set of 10 points in the lower* left *instead of the lower right.*



points = 10 Correlation coefficient = -0.9960

Sum of squares =

● Add data
○ Draw line
☑ Show least-squares line
□ Show mean X & mean Y lines

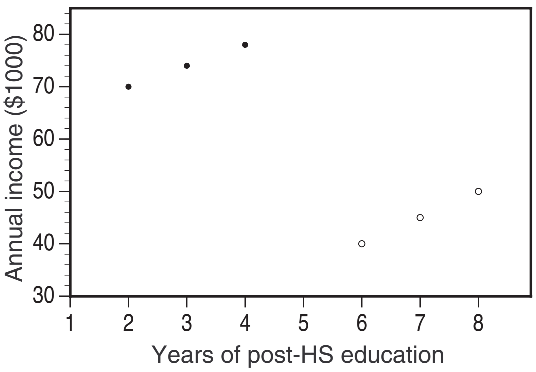


points = 10 Correlation coefficient = -0.0922

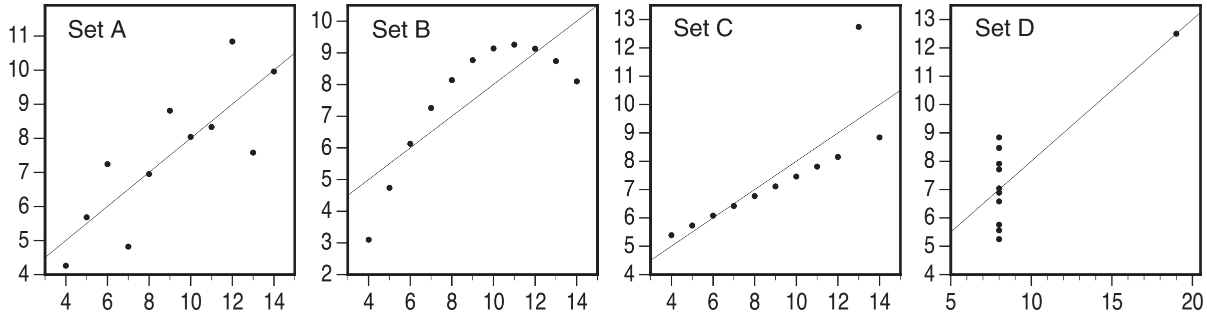
Sum of squares =

● Add data
○ Draw line
☑ Show least-squares line
□ Show mean X & mean Y lines

2.103. 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.



2.104. **(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)** Scatterplots 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, 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.)



2.105. There are 1684 female binge drinkers in the table; 8232 female students are not binge drinkers.

2.106. There are $1684 + 8232 = 9916$ women in the study. The number of students who are not binge drinkers is $5550 + 8232 = 13{,}782$.

2.107. Divide the number of non-bingeing females by the total number of students:

$$\frac{8232}{17{,}096} \doteq 0.482$$

2.108. Use the numbers in the right-hand column of the table in Example 2.28. Divide the counts of bingeing and non-bingeing students by the total number of students:

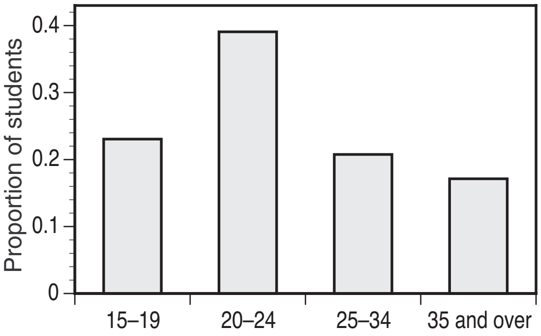
$$\frac{3314}{17{,}096} \doteq 0.194 \text{ and } \frac{13{,}782}{17{,}096} \doteq 0.806$$

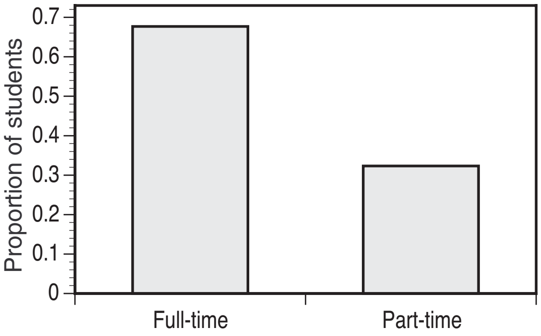
2.109. This is a conditional distribution; take the number of bingeing males divided by the total number of males: $\frac{1630}{7180} \doteq 0.227$.

2.110. The first computation was performed in the previous solution; for the second, take the number of non-bingeing males divided by the total number of males: $\frac{5550}{7180} \doteq 0.773$.

	FT	PT	
15–19	3388 *0.2067*	389 *0.0237*	3777 0.2305
20–24	5238 *0.3196*	1164 *0.0710*	6402 0.3907
25–34	1703 *0.1039*	1699 *0.1037*	3402 0.2076
35+	762 *0.0465*	2045 *0.1248*	2807 0.1713
	11091 0.6768	5297 0.3232	16388

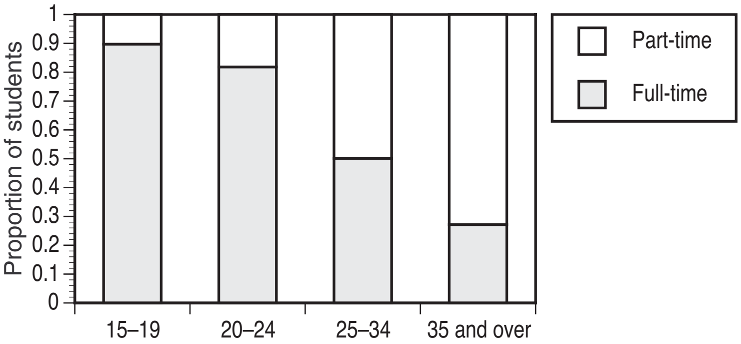
2.111. (a) There are about 3,388,000 full-time college students aged 15 to 19. (Note that numbers are in thousands.) **(b)** The joint distribution is found by dividing each number in the table by 16,388 (the total of all the numbers). These proportions are given in italics on the right. For example, $\frac{3388}{16388} \doteq 0.2067$, meaning that about 20.7% of all college students are full-time and aged 15 to 19. **(c)** The marginal distribution of age is found by dividing the *row* totals by 16,388; they are in the right margin of the table and the graph on the left below. For example, $\frac{3777}{16388} \doteq 0.2305$, meaning that about 23% of all college students are aged 15 to 19. **(d)** The marginal distribution of status is found by dividing the *column* totals by 16,388; they are in the bottom margin of the table and the graph on the right below. For example, $\frac{11091}{16388} \doteq 0.6768$, meaning that about 67.7% of all college students are full-time.





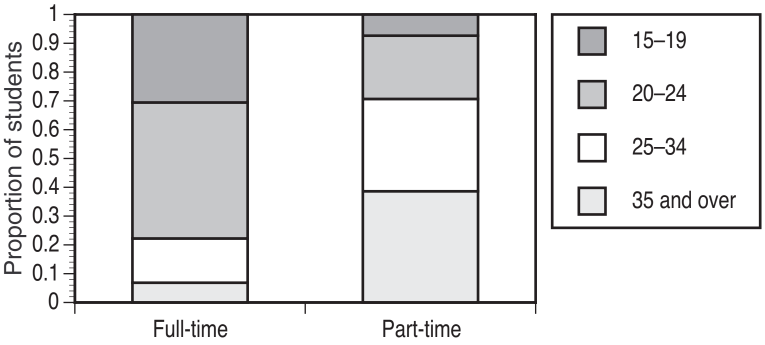
2.112. Refer to the counts in the solution to Exercise 2.111. For each age category, the conditional distribution of status is found by dividing the counts in that row by that row total. For example, $\frac{3388}{3777} \doteq 0.8970$ and $\frac{389}{3777} \doteq 0.1030$, meaning that of all college students in the 15–19 age range, about 89.7% are full-time, and the rest (10.3%) are part-time. Note that each pair of numbers should add up to 1 (except for rounding error, but with only two numbers, that rarely happens). The complete table is shown on the next page, along with one possible graphical presentation. We see that the older the students are, the more likely they are to be part-time.

	FT	PT
15–19	0.8970	0.1030
20–24	0.8182	0.1818
25–34	0.5006	0.4994
35+	0.2715	0.7285



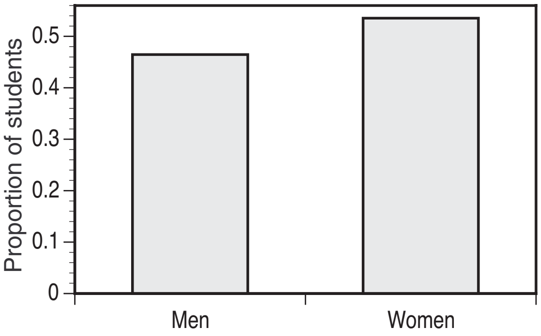
2.113. Refer to the counts in the solution to Exercise 2.111. For each status category, the conditional distribution of age is found by dividing the counts in that column by that column total. For example, $\frac{3388}{11091} \doteq 0.3055$, $\frac{5238}{11091} \doteq 0.4723$, etc., meaning that of all full-time college students, about 30.55% are aged 15 to 19, 47.23% are 20 to 24, and so on. Note that each set of four numbers should add up to 1 (except for rounding error). Graphical presentations may vary; one possibility is shown below. We see that full-time students are dominated by younger ages, while part-time students are more likely to be older. (This is essentially the same observation made in the previous exercise, seen from a different viewpoint.)

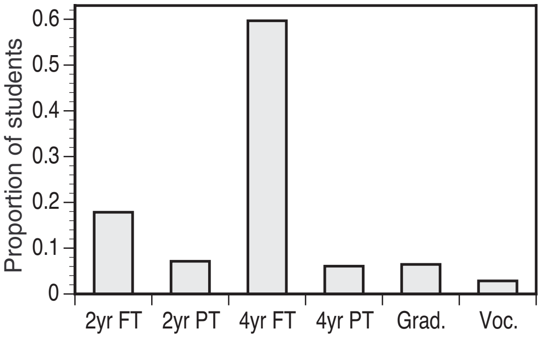
	FT	PT
15–19	0.3055	0.0734
20–24	0.4723	0.2197
25–34	0.1535	0.3207
35+	0.0687	0.3861



2.114. (a) There are about 890,000 male recent high school graduates aged 16 to 24 years enrolled full-time in two-year colleges. **(b)** The marginal distribution of gender is found by dividing the *column* totals by 10,421 (the grand total for the table); they are in the bottom margin of the table and the graph on the left on the next page. For example, $\frac{4842}{10421} \doteq 0.4646$, meaning that about 46.5% of all these students are men. **(c)** The marginal distribution of status is found by dividing the *row* totals by 10,421; they are in the right margin of the table and the graph on the right on the next page. For example, $\frac{1859}{10421} \doteq 0.1784$, meaning that about 17.8% of these students are enrolled full-time in two-year colleges.

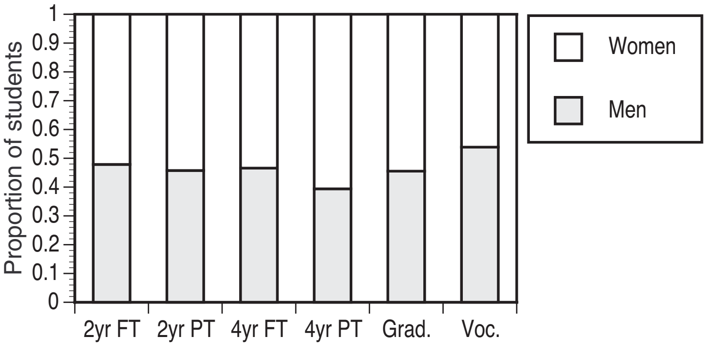
	Men	Women	
2yr FT	890	969	1859 0.1784
2yr PT	340	403	743 0.0713
4yr FT	2897	3321	6218 0.5967
4yr PT	249	383	632 0.0606
Grad	306	366	672 0.0645
Voc	160	137	297 0.0285
	4842 0.4646	5579 0.5354	10421





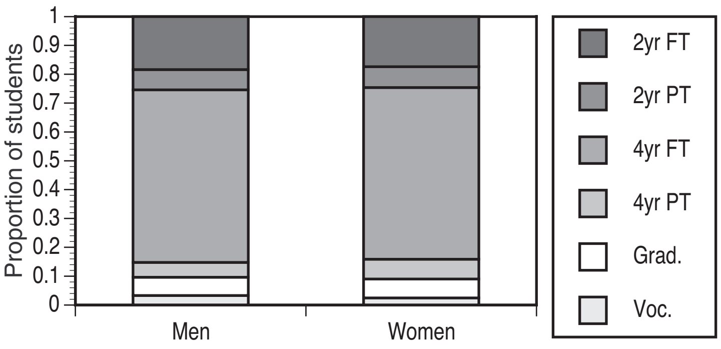
2.115. Refer to the counts in the solution to Exercise 2.114. For each status, the conditional distribution of gender is found by dividing the counts in that row by that row total. For example, $\frac{890}{1859} \doteq 0.4788$ and $\frac{969}{1859} \doteq 0.5212$, meaning that of all full-time students at two-year colleges, about 47.9% are men and the rest (52.1%) are women. Note that each pair of numbers should add to 1 (except for rounding error, but with only two numbers, that rarely happens). Graphical presentations may vary; one possibility is shown below. We see that women make up the majority of all groups except for vocational school students.

	Men	Women
2-yr FT	0.4788	0.5212
2-yr PT	0.4576	0.5424
4-yr FT	0.4659	0.5341
4-yr PT	0.3940	0.6060
Grad.	0.4554	0.5446
Voc.	0.5387	0.4613



2.116. Refer to the counts in the solution to Exercise 2.114. For each gender, the conditional distribution of status is found by dividing the counts in that column by that column total. For example, $\frac{890}{4842} \doteq 0.1838$, $\frac{340}{4842} \doteq 0.0702$, etc., meaning that of all male college students, about 18.38% are enrolled full-time in two-year colleges, 7.02% are attending a two-year college part-time, and so on. Note that each set of six numbers should add up to 1 (except for rounding error). Graphical presentations may vary; one possibility is shown below. We see that there is little difference between genders in the distribution of status: The percentages of men and women in each status category are quite similar.

	Men	Women
2-yr FT	0.1838	0.1737
2-yr PT	0.0702	0.0722
4-yr FT	0.5983	0.5953
4-yr PT	0.0514	0.0687
Grad.	0.0632	0.0656
Voc.	0.0330	0.0246



2.117. Two examples are shown on the right. In general, choose a to be any number from 0 to 200, and then all the other entries can be determined.

50	150
150	50

175	25
25	175

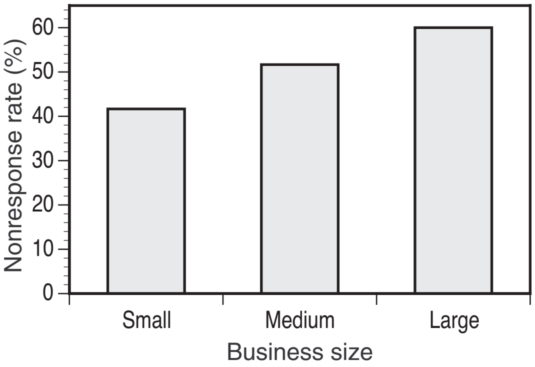
Note: *This is why we say that such a table has "one degree of freedom": We can make one (nearly) arbitrary choice for the first number, and then have no more decisions to make.*

2.118. To construct such a table, we can start by choosing values for the row and column sums $r1, r2, r3, c1, c2, c3$, as well as the grand total N. Note that the $N = r1 + r2 + r3 = c1 + c2 + c3$, so we only have five choices to make. Then, find each count $a, b, c, d, e, f, g, h, i$ by taking the corresponding *row* total, times the corresponding *column* total, divided by the *grand* total. For example, $a = r1 \times c1/N$ and $f = r2 \times c3/N$. Of course, these counts should be whole numbers, so it may be necessary to make adjustments in the row and column totals to meet this requirement.

a	b	c	$r1$
d	e	f	$r2$
g	h	i	$r3$
$c1$	$c2$	$c3$	N

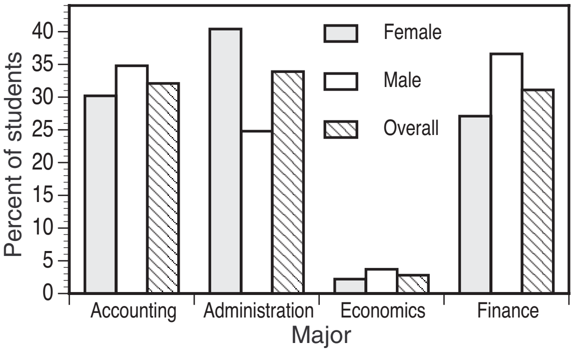
The simplest such table would have all nine counts $a, b, c, d, e, f, g, h, i$ equal to one another.

2.119. **(a)** Overall, $\frac{125+155+180}{900} \doteq 51.1\%$ did not respond. **(b)** Generally, the larger the business, the less likely it was to respond: $\frac{125}{300} \doteq 41.7\%$ of small businesses, $\frac{155}{300} \doteq 51.7\%$ of medium-sized businesses, and $\frac{180}{300} = 60.0\%$ of large businesses did not respond. **(c)** At right. **(d)** Of the 440 total responses, $\frac{175}{440} \doteq 39.8\%$ came from small businesses, $\frac{145}{440} \doteq 33.0\%$ from medium-sized businesses, and $\frac{120}{440} \doteq 27.3\%$ from large businesses. **(e)** No: Almost 40% of respondents were small businesses, while just over a quarter of all responses come from large businesses.

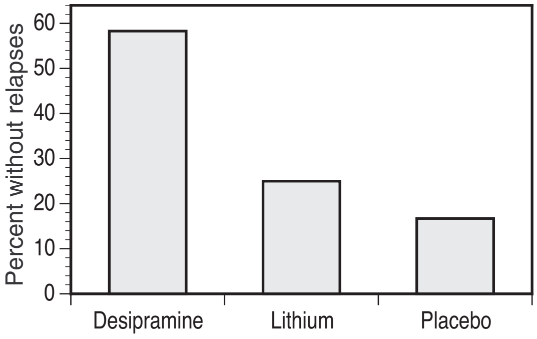


2.120. **(a)** Use column percents, e.g., $\frac{68}{225} \doteq 30.22\%$ of females are in administration. See table and graph below. The biggest difference between women and men is in Administration: A higher percentage of women chose this major. Meanwhile, a greater proportion of men chose other fields, especially Finance. **(b)** There were 386 responses; $\frac{336}{722} \doteq 46.5\%$ did not respond.

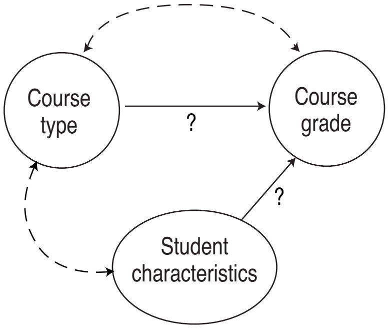
	Female	Male	Overall
Accting.	30.22%	34.78%	32.12%
Admin.	40.44%	24.843%	33.94%
Econ.	2.22%	3.7%	2.85%
Fin.	27.11%	36.65%	31.09%



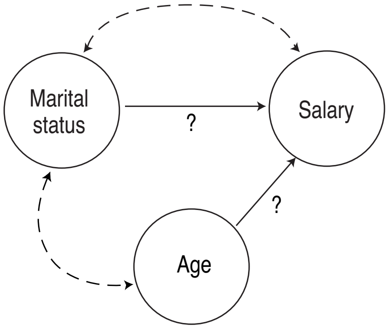
2.121. $\frac{14}{24} \doteq 58.33\%$ of desipramine users did not have a relapse, while $\frac{6}{24} = 25\%$ of lithium users and $\frac{4}{24} \doteq 16.67\%$ of those who received placebos succeeded in breaking their addictions. Desipramine seems to be effective. Note that use of percentages is not as crucial here as in other cases because each drug was given to 24 addicts.



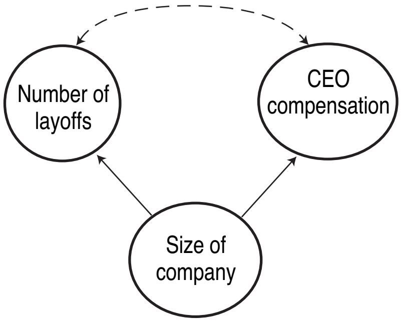
2.122. Responses will vary. For example, students who choose the online course might have more self-motivation or better computer skills. A diagram is shown on the right; the generic "Student characteristics" might be replaced with something more specific.



2.123. 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."



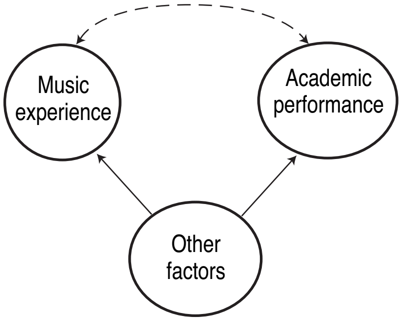
2.124. A large company has more workers who might be laid off and often pays its CEO a higher salary (because, presumably, there is more work involved in running a large company than a small one). Smaller companies typically pay less and have fewer workers to lay off.



2.125. 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.

2.126. 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.

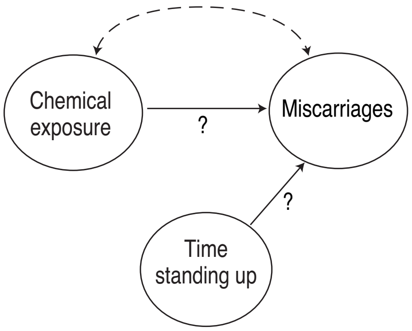
2.127. Students with music experience may have other advantages (wealthier parents, better school systems, and so forth.). That is, experience with music may have been a "symptom" (common response) of some other factor that also tends to cause high grades.



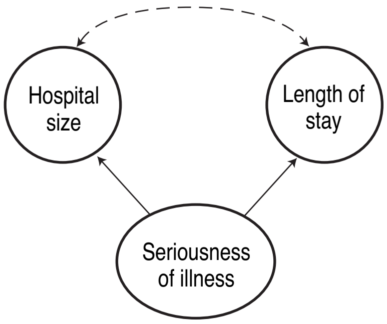
2.128. Two possibilities are that they might perform better simply because this is their second attempt or because they feel better prepared as a result of taking the course (whether or not they really *are* better prepared).

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

For 2.129



For 2.130



2.130. 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.

2.131. 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.

2.132. 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.)

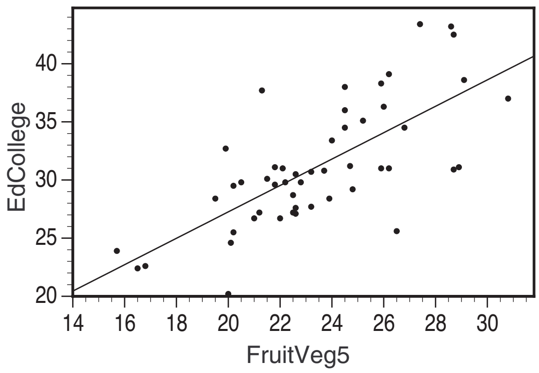
2.133. (a) Statements such as this typically mean that the risk of dying at a given age is half as great; that is, given two groups of the same age, where one group walks and the other does not, the walkers are half as likely to die in (say) the next year. **(b)** Men who choose to walk might also choose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

2.134. 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.

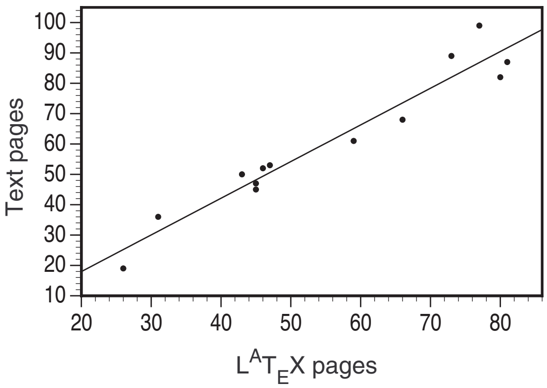
2.135. A school that accepts weaker students but graduates a higher-than-expected number of them would have a positive residual, while a school with a stronger incoming class but a lower-than-expected graduation rate would have a negative residual. It seems reasonable to measure school quality by how much benefit students receive from attending the school.

2.136. (a) The association is negative and roughly linear. This seems reasonable because a low number of smokers suggests that the state's population is health-conscious, so we might expect more people in that state to have healthy eating habits. **(b)** The correlation is $r \doteq -0.4798$. **(c)** Utah is the farthest point to the left (that is, it has the lowest smoking rate) and lies well below the line (i.e., the proportion of adults who eat fruits and vegetables is lower than we would expect). **(d)** California has the second-lowest smoking rate and one of the highest fruit/vegetable rates. This point lies above the line, meaning that the proportion of adults to eat fruits and vegetables is higher than we would expect.

2.137. **(a)** The scatterplot shows a moderate positive association. **(b)** The regression line ($y = 1.1353x + 4.5503$) fits the overall trend. **(c)** For example, a state whose point falls above the line has a higher percent of college graduates than we would expect based on the percent who eat 5 servings of fruits and vegetables. **(d)** No; association is not evidence of causation.

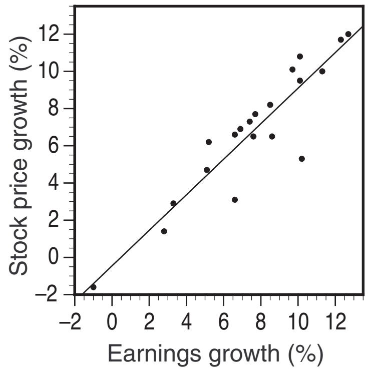


2.138. **(a)** The plot shows a fairly strong positive linear association. **(b)** The regression equation is $\hat{y} = -6.202 + 1.2081x$. **(c)** If $x = 62$ pages, we predict $\hat{y} \doteq 68.7$ pages.



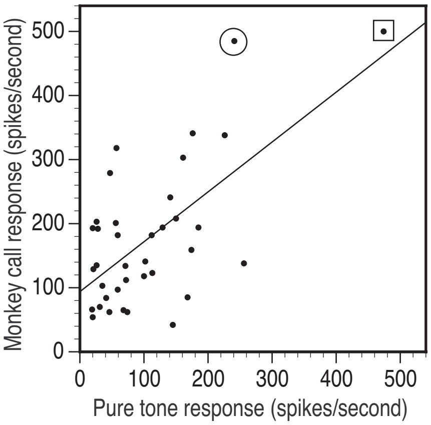
2.141. These results support the idea (the slope is negative, so variation decreases with increasing diversity), but the relationship is only moderately strong ($r^2 = 0.34$, so diversity only explains 34% of the variation in population variation).

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

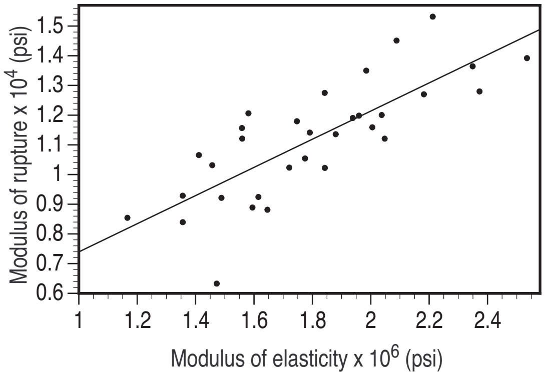


2.142. **(a)** A scatterplot of stock price growth against earnings growth shows a positive association, which supports the idea. Additionally, each y-value is fairly similar to its x-value, which indicates that stock price growth is roughly predicted by earnings growth (that is, $\hat{y} \approx x$)—this 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.

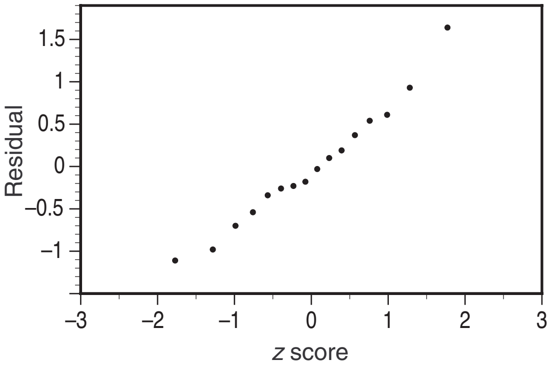
2.143. **(a)** One possible measure of the difference is the mean response: 106.2 spikes/second for pure tones and 176.6 spikes/second for monkey calls—an average of an additional 70.4 spikes/second. **(b)** The regression equation is $\hat{y} = 93.9 + 0.778x$. The third point (pure tone 241, call 485 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 $\hat{y} = 98.4 + 0.679x$.



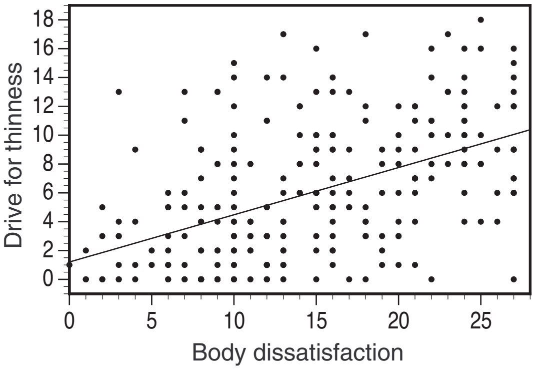
2.144. On the right is a scatterplot of MOR against MOE, showing a moderate, linear, positive association. The regression equation is $\hat{y} = 2653 + 0.004742x$; this regression explains $r^2 = 0.6217 \doteq 62\%$ of the variation in MOR. So, we can use MOE to get fairly good (though not perfect) predictions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



2.145. 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.

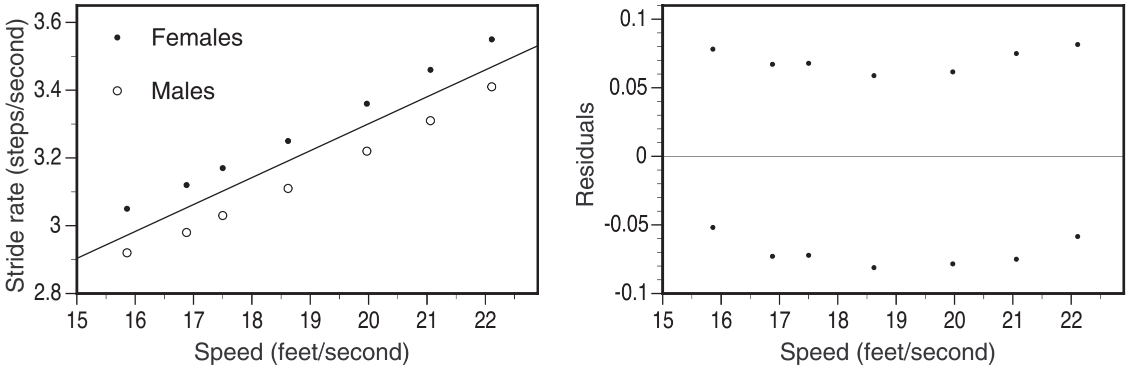


2.146. **(a)** The scatterplot is on the right. **(b)** The regression equation is $\hat{y} = 1.2027 + 0.3275x$. As we see from the scatterplot, the relationship is not too strong; the correlation ($r = 0.4916$, $r^2 = 0.2417$) confirms this.

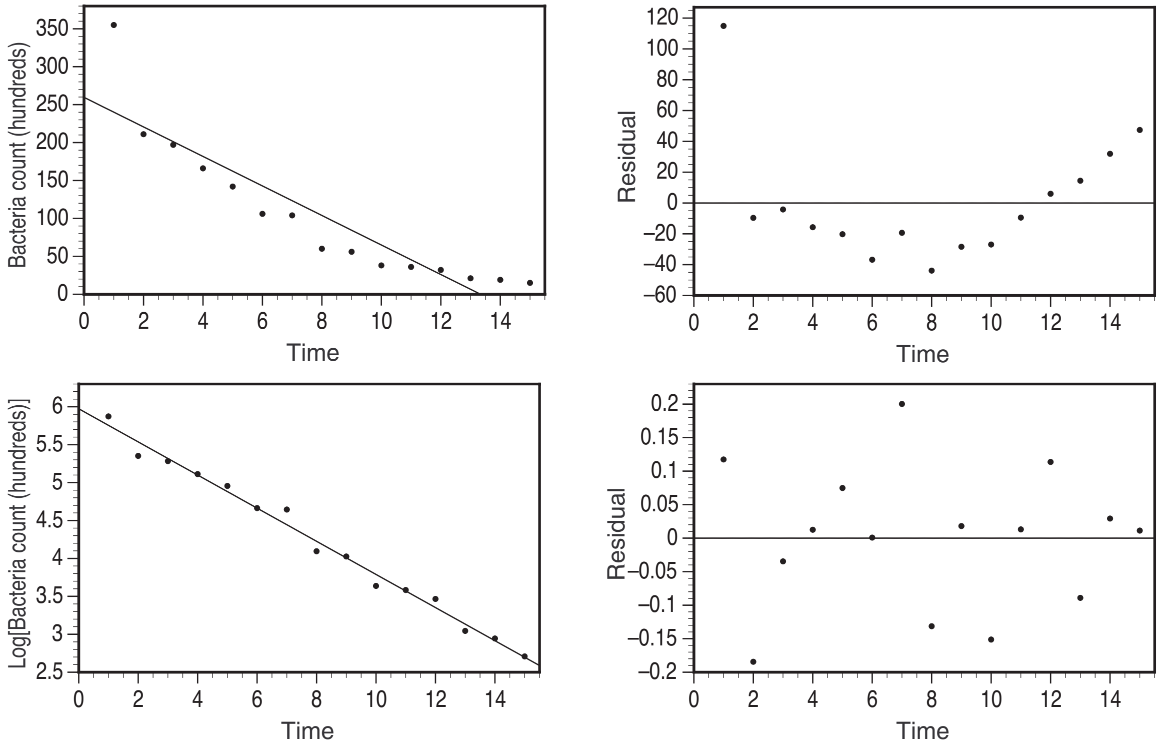


2.147. **(a)** Yes: The two lines appear to fit the data well. There do not appear to be any outliers or influential points. **(b)** Compare the slopes: before—0.189; after—0.157. (The units for these slopes are 100 ft^3/day per degree-day/day; for students who are comfortable with units, 18.9 ft^3 vs. 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.

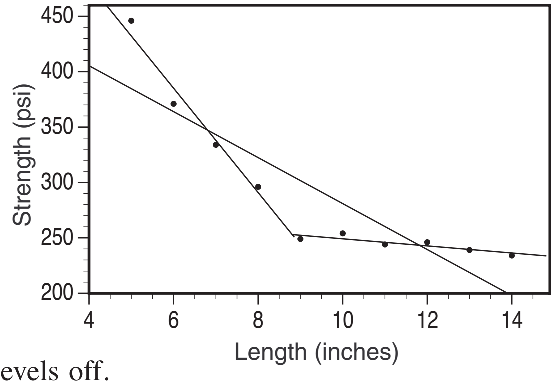
2.148. **(a)** Below, left. **(b)** The regression equation is $\hat{y} = 1.71 + 0.0795x$. **(c)** Below, right. The points for the residuals, like those of the original data, are split with women above the line (zero), and men below. (Men are taller on the average, so they have longer legs, and therefore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



2.149. **(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.



2.150. **(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 and then almost levels off.



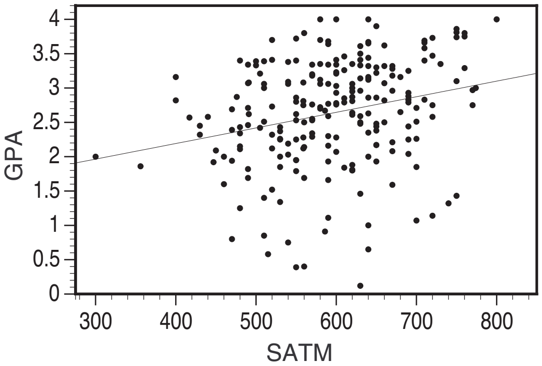
2.151. In the mid-1990s, European and American stocks were only weakly linked, but now it is more common for them to rise and fall together. Thus, investing in both types of stocks is not that much different from investing in either type alone.

2.152. 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.

2.153. Number of firefighters and amount of damage both increase with the seriousness of the fire (i.e., they are common responses to the fire's seriousness.)

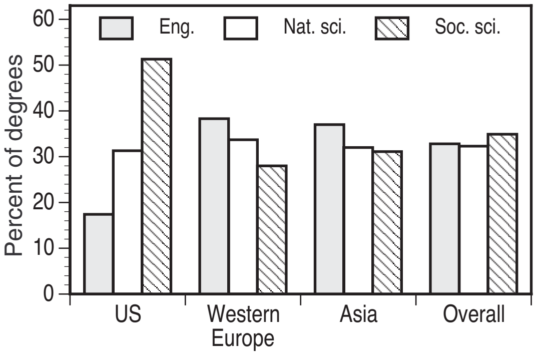
2.154. Note that $\bar{y} = 46.6 + 0.41\bar{x}$. We predict that Octavio will score 4.1 points above the mean on the final exam: $\hat{y} = 46.6 + 0.41(\bar{x} + 10) = 46.6 + 0.41\bar{x} + 4.1 = \bar{y} + 4.1$. (Alternatively, because the slope is 0.41, we can observe that an increase of 10 points on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

2.155. 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.

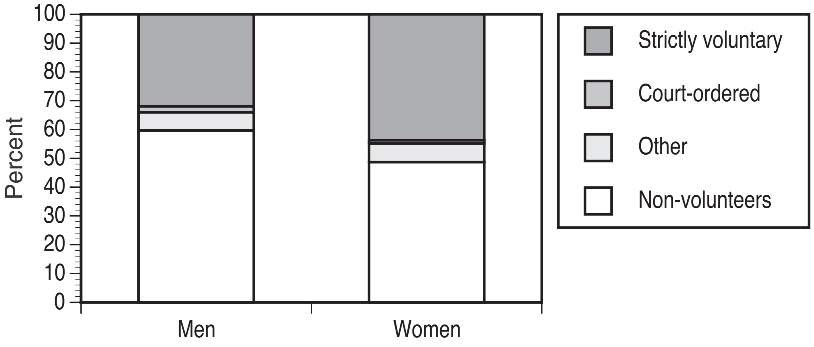


2.156. Compute column percents; for example, $\frac{61{,}941}{355{,}265} \doteq 17.44\%$ of those U.S. degrees considered in this table are in engineering. See table and graph at right. We observe that there are considerably more social science degrees and fewer engineering degrees in the United States. The Western Europe and Asia distributions are similar.

Field	United States	Western Europe	Asia	Overall
Eng.	17.44%	38.26%	36.96%	32.78%
Nat. sci.	31.29%	33.73%	31.97%	32.29%
Soc. sci.	51.28%	28.01%	31.07%	34.93%

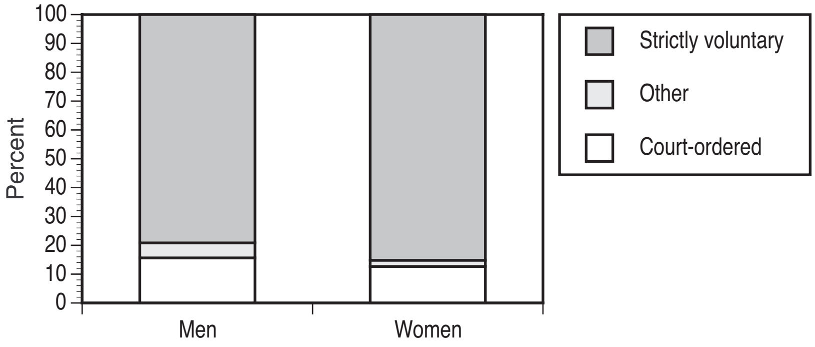


2.157. Different graphical presentations are possible; one is shown below. More women perform volunteer work; the notably higher percentage of women who are "strictly voluntary" participants accounts for the difference. (The "court-ordered" and "other" percentages are similar for men and women.)



2.158. Table shown on the right; for example, $\frac{31.9\%}{40.3\%} \doteq 79.16\%$. The percents in each row sum to 100%, with no rounding error for up to four places after the decimal. Both this graph and the graph in the previous exercise show that women are more likely to volunteer, but in this view, we cannot see the difference in the rate of non-participation.

Gender	Strictly voluntary	Court-ordered	Other
Men	79.16%	5.21%	15.63%
Women	85.19%	2.14%	12.67%



2.159. **(a)** At right. **(b)** $\frac{490}{800} = 61.25\%$ of male applicants are admitted, while only $\frac{400}{700} \doteq 57.14\%$ of females are admitted. **(c)** $\frac{400}{600} \doteq 66.67\%$ of male business school applicants are admitted; for females, this rate is the same: $\frac{200}{300} \doteq 66.67\%$. In the law school, $\frac{90}{200} = 45\%$ of males are admitted, compared to $\frac{200}{400} = 50\%$ of females. **(d)** A majority (6/7) of male applicants apply to the business school, which admits $\frac{400+200}{600+300} \doteq 66.67\%$ of all applicants. Meanwhile, a majority (3/5) of women apply to the law school, which admits only $\frac{90+200}{200+400} \doteq 48.33\%$ of its applicants.

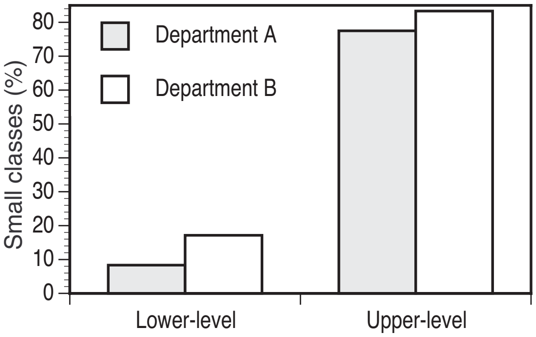
	Admit	Deny
Male	490	310
Female	400	300

2.160. Tables will vary, of course. The key idea is that one gender should be more likely to apply to the schools that are easier to get into. For example, if the four schools admit 50%, 60%, 70%, and 80% of applicants, and men are more likely to apply to the first two, while women apply to the latter two, women will be admitted more often.

A nice variation on this exercise is to describe two basketball teams practicing. You observe that one team makes 50% of their shots, while the other makes only 40%. Does that mean the first team is more accurate? Not necessarily; perhaps they attempted more lay-ups while the other team spent more time shooting three-pointers. (Some students will latch onto this kind of example much more quickly than discussions of male/female admission rates.)

2.161. If we ignore the "year" classification, we see that Department A teaches 32 small classes out of 52, or about 61.54%, while Department B teaches 42 small classes out of 106, or about 39.62%. (These agree with the dean's numbers.)

For the report to the dean, students may analyze the numbers in a variety of ways, some valid and some not. The key observations are: (i) When considering only first- and second-year classes, A has fewer small classes ($\frac{1}{12} \doteq 8.33\%$) than B ($\frac{12}{70} \doteq 17.14\%$). Likewise, when considering only upper-level classes, A has $\frac{31}{40} = 77.5\%$ and B has $\frac{30}{36} \doteq 83.33\%$ small classes. The graph on the right illustrates this. These numbers are given in the back of the text, so most students should include this in their analysis! (ii) $\frac{40}{52} \doteq 77.78\%$ of A's classes are upper-level courses, compared to $\frac{36}{106} \doteq 33.96\%$ of B's classes.



2.162. **(a)** The best numerical summary would note that we view target audience ("magazine readership") as explanatory, so we should compute the conditional distribution of model dress for each audience. This table and graph are shown below. **(b)** The sample is not an SRS: A set of magazines were chosen, and then all ads in three issues of those magazines were examined. It is not clear how this sampling approach might invalidate our conclusions, but it does make them suspect.

Minitab output

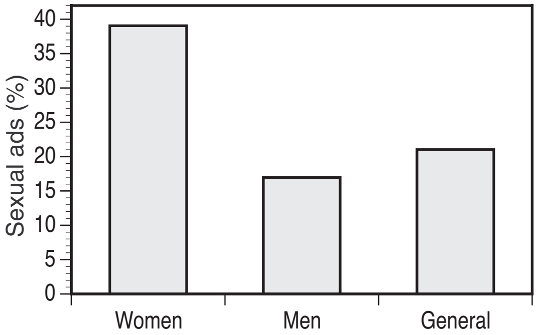
```
          Women      Men     Genl    Total
    1       351      514      248     1113
         424.84   456.56   231.60

    2       225      105       66      396
         151.16   162.44    82.40

Total       576      619      314     1509

ChiSq = 12.835 +  7.227 +  1.162 +
        36.074 + 20.312 +  3.265 = 80.874
df = 2, p = 0.000
```

	Magazine readership		
Model dress	Women	Men	General
Not sexual	60.94%	83.04%	78.98%
Sexual	39.06%	16.96%	21.02%



2.163. **(a)** As the conditional distribution of model dress for each age group has been given to us, it only remains to display this distribution graphically. One such presentation is shown below. **(b)** In order to perform the significance test, we must first recover the counts from the percents. For example, there were $(0.723)(1006) \doteq 727$ non-sexual ads in young adult magazines. The remainder of these counts can be seen in the Minitab output below, where we see $X^2 \doteq 2.59$, df $= 1$, and $P \doteq 0.108$—not enough evidence to conclude that age group affects model dress.

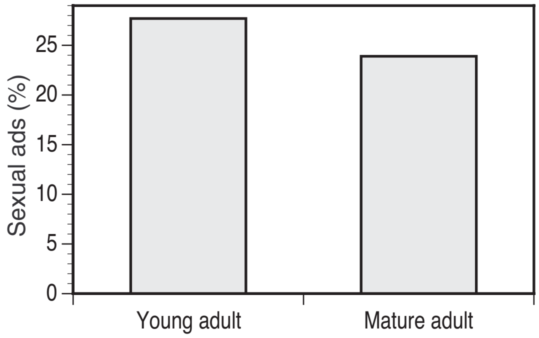
Minitab output

```
          Young   Mature    Total
    1       727      383     1110
         740.00   370.00

    2       279      120      399
         266.00   133.00

Total      1006      503     1509

ChiSq =  0.228 +  0.457 +
         0.635 +  1.271 = 2.591
df = 1, p = 0.108
```



2.164. **(a)** Subtract the "agreed" counts from the sample sizes to get the "disagreed" counts. The table is in the Minitab output on the right. (The output has been slightly altered to have more descriptive row and column headings.) We find $X^2 \doteq 2.67$, df $= 1$, and $P = 0.103$, so we cannot conclude that students and non-students differ in the response to this question. **(b)** For testing H_0: $p_1 = p_2$ vs. H_a: $p_1 \neq p_2$, we have $\hat{p}_1 \doteq 0.3607$, $\hat{p}_2 \doteq 0.5085$, $\hat{p} \doteq 0.4333$, $\text{SE}_{D_p} \doteq 0.09048$, and $z = -1.63$. Up to rounding, $z^2 = X^2$, and the P-values are the same. **(c)** The statistical tests in (a) and (b) assume that we have two SRSs, which we clearly do not have here. Furthermore, the two groups differed in geography (northeast/West Coast) in addition to student/non-student classification. These issues mean we should not place too much confidence in the conclusions of our significance test—or, at least, we should not generalize our conclusions too far beyond the populations "upper level northeastern college students taking a course in Internet marketing" and "West Coast residents willing to participate in commercial focus groups."

Minitab output

```
       Students   Non-st    Total
  Agr        22       30       52
          26.43    25.57

  Dis        39       29       68
          34.57    33.43

Total        61       59      120

ChiSq =  0.744 +  0.769 +
         0.569 +  0.588 = 2.669
df = 1, p = 0.103
```

2.165. **(a)** First we must find the counts in each cell of the two-way table. For example, there were about $(0.172)(5619) \doteq 966$ Division I athletes who admitted to wagering. These counts are shown in the Minitab output on the right, where we see that $X^2 \doteq 76.7$, df $= 2$, and $P < 0.0001$. There is very strong evidence that the percentage of athletes who admit to wagering differs by division. **(b)** Even with much smaller numbers of students (say, 1000 from each division), P is still very small. Presumably, the estimated numbers are reliable enough that we would not expect the true counts to be less than 1000, so we need not be concerned about the fact that we had to estimate the sample sizes. **(c)** If the reported proportions are wrong, then our conclusions may be suspect—especially if it is the case that athletes in some division were more likely to say they had not wagered when they had. **(d)** It is difficult to predict exactly how this might affect the results: Lack of independence could cause the estimated percents to be too large, or too small, if our sample included several athletes from teams which have (or do not have) a "gambling culture."

Minitab output

```
            Div1     Div2     Div3    Total
    1        966      621      998     2585
         1146.87   603.54   834.59

    2       4653     2336     3091    10080
         4472.13  2353.46  3254.41

Total       5619     2957     4089    12665

ChiSq = 28.525 +  0.505 + 31.996 +
         7.315 +  0.130 +  8.205 = 76.675
df = 2, p = 0.000
```