Chapter 7 – Exploring Data: Part I Review

Test Yourself exercise answers are sketches. All of these problems are similar to ones found in Chapters 1–6, for which the solutions in this manual provide more detail.

7.1 (c) sex and marital status are categorical variables.

7.2 Answers will vary. Some suggestions for questions with a categorical response: "What is your class level (freshman, sophomore, etc.)?" or "Is this your first statistics class?" Some suggestions for questions with a quantitative response: "How many hours per week do you work at a paying job?" or "How many hours do you spend studying in a typical week?"

7.3 (d) is all of the above.

7.4 (b) between 1000 and 2000.

7.5 (b) about 65%.

7.6 (a) The mean.

7.7 (b) roughly symmetric with one high outlier.

7.8 (c) 41.5%.

7.9 (c) 46%.

7.10 More than half of all American households do not carry credit card debt, but some have a great deal of credit card debt.

7.11 The units of measurement are: **(a)** centimeters; **(b)** centimeters; **(c)** centimeters; **(d)** grams^2.

7.12 (b) 130 millimeters.

7.13 (c) Strong El Niño years tend to have lower monsoon rainfalls than in other years.

7.14 (c) 8%.

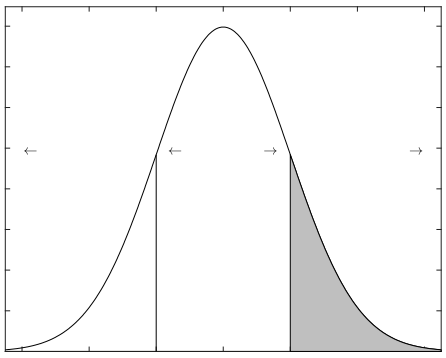
7.15 (b) 25%.

7.16 (a) By definition, the 85th percentile requires 85% of the scores to be lower than this score. Thus, $z = 1.04$, because the area to the left of 1.04 under the standard Normal curve is 0.8508 (software gives $z = 1.0364$). Then, unstandardizing, we have $x = 511 + (1.04)120 = 635.8$ points. **(b)** Joseph's z-score is $z = \frac{451 - 511}{120} = -0.5$. From Table A, the area below z = –0.5 is 0.3085. He scored better than about 31% of all MCAT takers. **(c)** The first quartile is equivalent to the

25th percentile. Thus, $z = -0.67$, because the area under the standard normal curve to the left of -0.67 is 0.2500 (software gives $z = -0.6745$). Then, unstandardizing, we have $x = 511 + (-0.67)120 = 430.6$ points.

7.17 (a) $40 < x < 50$ corresponds to $\frac{40 - 44.8}{2.1} < z < \frac{50 - 44.8}{2.1}$, or $-2.29 < z < 2.48$. This proportion is 0.9934 – 0.0110 = 0.9824 = 98.24%. **(b)** Approximately 95% of all values are within 2σ of μ in a Normal distribution; this becomes $44.8 - (2)2.1 = 40.6$ inches to $44.8 + (2)2.1 = 49$ inches. **(c)** The 70th percentile is the height where 70% of the heights are less than this value. Thus, $z = 0.52$, because the area to the left of 0.52 under the standard Normal curve is 0.6985 (software gives $z = 0.5244$). Then, unstandardizing, we have $x = 44.8 + (0.52)2.1 = 45.892$ inches.

7.18 (a) Approximately 99.7% of all values are within 3σ of μ in a Normal distribution; this becomes $266 - (3)16 = 218$ days to $266 + (3)16 = 314$ days. **(b)** $x > 282$ corresponds to $z > \frac{282 - 266}{16} = 1$. The provided Normal curve demonstrates that the percent greater than $z = 1$ is $\frac{100 - 68}{2} = 16\%$.



7.19 About 3.76%: Slots meeting specifications correspond to $0.8725 < x < 0.8775$, which for the $N(0.8750, 0.0012)$ distribution corresponds to $\frac{0.8725 - 0.8750}{0.0012} < z < \frac{0.8775 - 0.8750}{0.0012}$, or $-2.08 < z < 2.08$, for which Table A gives 0.9812 − 0.0188 = 0.9624. Thus, the proportion of slots that do not meet these specifications is 1 - 0.9624 = 0.0376.

7.20 (a) $x < 50$ corresponds to $z < \frac{50 - 69}{8.5}$, or $z < -2.24$, for which Table A gives 0.0125, or 1.25%. **(b)** $x > 85$ corresponds to $z > \frac{85 - 69}{8.5}$, or $z > 1.88$, for which Table A gives (1 – 0.9699) = 0.0301, or 3.01%. **(c)** The central 80% of data corresponds to $z = -1.28$ (the $\frac{100 - 80}{2} = 10$th percentile) and $z = 1.28$ (the $80 + \frac{100 - 80}{2} = 90$th percentile). Unstandardizing gives that $-1.28 < z < 1.28$ corresponds to $69 + (-1.28)8.5 < x < 69 + (1.28)8.5$, or $58.12 < x < 79.88$ beats per minute.

7.21 (a) Minimum = 7.2, Q_1 = 8.5, M = 9.3, Q_3 = 10.9, and Maximum = 12.8. **(b)** M = 27. **(c)** 25% of values exceed Q_3 = 30. **(d)** Yes—virtually all Torrey pine needles are longer than virtually all Aleppo pine needles. There is no overlap in the distributions, as seen by comparing, say, the minimum for Torrey pine needles (about 21) to the maximum for Aleppo pine needles (12.8).

7.22 (c) A very weak association.

7.23 (a) A team gains about 0.85 point per million dollars spent.

7.24 (c) 84.7.

7.25 (c) the prediction is not sensible because no money is far outside the range of values of the explanatory variable.

7.26 (b) –0.6.

7.27 (d) For each degree increase in mean sea surface temperature, the predicted mean coral growth of a reef decreases by 0.22 centimeter.

7.28 (c) 0.82.

7.29 (d) We conclude that memory of food intake in the distant past is fair to poor.

7.30 (b) The Fidelity Technology Fund has a closer relationship to returns from the stock market as a whole, but we cannot say that it has higher returns than the Fidelity Real Estate Fund.

7.31 (d) 0.6.

7.32 (c) There are one or two outliers and at least one of these may also be influential.

7.33 (c) 17.4%.

7.34 (b) 20.9%.

7.35 (a) A greater percentage of females spend five or more hours per day playing video or computer games or using the computer for something that is not school work, on an average school day, than males.

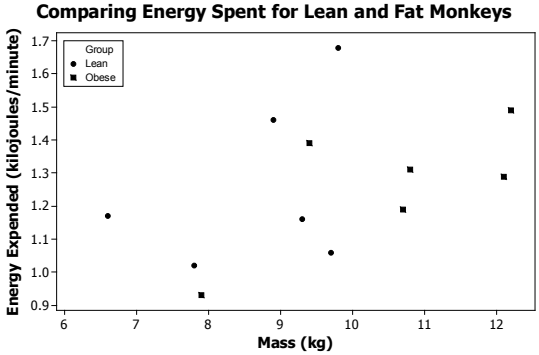
7.36 The increased correlation suggests that the two types of stocks (American and European) now tend to rise together and fall together, which reduces the ability of one to hedge risk of the other.

7.37 (a) No. **(b)** r^2 = 0.64, or 64%.

7.38 (a) Using the summary statistics, b = 0.448(8.69/0.045) = 86.514; a = 65.897 – 86.514(0.649); the least-squares regression line is given by $\hat{y}$ = 9.749 + 86.514x.

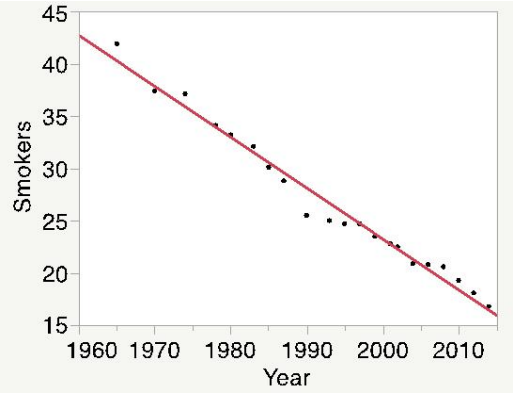
(b) If $x = 0.60$, we predict $\hat{y} = 9.749 + 86.514\,(0.60) = 61.7$ volume. **(c)** For $x = 0.99$, we predict $\hat{y} = 95.4$ volume. Since $r^2 = (0.448)^2 = 0.201$, only 20.1% of the variation in brain volume is explained by our regression model. Predictions using this model aren't very reliable. Also, because the largest value for introspection is 0.75 in the data set, using the regression equation for $x = 0.99$ would be extrapolation.

7.39 (a) 8.683 kg. **(b)** 10.517 kg. **(c)** Such a comparison would be unreasonable, because the lean group is less massive and, therefore, would be expected to burn less energy on average. **(d)** The scatterplot is provided.



(e) Based on the plot, it appears that the rate of increase in energy burned per kilogram of mass is about the same for both groups. The obese monkeys burn less energy than the lean monkeys, because their points tend to be below the others. Do they expend less energy because they are obese, or are they obese because they expend less energy?

7.40 (a) The scatterplot is provided, along with the regression line for part (c).



(b) There is a very strong, negative linear relationship between year and percent of smokers. There are no real outliers, but the rate of decline in smoking has varied over time. From 1965 to the late 1980s, for example, there was a very sharp decline, while in the 1990s the decline slowed. **(c)** Using the summary statistics, $b = -0.99(7.0/14.3) = -0.485$, $a = 26.7 - (-0.485)1993.0 = 993.305$; the least-

squares regression line is given by $\hat{y} = 993.305 - 0.485x$. See the scatterplot for the regression line. This matches the line provided in the JMP output, to rounding error.

Linear Fit

Smokers = 995.00419 - 0.4858717*Year

Summary of Fit

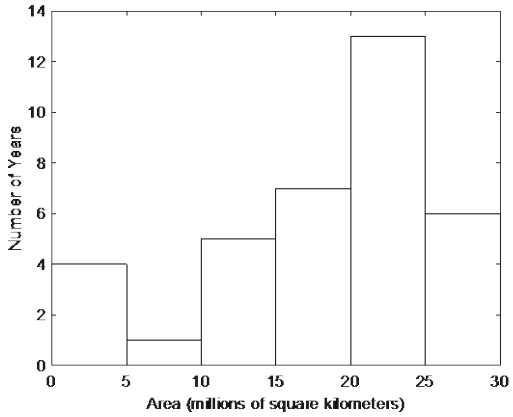
RSquare	0.980181
RSquare Adj	0.979138
Root Mean Square Error	1.012925
Mean of Response	26.6619
Observations (or Sum Wgts)	21

(d) Over this time, on average, smoking declined by 0.485% per year (almost 1% every two years). **(e)** $r^2 = (-0.99)^2 = 0.9801$, or 98.01%. See the JMP output for confirmation. **(f)** In the year 2020, we predict that $993.305 - 0.485(2020) =$ 13.605% of adults will smoke. The goal would not be achieved. **(g)** $993.305 - 0.485(2075) = -13.070\%$. This negative percent is impossible to achieve. Such a prediction is foolish, because the year 2075 is far outside the range of years on which this model was based. This would be terrible extrapolation.

7.41 (a) 190/8474 = 0.0224, or 2.24%. **(b)** 633/8474 = 0.0747, or 7.47%. **(c)** 27/633 = 0.0427, or 4.27%. **(d)** 4621/8284 = 0.5578, or 55.78%. **(e)** The conditional distribution of CHD for each level of anger is tabulated below. The result for the high anger group was computed in part (c), for example. Clearly, angrier people are at greater risk of CHD.

Low anger	Moderate anger	High anger
1.70%	2.33%	4.27%

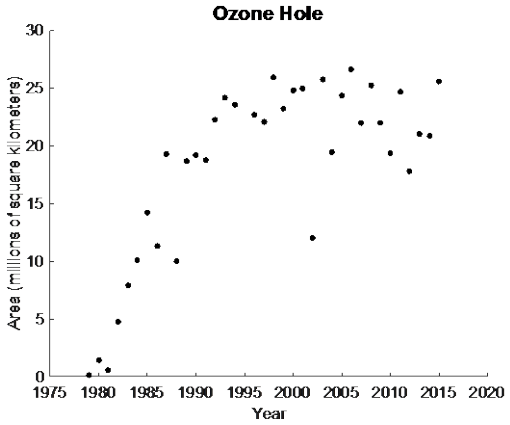
7.42(a) A histogram is provided. The distribution is skewed to the left and has several observations with small areas.



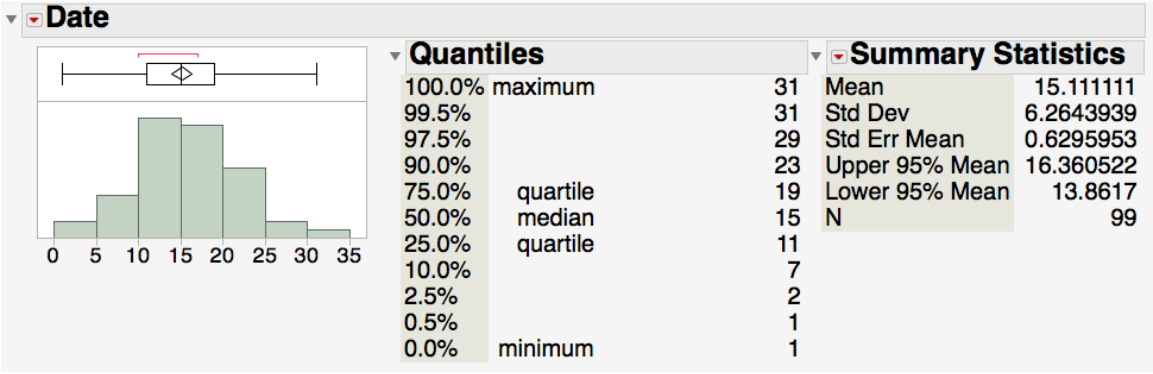
(b) Because the distribution is skewed to the left, we expect the mean to be clearly less than the median. This is, in fact, the case; the mean is 18.253 millions of km^2 and the median is 20.95 millions of km^2.

7.43 PLAN: Make a time plot to show how the size of the ozone hole changed

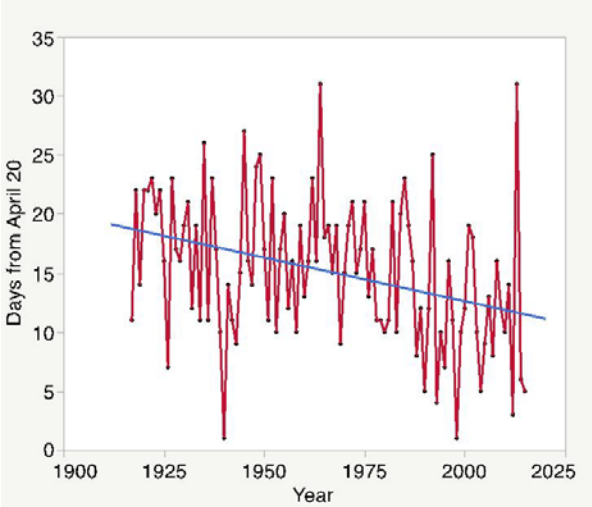
between 1979 and 2015. SOLVE: The time plot is provided. CONCLUDE: In addition to year-to-year variation, the time plot shows two distinct trends between 1979 and 2015. From 1979 to the mid 1990's, there is a very strong, positive linear relationship between the year and the ozone hole area. The slope of a regression line fit to this portion of the data would be quite large. From 1995 to 2015, the relationship between the year and the ozone hole area is quite different; here the linear relationship is much weaker and is negative. Additionally, there is a low outlier for the 2002 data point. No cyclical fluctuation is present.



7.44 JMP output is given below. The distribution is roughly symmetric. Based on the histogram and boxplot, we will not call any values outliers. The mean, standard deviation, and five-number summary (all in days) are $\bar{x}$ = 15.111, s = 6.264, and Min = 1, Q_1 = 11, M = 15, Q_3 = 19, and Max = 31. The median date is May 4 (day 15).



7.45 (a) The plot is provided, along with the regression line for part (b).



(b) From the JMP output, it can be seen that the least-squares regression line is $\hat{y}$ = 160.11 – 0.0738x. The slope is negative, suggesting that the ice breakup day is decreasing (by 0.0738 day per year). See the plot for the regression line.

Linear Fit

Date = 160.11121 - 0.0737539*Year

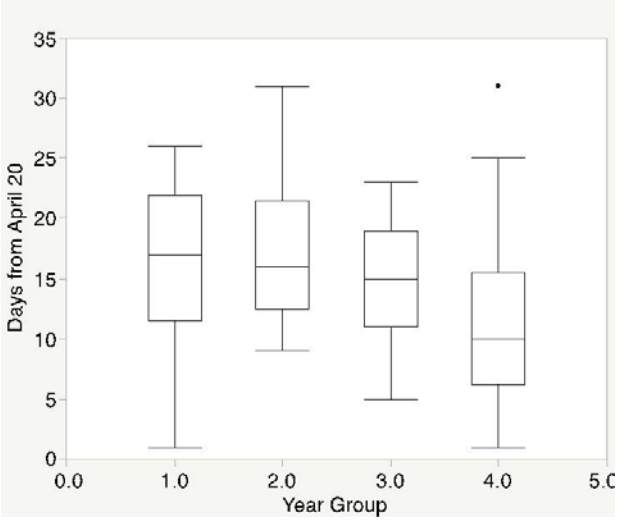
Summary of Fit

RSquare	0.114358
RSquare Adj	0.105227
Root Mean Square Error	5.925642
Mean of Response	15.11111
Observations (or Sum Wgts)	99

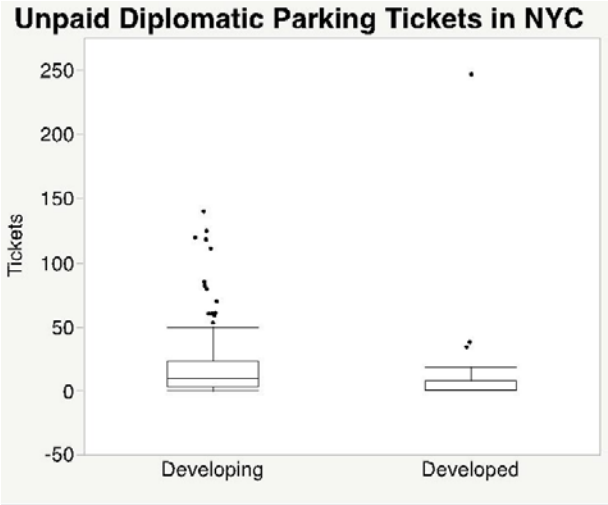
(c) The regression line is not very useful for prediction, as it accounts for only about 11% (r^2 = 0.1144) of the variation in ice breakup time.

7.46 The five-number summaries and the boxplots are provided. Clearly, breakup has tended to come earlier in the 1992–2015 time segment (as evidenced by the median and quartiles), but the minimum 1 day also occurred in the 1917–1941 time segment. The year 2013 was a high outlier.

	Min	Q_1	Med	Q_3	Max
1917–1941	1	11.5	17	22.0	26
1942–1966	9	12.5	16	21.5	31
1967–1991	5	11.0	15	19.0	23
1992–2015	1	6.25	10	15.5	31



7.47 PLAN: Create side-by-side boxplots to compare the distributions for countries identified as "developing" and "developed" and compute appropriate summary statistics. SOLVE: Both groups (developing countries and developed countries) have right-skewed distributions for unpaid parking tickets. Undeveloped countries have more outliers than developed countries, but the most unpaid tickets are for a developed country (Kuwait). Because of the outliers, the five-number summary is appropriate for these distributions. CONCLUDE: Comparing the distributions, developing countries' diplomats tend to have more unpaid tickets (More than 75% of developed countries have fewer unpaid tickets than the median number of developing countries.) National income alone, however, does not explain countries whose diplomats have more or fewer unpaid tickets; the country with the largest number of unpaid tickets is classified as "developed," but it is an Arab emirate; perhaps the culture there has an impact on how their diplomats view a parking ticket.



	Min	Q_1	M	Q_3	Max
Developed	0	0	0.7	8.13	246.2
Developing	0	3.2	9.5	22.8	139.6

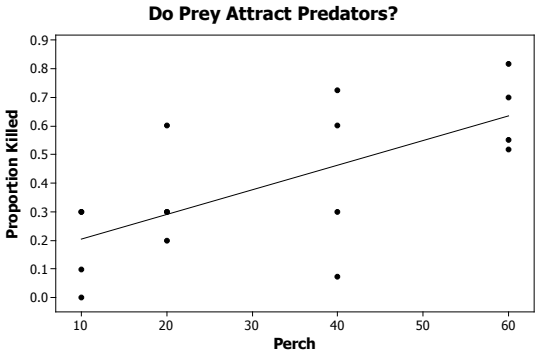
7.48 PLAN: Compare the seed masses for the two groups of plants (with and without cicadas), both graphically (using stemplots, histograms, or boxplots) and numerically (with appropriate statistics). SOLVE: Back-to-back stemplots are given with the thousands place being truncated, the stems representing the tenths place, and the leaves representing the hundredths place. These plots show little difference overall. Both shapes are somewhat irregular, but neither is clearly higher or lower. Means and medians are also similar. CONCLUDE: The data give little reason to believe that cicadas make good fertilizer, at least on the basis of this experiment.

Cicada plants		Control plants
0	1	
	1	3
4	1	445
7	1	77
99	1	89999
111100	2	0111
3333332222	2	2
5544	2	4444445555
7777666	2	66666
999	2	89
110	3	
	3	
5	3	

	$\bar{x}$	M
Cicada group	0.2426 mg	0.2380 mg
Control group	0.2221 mg	0.2410 mg

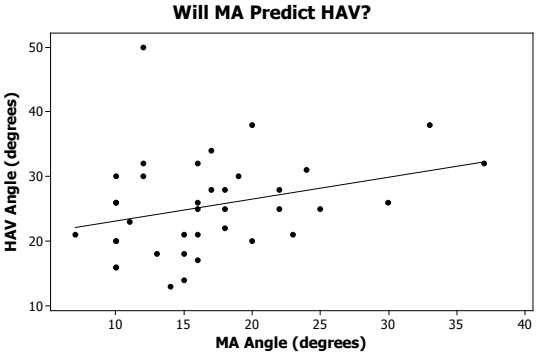
7.49 PLAN: Display the distribution with a graph, and compute appropriate numerical summaries. SOLVE: A stemplot is shown; a histogram could also be used. The distribution seems to be fairly Normal, apart from a high outlier of 50°. The five-number summary is preferred because of the outlier: Min = 13°, Q_1 = 20°, M = 25°, Q_3 = 30°, and Max = 50°. (The mean and standard deviation are $\bar{x}$ = 25.4211° and s = 7.4748°.) CONCLUDE: Student descriptions of the distribution will vary. Most patients have a deformity angle in the range of 15° to 35°.

7.50 PLAN: We will examine the relationship with a scatterplot and (if appropriate) provide correlation and regression lines. SOLVE: The scatterplot suggests a positive linear association, albeit with lots of scatter, so correlation and regression are reasonable tools to summarize the relationship. The correlation is r = 0.682, and the least-squares regression line is $\hat{y}$ = 0.1205 + 0.008569x. The regression line explains r^2 = 46.5% of the variation in the proportion killed. CONCLUDE: The analysis

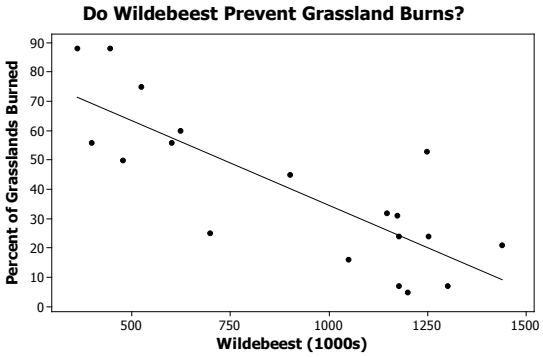


provides weak support for the idea that the proportion of perch killed rises with the number of perch present.

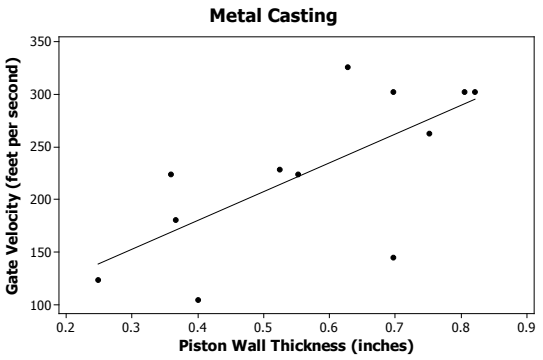
7.51 PLAN: We examine the relationship with a scatterplot and (if appropriate) a correlation and regression line. SOLVE: MA angle is the explanatory variable, so it should be on the horizontal axis of the scatterplot. The scatterplot shows a moderate to weak positive linear association, with one clear outlier (the patient with HAV angle 50°). The correlation is $r = 0.302$, and the regression line is $\hat{y} = 19.723 + 0.3388x$. CONCLUDE: MA angle can be used to give (very rough, imprecise) estimates of HAV angle, but the spread is so wide that the estimates would not be very reliable. The linear relationship explains only $r^2 = 9.1\%$ of the variation in HAV angle.



7.52 PLAN: We will examine the relationship with a scatterplot and (if appropriate) a correlation and regression line. SOLVE: The scatterplot suggests a fairly strong, negative linear association, so correlation and regression are reasonable tools to use here. The correlation is $r = -0.803$, and the regression equation is $\hat{y} = 92.29 - 0.05762x$; the equation explains $r^2 = 64.6\%$ of the variation in burned grassland. CONCLUDE: The claim is supported: When wildebeest numbers are higher, the percent of grassland burned tends to be lower. Each additional 1000 wildebeest decrease burned area by about 0.058% on the average.

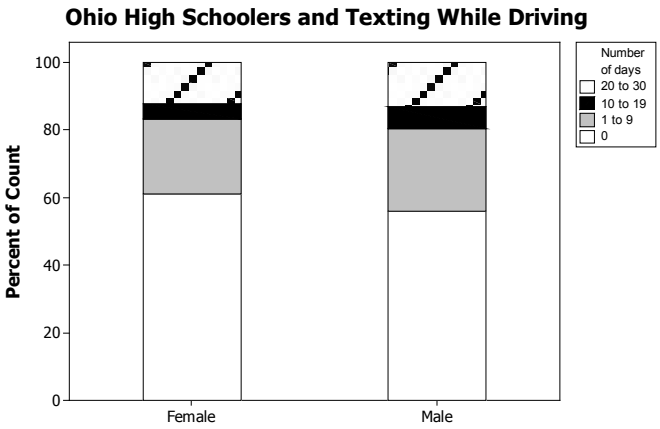


7.53 PLAN: We will examine the relationship with a scatterplot and (if appropriate) a correlation and regression line. SOLVE: The scatterplot, shown with the regression line $\hat{y} = 70.44 + 274.78x$, shows a moderate, positive linear relationship. The linear relationship explains about $r^2 = 49.3\%$ of the variation in gate velocity. CONCLUDE: The regression formula might be used as a rule of thumb for new workers to follow, but the wide spread in the scatterplot suggests that there

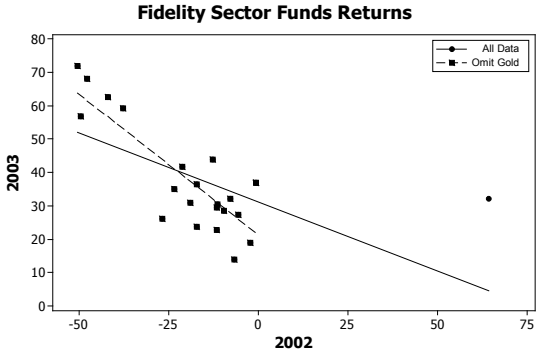


may be other factors that should be taken into account in choosing the gate velocity.

7.54 PLAN: We will compare males and females using a segmented bar graph of appropriate conditional distributions. SOLVE: Student analyses will vary, as will graph choices. Some possible observations: Over half of both sexes report not having texted (or emailed) while driving during the past 30 days. Males are somewhat more likely to have texted or emailed at least once (their relative percents in each category that represents having texted are more than for females, especially in the 1 to 9 category). CONCLUDE: Student conclusions will vary also.



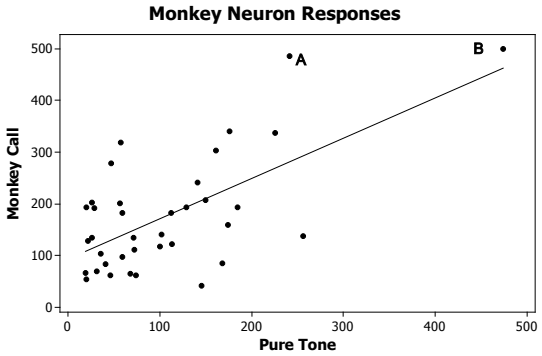
7.55 (a) The scatterplot of 2003 returns against 2002 returns shows (ignoring the outlier) a strong negative association.



(b) The correlation for all 23 points is $r = -0.616$; with the outlier removed, $r = -0.838$. The outlier deviates from the linear pattern of the other points; removing it makes the negative association stronger, and so r moves closer to –1. **(c)** Regression formulas are given in the table. The first line is solid in the plot; the second is the dashed line. The least-squares regression line makes the sum of the squares of the vertical deviations of the points from the line as small as possible. The line for the 22 other funds is so far below Fidelity Gold that the squared deviation is very large. The line must pivot up toward Fidelity Gold in order to minimize the sum of squares for all 23 deviations. Fidelity Gold is very influential.

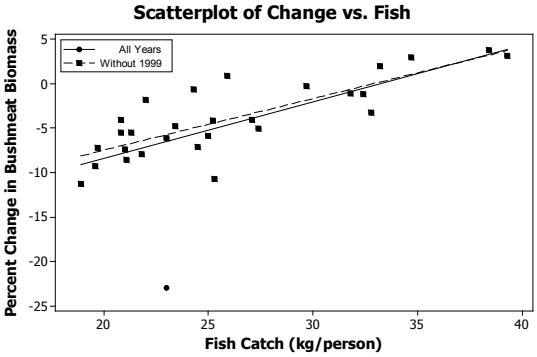
	r	Equation
All 23 funds	−0.616	$\hat{y} = 31.1167 - 0.4132x$
Without Gold	−0.838	$\hat{y} = 21.4616 - 0.8403x$

7.56 (a) The regression equation is $\hat{y} = 93.92 + 0.7783x$. The third point (pure tone 241, call 485 spikes/second) is A. The first point (474 and 500 spikes/second) is B.



(b) The correlation drops only slightly (from 0.639 to 0.610) when A is removed; it drops more drastically (to 0.479) without B. **(c)** When either point is removed, the slope decreases. Without A, the line is $\hat{y} = 98.42 + 0.6792x$; without B, it is $\hat{y} = 101.1 + 0.6927x$.

7.57 (a) Fish catch (on the horizontal axis) is the explanatory variable. The point for 1999 is at the bottom of the plot.



(b) The correlations are given in the table below. The outlier decreases r because it weakens the strength of the association.

	r	Equation
All points	0.672	$\hat{y} = -21.09 + 0.6345x$
Without 1999	0.804	$\hat{y} = -19.05 + 0.5788x$

(c) The two regression lines are given in the plot in part (a); the solid line in the plot uses all points, while the dashed line omits the outlier. The equations for these lines are given in the table in part (b). The plot in part (a) shows that the effect of the outlier on the line is small. This occurs because there are several other years with similar changes in bushmeat biomass. Also, this year was not particularly extreme in the amount of fish caught.

7.58 (a) The two-way table is provided.

	Temperature		
	Cold	Neutral	Hot
Hatched	16	38	75
Did not hatch	11	18	29

(b) PLAN: Compare the conditional distribution of hatching given temperature. SOLVE: In order of increasing temperature, the proportions hatching are 16/27 = 0.593, or 59.3%; 38/56 = 0.679, or 67.9%; and 75/104 = 0.721, or 72.1%. (We could also construct a bar graph of these percents.) CONCLUDE: The percent hatching increases with temperature; the cold temperature did not prevent hatching but made it less likely. The difference between the percents hatching at hot and neutral temperatures is fairly small, and may not be big enough to be called significant. (Statistical tests say that it is not.)

7.59 (a) There are two somewhat low IQs: 72 qualifies as an outlier by the 1.5 × *IQR* rule, while 74 is on the boundary. However, for a small sample, this stemplot looks reasonably Normal.

7	24
7	
8	
8	69
9	13
9	68
10	023334
10	578
11	11222444
11	89
12	0
12	8
13	02

(b) We compute $\bar{x}$ = 105.84 and s = 14.27, and we find 23/31 = 74.2% of the scores in the range $\bar{x} \pm 1s$, or 91.6 to 120.1, and 29/31 = 93.5% of the scores in the range $\bar{x} \pm 2s$, or 77.3 to 134.4. For an exactly Normal distribution, we would expect these proportions to be 68% and 95%. Given the small sample, this is reasonably close agreement.