Statistics 6126

Brief Solutions to Homework Exercises

These solutions are solely for the use of students in STA 6126 at the University of Florida and are not to be distributed elsewhere. Please report any errors in these solutions to Alan Agresti, aa@stat.ufl.edu, so they can be corrected.

Chapter 1

1.2. (a) Population was all 7 million voters, and sample was 2705 voters in exit poll. (b) A statistic is the 56.5% who voted for Schwarzenegger from the exit poll sample of size 2705; a parameter is the 55.9% who actually voted for Schwarzenegger.

1.3. (a) All students at the University of Wisconsin. (b) A statistic, since it's calculated only for the 100 sampled students.

1.5. (a) All adult Americans. (b) Proportion of all adult Americans who would answer definitely or probably true. (c) The sample proportion 0.523 estimates the population proportion. (d) No, it is a *prediction* of the population value but will not equal it exactly, because the sample is only a very small subset of the population.

1.17. (a) A statistic is the 45% of the sample of subjects interviewed in the UK who said *yes*. (b) A parameter is the true percent of the 48 million adults in the UK who would say *yes*. (c) A descriptive analysis is that the percentage of *yes* responses in the survey varied from 10% (in Bulgaria) to 60% in Luxembourg). (d) An inferential analysis is that the percentage of adults in the UK who would say *yes* falls between 41% and 49%.

Chapter 2

2.1. (a) Discrete variables take a set of separate numbers for their values (such as nonnegative integers). Continuous variables take an infinite continuum of values. (b) Categorical variables have a scale that is a set of categories; for quantitative variables, the measurement scale has numerical values that represent different magnitudes of the variable. (c) Nominal variables have a scale of *unordered* categories, whereas ordinal variables have a scale of *ordered* categories. The distinctions among types of variables are important in determining the appropriate descriptive and inferential procedures for a statistical analysis.

2.2. (a) Quantitative (b) Categorical (c) Categorical (d) Quantitative (e) Categorical (f) Quantitative (g) Categorical (h) Quantitative (i) Categorical

2.3. (a) Ordinal (b) Nominal (c) Interval (d) Nominal (e) Nominal (f) Ordinal (g) Interval (h) Ordinal (i) Nominal (j) Interval (k) Ordinal

2.5. (a) Interval (b) Ordinal (c) Nominal

2.7. (a) Ordinal, since there is a sense of order to the categories. (b) Discrete, since separate values rather than continuum of numbers. (c) These values are statistics since they come from a sample.

2.13. (a) Observational study (b) Experiment (c) Observational study (d) Experiment

2.14. (a) Experimental study, since the researchers are assigning subjects to treatments. (b) An observational study could observe those who grew up in nonsmoking or smoking environments and examine incidence of lung cancer for each group.

2.27. (a) This is a volunteer sample, so results are unreliable; e.g., there is no way of judging how close 93% is to the actual population who believe that benefits should be reduced. (b) This is a volunteer sample; perhaps an organization opposing gun control laws has encouraged members to send letters, resulting in a distorted picture for the congresswoman. The results are completely unreliable as a guide to views of the overall population. She should take a probability sample of her constituents to get a less biased reaction to the issue. (c) The physical science majors who take the course might tend to be different from the entire population of physical science majors (perhaps more liberal minded on sexual attitudes, for example). Thus, it would be better to take random samples of students of the two majors from the population of all social science majors and all physical science majors at the college. (d) There would probably be a tendency for students within a given class to be more similar than students in the school as a whole. For example, if the chosen first period class consists of college-bound seniors, the members of the class will probably tend to be less opposed to the test than would be a class of lower achievement students planning to terminate their studies with high school. The design could be improved by taking a simple random sample of students, or a larger random sample of classes with a random sample of students then being selected from each of those classes (a two-stage random sample).

2.34. (b)

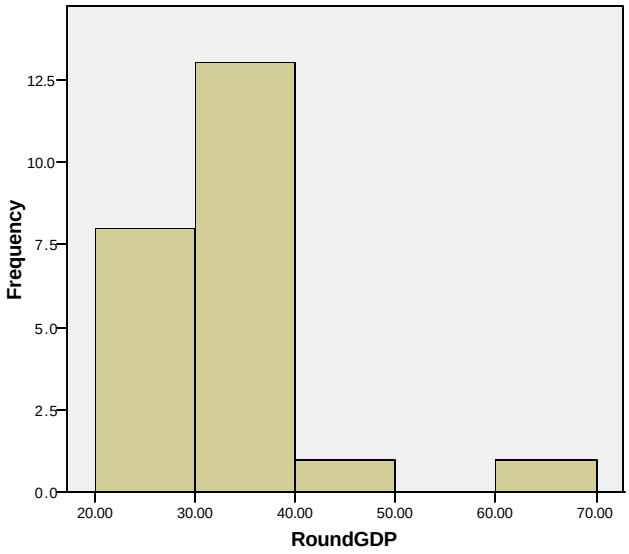
2.36. (c)

Chapter 3

3.6. (a) GDP is rounded to the nearest thousand

Stem (10 thousands)	Leaves (thousands
2	023
2	58899
3	00011122233
3	89
4	0
4	
5	
5	
6	
6	
7	0

(b)



(c) The outlier in each plot is Luxembourg.

3.10. (a)

Stem	Leaves
0	4679
1	133
2	0
3	9
4	4

(b) The mean is 16.6 days, and the standard deviation is 13.9.

3.11. (a)

TV Hours	Frequency	Relative Frequency
0	79	4.0
1	422	21.2
2	577	29.0
3	337	17.0
4	226	11.4
5	136	6.8
6	99	5.0
7	23	1.2
8	34	1.7
9	4	0.2
10	23	1.2
12	14	0.7
13	1	0.1
14	7	0.4
15	2	0.1
18	2	0.1
24	1	0.1
Total	1987	100.0

(b) The distribution is unimodal and right skewed. (c) The median is the 994^{th} data value, which is 2. (d) The mean is larger than 2 because the data is skewed right by a few high values.

3.19. (a) Median: \$10.13; mean: \$10.18; range: \$0.46; standard deviation: \$0.22. (b) Median: \$10.01; mean: \$9.17; range: \$5.31; standard deviation: \$2.26. The median is resistant to outliers, but the mean, range, and standard deviation are highly impacted by outliers.

3.22. (a) The life expectancies in Africa vary more than the life expectancies in Western Europe, because the life expectancies for the African countries are more spread out than those for the Western European countries. (b) The standard deviation is 1.1 for the Western European nations and 7.1 for the African nations.

3.24. (a) Approximately 68% of the values are contained in the interval 32 to 38 days; approximately 95% of the values are contained in the interval 29 to 41 days; all or nearly all of the values are contained in the interval 26 to 44 days. (b) (i) The mean would decrease if the observation for the U.S. was included. (ii) The standard deviation would increase if the observation for the U.S. was included. (c) The U.S. observation is 5.3 standard deviations below the mean.

3.25. (a) 88.8% of the observations fall within one standard deviation of the mean. (b) The Empirical Rule is not appropriate for this variable, since the data are highly skewed to the right.

3.28. (d)

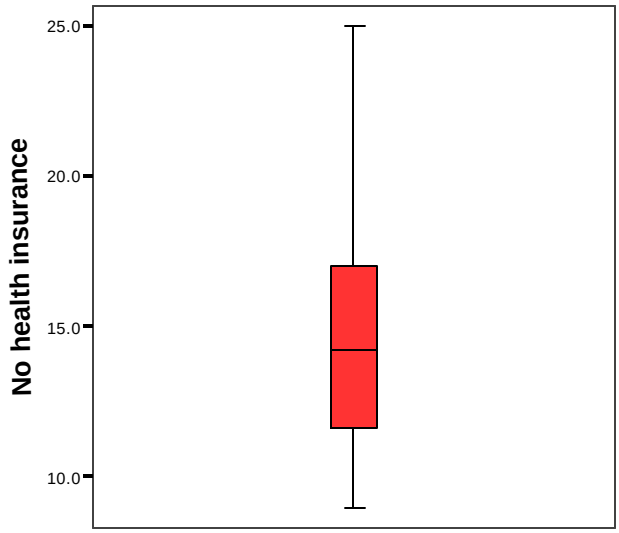
3.30. The distribution is most likely skewed to the right since the minimum water consumption (0 thousands of gallons) is less than one standard deviation below the mean.

3.33. The mean, standard deviation, maximum, and range all increase, because the observation for D.C. was a high outlier. Note that these statistics are not resistant to outliers. On the other hand, the median, Q3, Q1, the interquartile range, and the mode remain the same, as these are all resistant to outliers. The minimum remains the same since D.C. was a high outlier and not a low outlier.

3.35. (a) The sketch should show a right-skewed distribution. (b) The sketch should show a right-skewed distribution. (c) The sketch should show a left-skewed distribution. (d) The sketch should show a right-skewed distribution. (e) The sketch should show a left-skewed distribution.

3.39. (a) Minimum = 0, Q1 = 2, median = 3, Q3 = 5, maximum = 14. (b) Same as part (a). (c) The observations with values 12 and 14 are outliers. (d) The standard deviation is 3.

3.41. (a)



(b) The distribution appears to be skewed to the right, because of the long distance between the upper quartile and the maximum.

3.48. (a) Response variable: happiness; explanatory variable: religious attendance. (b) For those who attend religious services nearly every week or more, 44.5% reported being very happy. For those who attend religious services never or less than once a year, 23.2% reported being very happy. (c) There appears to be an association between happiness and religious attendance since the percentages that reported being very happy differed greatly by attendance at religious services.

3.49. (a) United States: predicted fertility = 3.2 – 0.04(50) = 1.2; Yemen: predicted fertility = 3.2 – 0.04(0) = 3.2. (b) The negative value implies that the fertility rate decreases as Internet use increases.

3.50. (a) Points in a scatterplot for these data should have a negative association and be fairly tightly clustered in a straight-line pattern. (b) Contraceptive use is more strongly associated with fertility than is Internet use because –0.89 is larger in absolute value than –0.55.

3.59. The distribution of cost for New York and Boston are similar, and both cities have high and low outliers. The distributions for all three cities are roughly symmetric. The distribution for cost in London is higher than the distributions in both New York and Boston, with 75% of the costs in London being higher than all costs in Boston and almost all costs in New York.

3.64. The median is not impacted by gains made by the wealthiest Americans because the wealthiest Americans are at the high end of net worth, and the median is the value at the center of the data.

3.69. (a) The median is preferred over the mean when the data are skewed and/or there are outliers that will affect the mean. One example is income. (b) The mean is preferred over the mean when the distribution is approximately symmetric or when it is very highly discrete, such as the number of times you have been married.

3.70. (a) The standard deviation s is generally preferred over the range because it is calculated from all of the data and will not be impacted as much as the range when there are outliers. (b) The IQR is preferred to the standard deviation s when the distribution is very highly skewed or there are severe outliers, because the IQR is less sensitive to these features than s is.

3.72. (c)

3.73. (c)

3.74. (a)

3.78. (a) The mean is now 77, while the standard deviation stays at 20. (b) The mean is 50,000£, and the standard deviation is 15,000£.

Chapter 4

4.8. (a) 0.1587
(b) 0.1587
(c) 0.2514

4.9. a) z = 1 gives tail probability .1587, two-tail prob = 0.317, and prob within a standard deviation of the mean = 1 - .317 = .68.
b) z = 1.96 gives tail probability .025, two-tail prob = 0.05, and prob within 2 standard deviations of the mean = 1 - .05 = .95.
c) z = 3.0 gives tail probability .00135, two-tail prob = 0.0027, and prob within 3 standard deviations of the mean = 1 - .0027 = .997.
d) z = 0.67 gives tail probability .25, two-tail prob = 0.50, and prob within 0.67 standard deviations of the mean = 1 - .50 = .50.

4.10. (a) 2.33 (b) 1.96 (c) 1.64 (d) 1.28 (e) 0.67 (f) 0

4.11. (a) 0.67 (b) 1.64 (c) 1.96 (d) 2.33 (e) 2.58

4.12. (a) 1.28 (b) 1.64 (c) 2.06 (d) 2.33

4.17. (a) The 98th percentile is 2.05 standard deviations above the mean. (b) The IQ score for the 98th percentile is 100 + 2.05(16) = 132.8, or about 133.

4.19. (a) An MDI of 120 has z = (120 – 100)/16 = 1.25, and is 1.25 standard deviations above the mean. The proportion of children with an MDI of 120 or more is 0.1056. (b) The MDI score that is the 90th percentile is 1.28 standard deviations above the mean, so this score is 100 + 1.28(16) = 120.48, or 120. (c) The lower quartile is 0.67 standard deviations below the mean, which gives a lower quartile of 100 – 0.67(16) = 89.28, or 89. Similarly, the upper quartile is 0.67 standard deviations above the mean, which gives an upper quartile of 100 + 0.67(16) = 110.72, or 111. Since the MDI scores are approximately normal, the median will be equal to the mean of 100.

4.21. (a) 20 gallons per week has z = (20 – 16)/5 = 0.8, and is 0.8 standard deviations above the mean. The proportion of adults who use more than 20 gallons per week is 0.2119. (b) The 95th percentile is 1.645 standard deviations above normal. We need to solve for μ where 1.645 = (20 – μ)/5. The value of μ is 11.8. So, the mean would need to be about 11.8 gallons so that only 5% of adults use more than 20 gallons per week. (c) If the distribution of gasoline use is not actually normal, we should expect it to be

right-skewed, since there will be some adults with very high gasoline usages that will cause the distribution to have a long right tail.

4.23. An SAT score of 600 is (600 – 500)/100 = 1.0 standard deviations above the mean. An ACT score of 29 is (29 – 21)/4.7 = 1.70 standard deviations above the mean. Relatively speaking, an ACT score of 29 is higher than an SAT score of 600.

4.27. (a) The sampling distribution of the sample proportion of heads for flipping a balanced coin once is

p	0	1
Probability	0.50	0.50

(b) The sampling distribution of the sample proportion of heads for flipping a balanced coin twice is

p	0	0.5	1
Probability	0.25	0.50	0.25

(c) The sampling distribution of the sample proportion of heads for flipping a balanced coin three times is

p	0	1/3	2/3	1
Probability	0.125	0.375	0.375	0.125

(d) The sampling distribution of the sample proportion of heads for flipping a balanced coin four times is

p	0	0.25	0.50	0.75	1
Probability	0.0625	0.25	0.375	0.25	0.0625

(e) As the number of flips increases, the sampling distribution of the sample proportion of heads seems to be getting more normal, with the probabilities concentrating more closely around 0.50.

4.29. (a) $s_{\bar{y}} = \dfrac{s}{\sqrt{n}} = \dfrac{0.5}{\sqrt{2293}} = 0.0104$. (b) If actually 50% of the population voted for DeWine, it would be surprising to obtain 44% in this exit poll, since 44% is 6% lower than 50%, and the standard error for the sampling distribution is 1.04%; that is, the sample proportion of 0.44 is nearly 6 standard errors below 0.50. (c) Based on the information from the exit poll, I would be willing to predict that Sherrod Brown would win the Senatorial election.

4.33. (a) The probability that PDI is below 90 is

$$P(Y < 90) = P\left(Z < \frac{90-100}{15}\right) = P(Z < -0.67) = 0.2514.$$

(b) The probability that the sample mean PDI is below 90 is

$$P(\bar{Y} < 90) = P\left(Z < \frac{90-100}{15/\sqrt{25}}\right) = P(Z < -3.33) < 0.00135.$$

(c) An individual PDI of 90 is not surprising, since the probability is 0.2514 of that value or lower. However, a sample mean PDI of 90 would be surprising since this value would happen almost never. (d) The sketch of the sampling distribution should be less spread out and have a taller peak and thinner tails than the sketch of the population distribution.

4.36. (a) The population distribution is skewed to the right with mean 5.2 and standard deviation 3.0. (b) The sample data distribution based on the sample of 36 families and is skewed to the right with mean 4.6 and standard deviation 3.2. (c) The sampling distribution of $\bar{y}$ is approximately normal with mean 5.2 and standard error $3.0/\sqrt{36} = 0.5$. This distribution describes the theoretical distribution for the sample mean.

4.41 (b) Even though the population distribution is not normal (there are only two possible values), the sample proportions for the 1000 samples of size 100 each should have a histogram with an approximately bell shape.

4.42 (a) The population distribution is skewed, but the empirical distribution of sample means probably has a bell shape, reflecting the Central Limit Theorem.
(b) The Central Limit Theorem applies to relatively *large* random samples, but here $n = 2$ for each sample.

4.46. (a) The sample data distribution tends to resemble the population distribution more closely than the sampling distribution. A random sample of data from a population should be representative of the population, and its distribution should be similar to the population distribution. (b) The sample data distribution is the distribution of data that we actually observe. The sampling distribution of $\bar{y}$ is the probability distribution for the possible values of the sample statistic $\bar{y}$.

4.47. (a) A lower bound for the mean is

$$m = \sum yP(y) = 1(0.01) + 2(0.10) + 3(0.09) + 4(0.31) + 5(0.19) + 6(0.29) = 4.41.$$

(b) Since we know the category of ideal number of children that falls at the 50% point, we can find the median. The median is 4 children.

4.50. When $n = 100$, $s_{\bar{y}} = \frac{0.5}{\sqrt{n}} = \frac{0.5}{\sqrt{100}} = 0.05$. The interval 0.35 to 0.65 is the interval within which the sample proportion is almost certain to fall. When $n = 1000$, $s_{\bar{y}} = \frac{0.5}{\sqrt{1000}} = 0.016$. The interval 0.453 to 0.547 is the interval within which the sample proportion is almost certain to fall. When $n = 10{,}000$, $s_{\bar{y}} = \frac{0.5}{\sqrt{10{,}000}} = 0.005$. The interval 0.485 to 0.515 is the interval within which the sample proportion is almost certain to fall.

4.51. a, c, d

4.52. c

4.53. False. As the sample size increases, the standard error of the sampling distribution of $\bar{y}$ decreases, since $s_{\bar{y}} = \frac{s}{\sqrt{n}}$ decreases as n increases.

4.54. (a) Group A: $P(y < 400) = P\left(z < \frac{400 - 500}{100}\right) = P(z < -1) = 0.1587$. Almost 16% of students from Group A are not admitted to Lake Wobegon Junior College. Group B:

$P(y<400)=P\left(z<\dfrac{400-450}{100}\right)=P(z<-0.5)=0.3085$. Almost 31% of students from Group B are not admitted to Lake Wobegon Junior College. (b) Of the students who are not admitted, 0.3085/(0.3085 + 0.1587) = 0.3085/0.4672 = 0.6603, or about 66%, are from Group B. (c) If the new policy is implemented, the proportion of students from Group A that are not admitted would be 0.0228, while the proportion of students from Group B that are not admitted would be 0.0668. In this case, about 75% of the students who are not admitted would be from Group B. Relatively speaking, this policy would hurt students from Group B more than the current policy.

4.55. (a) $s=\sqrt{\sum(y-m)^2P(y)}=\sqrt{(0-0.5)^2(0.5)+(1-0.5)^2(0.5)}=\sqrt{0.25}=0.5$. (b) $m=\sum yP(y)=0(1-p)+1(p)=p$;

$s=\sqrt{(0-p)^2(1-p)+(1-p)^2(p)}=\sqrt{p^2-p^3+p-2p^2+p^3}=\sqrt{p-p^2}=\sqrt{p(1-p)}$. (c) The standard error for a sample proportion for a random sample of size n is $\dfrac{s}{\sqrt{n}}=\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}=\sqrt{\dfrac{p(1-p)}{n}}$.

4.57. (a) The finite population correction is $\sqrt{(30{,}000-300)/(30{,}000-1)}=\sqrt{0.99}=0.995$. (b) If n = N, the finite population correction is $\sqrt{(N-N)/(N-1)}=0$, so $s_{\bar{y}}=0$. (c) When n = 1, the finite population correction is $\sqrt{(N-1)/(N-1)}=1$, so $s_{\bar{y}}=\dfrac{s}{\sqrt{n}}=\dfrac{s}{\sqrt{1}}=s$. Thus, the sampling distribution of $\bar{y}$ and its standard error are identical to the population distribution and its standard deviation.

Chapter 5

5.4. The estimated standard error is $\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}=\sqrt{\dfrac{0.54(0.46)}{2003}}=0.011$.

5.7. (a) The estimated standard error in 2004 is $\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}=\sqrt{\dfrac{0.36(0.64)}{833}}=0.016$. (b) The margin of error is $z\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}=1.96(0.016)=0.03$, or 3%. (c) The 95% confidence interval is 36% - 3% = 33% to 36% + 3% = 39%. We are 95% confident that the population proportion of people agreeing that it is much better for everyone involved if the man is the achiever outside the home and the woman takes care of the home and family falls in the interval 33% to 39%.

5.8. $\hat{p} = 366/598 = 0.612$ The 99% confidence interval is

$$\hat{p} \pm z\sqrt{\frac{\hat{p}\,(1-\hat{p}\,)}{n}} = 0.612 \pm 2.576\sqrt{\frac{0.612\,(0.388)}{598}} = 0.56 \text{ to } 0.66.$$

5.12. (a) "Sample prop" = 1885/2815 = 0.6696. (b) Since we are 95% confident that the interval 65.2% to 68.7% contains the population proportion of American adults who are in favor of the death penalty and the entire interval exceeds 50%, it is reasonable to conclude that more than half of all American adults are in favor of the death penalty. (c) A 95% confidence interval for the proportion of American adults who opposed the death penalty is 31.3% to 34.8%.

5.13. (a) The proportion that said *legal* is 0.364; the proportion that said *not legal* is 0.636. (b) The 95% confidence interval is $0.364 \pm 1.96\sqrt{\frac{0.364(0.636)}{802}} = 0.331$ to 0.397. We are 95% confident that the interval 0.331 to 0.397 contains the population proportion that thinks marijuana should be made legal. Since this interval is entirely below 50%, we can conclude that a minority of Americans felt this way. (c) The proportion that said marijuana should be legal dropped until 1990 and has increased each year since.

5.18. If the sample size had been one-fourth as large, the confidence interval would be twice as wide and would be 0.23 to 0.31.

5.21. (a) The standard error is $\frac{s}{\sqrt{n}} = \frac{52.554}{\sqrt{1007}} = 1.656$. (b) We are 95% confident that the interval 21.5 to 28.0 contains the population mean number of female partners males have had sex with since their eighteenth birthday. (c) The mean is quite high compared to the median and the mode, which means that there were a few male respondents with a very large number of female sex partners. In addition, the standard deviation is more than twice the mean, confirming the right skew of the distribution of the number of female sex partners. A confidence interval based on the mean does not seem to be the best idea.

5.22. (a) The point estimate is 3.02 children. (b) The standard error is $\frac{s}{\sqrt{n}} = \frac{1.81}{\sqrt{497}} = 0.081$. (c) We are 95% confident that the interval 2.9 to 3.2 contains the population mean ideal number of children for a family to have. (d) Since the confidence interval is entirely above 2.0, it does not seem plausible that the population mean equals 2.0 children.

5.24. (a)

$$\bar{y} = \frac{11+11+6+9+14-3+0+7+22-5-4+13+13+9+4+6+11}{17} = \frac{124}{17} = 7.29;$$

$s = \sqrt{\frac{\sum (y-\bar{y})^2}{n-1}} = \sqrt{51.596} = 7.18$. (b) The standard error is $\frac{s}{\sqrt{n}} = \frac{7.18}{\sqrt{17}} = 1.74$. (c) The *t*-score that

is in the df = 16 row and $t_{0.025}$ column is 2.120. (d) The 95% confidence interval is $\bar{y} \pm t\frac{s}{\sqrt{n}} = 7.29 \pm 2.120\frac{7.18}{\sqrt{17}} = 3.6$ to 11.0. We are 95% confident that the interval 3.6 to 11.0 pounds contains the population mean change in weight for this therapy.

5.25. A confidence is not about any one subject or about 95% of the subjects, it is an interval estimate for our population parameter. The correct interpretation is that we are 95% confident that the interval 2.60 to 2.93 hours contains the population mean number of hours of TV watched on the average day.

5.28. (a) The 95% confidence interval is $1.81 \pm 1.96\frac{1.98}{\sqrt{816}} = 1.67$ to 1.95. We are 95% confident that the interval 1.67 to 1.95 contains the population mean number of days in the past 7 days that women have felt sad. (b) Since the standard deviation is larger than the mean, the variable is most likely skewed to the right. Since t procedures are robust against violations of normality and our sample size is large, our findings in part (a) are probably okay, unless there are extreme outliers.

5.32. (a) $\bar{y}$ = 1.5 days. (b) The 95% confidence interval is $1.5 \pm 1.96\frac{2.21}{\sqrt{1450}}$ = 1.4 to 1.6. We are 95% confident that the interval 1.4 to 1.6 contains the population mean number of days in the past 7 days that people have felt lonely.

5.34. (a) The confidence interval is 4.3 to 6.3. We are 95% confident that the interval 4.3 to 6.3 days contains the population mean length of stay for all inpatients in that hospital. (b) If the administrator wants the confidence interval to be half as wide, she needs to take a random sample of 400 records.

5.35. $n = p(1-p)\left(\frac{z}{M}\right)^2 = 0.30(0.70)\left(\frac{1.645}{0.06}\right)^2 = 156.89$ The necessary sample size is 157.

5.39. $n = 0.83(0.17)\left(\frac{1.96}{0.03}\right)^2 = 602.3$ The sample size was about 602.

5.41. We estimate the standard deviation to be (18 – 0)/6 = 3. The sample size calculation is $n = s^2\left(\frac{z}{M}\right)^2 = 3^2\left(\frac{1.96}{1}\right)^2 = 34.57$, so a sample of size 35 is needed.

5.44. (a) $\hat{p} = 0/5 = 0$, $se = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0(1)}{5}} = 0$. (b) Since the number in each category (0 like tofu and 5 do not like tofu) is less than 15, we cannot use the large-sample formula for a confidence interval. An appropriate confidence interval uses $\hat{p} = 2/9 = 0.222$ and the 95% confidence interval is

$0.222 \pm 1.96\sqrt{\frac{0.222(0.778)}{9}} = 0.0$ to 0.49. We are 95% confident that the interval 0 to 0.49 contains the population proportion of students who like tofu.

5.47. (a) You would expect about 95% to contain the parameter (but not necessarily exactly 95% because of random sampling fluctuations).

5.48. (a) For the 95% confidence interval, only about 5 of the confidence intervals should have failed to contain the parameter value of 0.90. You probably observed more than that, because the method does not work well with such a small sample size (n = 10). Recall that this formula is designed for cases in which at least 15 observations occur in each of the two categories.
(c) When the population proportion is 0.90, you probably observed that the sampling distribution is skewed to the left, as the sample proportion can not be much larger than 0.90 (since its upper bound is 1.0) but it could be much smaller than 0.90 when the sample size is small.

5.62. The discussion at the bottom of p. 126 of the textbook explains how the necessary sample size for estimating a mean to within a particular margin of error M depends on M, the degree of confidence (which affects the z or t score) and the population variance. The sample size must increase as the population is more variable. For entry-level employees at McDonalds, the income would not vary much, whereas for medical doctors it would vary a lot, so a larger sample size would be necessary to estimate population mean income for doctors than for McDonalds' employees.

5.66. (a)

5.67. (a)

5.68. (b)

5.69. (b) and (e) are correct

5.70. (a) A confidence interval for the mean is about the population mean, not the sample mean. (b) A confidence interval for the mean is about the population mean, not individuals. (c) We can actually be 100% confident that the sample mean is in the interval we construct (it is the midpoint of the interval). (d) This statement implies that the population mean changes.

5.71. We are 95% confident that the interval 21.5 to 23.0 years contains the mean age at first marriage of women in a certain country.

5.73. Since $\bar{y} = \frac{\sum y}{n}$, $\sum y = n\bar{y}$. If we know $(n-1)$ of the observations, we can add these up and subtract from $n\bar{y}$ to find the value of the remaining observation.

5.77. Given $\hat{p} = 0$ and n = 20, $|0 - p| = 1.96\sqrt{\frac{p(1-p)}{20}}$. Squaring both sides gives us $p^2 = \frac{1.96^2}{20}p - \frac{1.96^2}{20}p^2 = 0.19208p - 0.19208p^2$. This equation simplifies to $1.19208p^2 - 0.19208p = p(1.19208p - 0.19208) = 0$. The roots that solve this equation are $p = 0$ and $p = 0.161$.

Chapter 6

6.1. (a) null hypothesis (b) alternative hypothesis (c) alternative hypothesis (d) $H_0 : p = 0.50$; $H_a : p < 0.24$; $H_a : m > 100$.

6.2. (a) Let μ = the population mean ideal number of children. $H_0 : m = 2$ and $H_a : m \neq 2$. (b) The test statistic is t = 20.80. This test statistic was obtained with the following: $t = \frac{\overline{y} - m_0}{se} = \frac{2.49 - 2}{0.850/\sqrt{1302}} = 20.80$. (c) The *P*-value is the probability, assuming the null hypothesis is true, that the test statistic equals the observed value or a value even more extreme in the direction predicted by the alternative hypothesis. In this case, the *P*-value = 0.0000 (rounded to 4 decimal places), which means that if the true mean were 2, we would see results as extreme as or more extreme than we did almost never.

6.3. (a) $P = 2P(t > 1.04) \approx 2P(z > 1.04) = 2(0.1492) = 0.30$. If the true mean were 0, we would see results at least as extreme as we did about 30% of the time. (b) Now, $P = 2P(t < -2.50) \approx 2P(z < -2.50) = 2(0.0062) = 0.012$. Since the *P*-value is smaller than the *P*-value calculated in part (a), we have stronger evidence against the null hypothesis. (c) (i) 0.15 (ii) 0.85.

6.9. (a) The null hypothesis is $H_0 : m = 0$, and the alternative hypothesis is $H_a : m \neq 0$. (b) $t = \frac{\overline{y} - m_0}{se} = \frac{-0.052 - 0}{0.0397} = -1.31$. The *P*-value is P = 0.19. Since P > 0.05, we fail to reject the null hypothesis. There is not enough evidence to conclude that the mean score differs from the neutral value of zero. (c) We cannot accept $H_0 : m = 0$, since failing to reject the null hypothesis does not prove the null hypothesis to be true. We just did not have enough evidence to reject the null hypothesis. (d) A 95% confidence interval is $\overline{y} \pm t\frac{s}{\sqrt{n}} = -0.052 \pm 1.96(0.0397) = -0.13$ to 0.03. Note that zero falls within our confidence interval. When the null hypothesis value falls within the confidence interval, we fail to

reject the null hypothesis; when the null hypothesis value falls outside of the confidence interval, we reject the null hypothesis (with Type I error probability = the error probability for the CI).

6.11. Results of 99% confidence intervals for means are consistent with results of two-sided tests at $a = 1 - 0.99 = 0.01$ level.

6.15. (a) $H_0: p = 0.50$ and $H_a: p \neq 0.50$. (b) $z = -3.91$ The sample proportion falls 3.91 standard errors below the null hypothesis value of the proportion. (c) $P = 0.000$ (rounded). If the null hypothesis were true, then there is almost no chance that we would see results as extreme as or more extreme than our sample results. We would conclude that the proportion of Americans who would be willing to pay higher taxes in order to protect the environment differs from 0.50. (d) The confidence interval is entirely below 0.50, which leads us to believe that a minority of Americans would be willing to pay higher taxes in order to protect the environment. It shows just how far from 0.50 the parameter could plausibly be.

6.17. $\hat{p} = \dfrac{40}{116} = 0.345$ The test statistic is $z = \dfrac{0.345 - 1/3}{\sqrt{\dfrac{1/3(2/3)}{116}}} = 0.26$. The P-value is $P = P(z > 0.26) = 0.40$. We fail to reject the null hypothesis at $a = 0.05$ level. There is not enough evidence to conclude that astrologers are correct with their predictions more than 1/3 of the time.

6.18. (a) A Type I error would be concluding that astrologers are correct with their predictions more than 1/3 of the time when they really are not. (b) A Type II error would be failing to conclude that the astrologers are correct with their predictions more than 1/3 of the time when they really are.

6.22. (a) (i) A Type I error would be concluding that the population mean weight change was positive when it was actually 0. (ii) A Type II error would be failing to conclude that the population mean weight change was positive when it really was. (b) This would be a Type I error. (c) If $a = 0.01$, we would fail to reject the null hypothesis. If this decision were in error, it would be a Type II error.

6.23. (a) Jones: $t = \dfrac{\bar{y} - m_0}{se} = \dfrac{519.5 - 500}{10.0} = 1.95$, $P = 2P(t > 1.95) \approx 2P(z > 1.95) = 0.051$; Smith: $t = \dfrac{519.7 - 500}{10.0} = 1.97$, $P = 2P(t > 1.97) \approx 2P(z > 1.97) = 0.049$. (b) Jones's result is not statistically significant (since P-value > 0.050), but Smith's result is. (c) These two studies give such similar results that they should not yield different conclusions. Reporting the actual P-value shows that each study has moderate evidence against H_0 and shows that the results are very similar in practical terms.

6.24. (a) $se_0 = \sqrt{\dfrac{0.50(0.50)}{400}} = 0.025$;

Jones: $z = \dfrac{\hat{p} - p_0}{se_0} = \dfrac{0.55 - 0.50}{0.025} = 2.0$, $P = 2P(z > 2.0) = 0.046$;

Smith: $z = \dfrac{\hat{p} - p_0}{se_0} = \dfrac{0.5475 - 0.50}{0.025} = 1.90$, $P = 2P(z > 1.90) = 0.057$. (b) Jones's result is statistically significant (since P-value < 0.05), but Smith's result is not. (c) These two studies give such similar results that they should not yield different conclusions. Reporting the actual *P*-value shows that each study has moderate evidence against H_0 and shows that the results are very similar in practical terms. (d) The two confidence intervals almost entirely overlap, showing that the results of Jones and Smith are very similar.

6.25. $t = \dfrac{\bar{y} - m_0}{se} = \dfrac{497 - 500}{100/\sqrt{10{,}000}} = -3.0$, $P = 2P(t > 3.0) \approx 2P(z > 3.0) = 0.003$ Since the *P*-value is so small, we reject the null hypothesis at the usual alpha-levels and find the results to be highly statistically significant. However, the difference between the sample mean of 497 and the hypothesized mean of 500 is not important in a practical sense.

6.29. (a) H_0 is rejected when $P = 0.05$, which happens when $z = 1.64$, that is when the sample proportion falls at least 1.64 standard errors above the null hypothesis value of 0.50. This region is $\hat{p} \geq 0.50 + 1.64\sqrt{\dfrac{0.50(0.50)}{25}} = 0.664$. (b) z = (0.664 – 0.60)/0.10 = 0.64; We fail to reject if $\hat{p} < 0.664$, which happens with probability 0.74.

6.30. (a) H_0 is rejected when $P = 0.05$, which happens when $t = 1.699$ (for df=29), that is when the sample mean falls at least 1.699 standard errors above the null hypothesis value of 0. (b) This region is $\bar{y} \geq 0 + 1.699\dfrac{18}{\sqrt{30}} = 5.6$, so we fail to reject if $\bar{y} < 5.6$ (c) $t = \dfrac{5.6 - 10}{18/\sqrt{30}} = -1.33$; (d) *P*(Type II error) = $P(t < -1.33) = 0.10$.

6.33. (a) Let p = probability she guesses correctly on a particular flip. We test $H_0: p = 0.50$ and $H_a: p > 0.50$. The null says she actually does no better than random guessing, whereas the alternative says she does better than that. (b) Find the right-tail probability for the binomial distribution with $n = 5$ and $p = 0.50$. That is, the *P*-value is $P(4) + P(5) = 0.156 + 0.031 = 0.187$. This outcome is not unusual if she does not actually possess ESP. It can be explained by chance. We cannot reject H_0, and thus her claim is not convincing.

6.34. (a) We have a fixed number of observations (1336), each of which falls into one of two categories (voted for Clinton, did not vote for Clinton). The probability of falling in each category is the same for every observation. The outcomes of successive observations are independent (no person's vote impacted any other person's vote).

(b) $m = np = 1336(0.50) = 668$; $s = \sqrt{np(1-p)} = \sqrt{1336(0.50)(0.50)} = 18.3$. (c) We would almost certainly expect x to fall in the interval 613 to 723 (i.e., within 3 standard deviations). (d) Since x = 895 exceeds the upper bound of the interval in part (c), it appears that the population proportion of voters who voted for Hillary Clinton in the 2006 Senatorial election in New York is greater than 0.50.

6.39. The report should include $\bar{y}$ = 4, s = 2, $H_0: m = 0$, $H_a: m > 0$, t = (4 – 0)/1 = 4.0, P = 0.014.

6.45. (a) A Type I error occurs when one convicts the defendant, when he or she is actually innocent; a Type II error occurs when one acquits the defendant even though he or she is actually guilty. Most people would consider convicting an innocent defendant to be more serious. (b) To decrease the chance of a Type I error, one gives the defendant additional rights and makes it more difficult to introduce evidence that may be inadmissible in some way; this makes it more likely that the defendant will not be convicted, hence the relatively more guilty parties will be incorrectly acquitted. (c) If this strategy is used, a = 0.000000001, which means that the jury would almost never find the defendant guilty. It would be almost impossible to convict a defendant, and many guilty defendants would be acquitted.

6.50. If H_0 is true every time, the expected number of times one would get P = 0.05 just by chance is 60(0.05) = 3, which is the mean of the binomial distribution with n = 60 and p = 0.05; thus, the 3 cases could all simply be Type I errors.

6.51. (a) For each test, the probability equals 0.05 of falsely rejecting H_0 and committing a Type I error. This policy encourages the publishing of Type I errors, if actually many researchers are conducting studies in which the null hypotheses are correct. (b) Of all the studies conducted, the one with the most extreme or unusual results is the one that gets substantial attention. That result may be an unusual sample, with the sample far from the actual population mean. Further studies in later research would reveal that the true mean is not so extreme.

6.52. (a)

6.53. (b) and (e) are both correct

6.54. (b)

6.55. (a) and (c) are both correct

6.56. (b) and (d) are both correct

6.57. (a) False; P(Type I error) is a single value, the fixed a value (such as 0.05), whereas P(Type II error) decreases as the true parameter value falls farther from the H_0 value in the direction of values in H_a. (b) True; If we reject using a = 0.01, then P = 0.01. Thus, P = 0.05 also, so we also reject using a = 0.05. (c) False; The P-value is the probability of obtaining a sample statistic as extreme as or more extreme than we did, given H_0 is true. (d) True; Since P = 0.063 is greater than 0.05, we fail to reject H_0 at that level

and do not conclude that μ differs from 0. Similarly, the 95% confidence interval would contain 0, indicating that 0 is plausible for μ. The confidence interval shows precisely which values *are* plausible.

6.59. The value in H_0 is only one of many plausible values for the parameter. A confidence interval displays a range of possible values for the parameter. The terminology "accept H_0" makes it seem as if the null value is the only plausible one.

6.61. (a) H_0 either is, or is not, correct. It is not a variable, so one cannot phrase probability statements about it. (b) If H_0 is true, the probability that $\bar{y}$ = 120 or that $\bar{y}$ = 80 (i.e., that $\bar{y}$ is at least 20 from m_0 = 100, so that |*z*| is at least as large as observed) is 0.057. (c) This is true if the statement "$m = 100$" is substituted for "$m \neq 100$." (d) The probability of a Type I error equals a (which is not specified here), not the *P*-value. The *P*-value is compared to a in determining whether one can reject H_0. (e) It would be better to say "We do not reject H_0 at the a = 0.05 level." (f) No, we need P = 0.05 to be able to reject H_0.

Chapter 7

7.1. These are independent samples, because the subjects in the two samples are different, with no matching between one sample with the other sample.

7.9. (a) We can be 95% confident that the interval 0.18 to 0.26 contains the difference between the population proportion of teens who listen to lots of music with degrading sexual messages and have intercourse and the population proportion of teens who listen to little or no music with degrading sexual messages and have intercourse. (b) If the two population proportions were equal, it would be very unlikely to observe a difference as large as we did. It appears that the proportion of teens who listen to lots of music with degrading sexual messages and have intercourse is greater than the proportion of teens who listen to little or no degrading music and have intercourse.

7.11. (a)

$$se = \sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2} = \sqrt{0.399(0.601)/12{,}708 + 0.482(0.518)/8783} = 0.0069$$ (b)

$(\hat{p}_2 - \hat{p}_1) \pm z(se) = (0.482 - 0.399) \pm 1.96(0.0069) = 0.083 \pm 0.014 = 0.069$ to 0.097. We can be 95% confident that interval 0.069 to 0.097 contains the difference in the population proportion of college students who drank to get drunk in 2001 and the population proportion of college students who drank to get drunk in 1993.

7.16. (a) Let p_1 = the population proportion of female senior high school students who have ever used marijuana and p_2 = the population proportion of male senior high school students who have ever used marijuana. We could investigate whether there is a difference by testing $H_0: p_2 - p_1 = 0$, against

$H_a: p_2 - p_1 \neq 0$. (b) We are 95% confident that the population proportion of female senior high school students who have used marijuana is between 0.089 and 0.008 lower than the population proportion of male senior high school students who have used marijuana. (c) If the difference between the true population proportions of female and male senior high school students who have ever used marijuana were actually 0, we would see this large a difference or even larger with probability 0.02 (i.e., quite unlikely). It appears that the population proportion of male senior high school students who have ever used marijuana is greater than the population proportion of female senior high school students who have ever used marijuana.

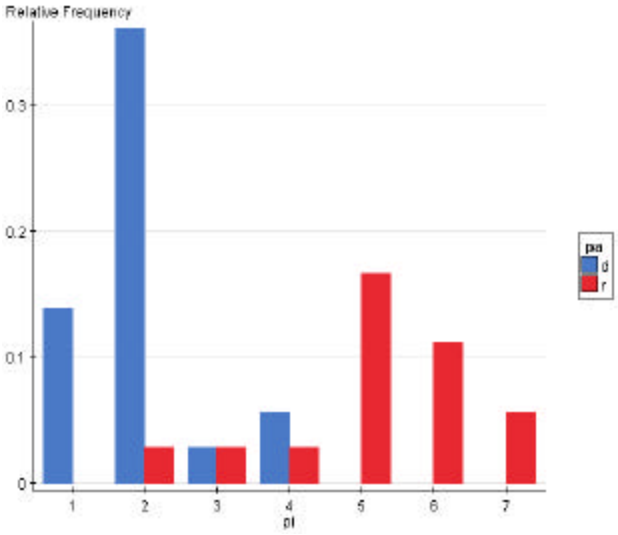
7.21. (a) The estimated difference between the HONC means for smokers and ex-smokers is 4.9. (b) We can be 95% confident that the HONC population mean for smokers is between 4.1 and 5.7 higher than the HONC population mean for ex-smokers. (c) THE HONC sample data distribution for ex-smokers appears to be right-skewed, since the standard deviation is greater than the mean. If there are no severe outliers, the inference might not be affected too much, since two-sided *t* procedures are robust against violations of normality and the sample size is large (so the sampling distribution is bell-shaped by the CLT).

7.23. We can be 95% confident that the population mean number of days in the past 7 days that women have felt sad is between 0.2 and 0.6 greater than the population mean number of days in the past 7 days that men have felt sad. If the population means were identical, the probability would be 0.000 (rounded to 3 decimal places) of observing a sample difference this large or even larger (in either direction). Both the confidence interval and the hypothesis test lead us to believe that there is a difference in the population mean number of days in the past 7 days that women have felt sad and the population mean number of days in the past 7 days that men have felt sad. These results are statistically significant, but the confidence interval indicates that the results may not be practically significant because the plausible differences are close to 0.

7.26. $df = 29 - 1 = 28$. The one-sided *P*-value is between 0.005 and 0.01 (software indicates $P = 0.006$), and the two-sided *P*-value is between 0.01 and 0.02 (software indicates $P = 0.012$). There is strong evidence that those who are not children of alcoholics have a higher population mean well-being than those who are children of alcoholics.

7.32. (a) $s = \sqrt{\dfrac{(12)2.1^2 + (16)3.2^2}{28}} = 2.78$, $se = 2.78\sqrt{\dfrac{1}{13} + \dfrac{1}{17}} = 1.025$. The 95% confidence interval is $2.8 \pm 2.048(1.025) = 0.7$ to 4.9. We can be 95% confident that the population mean family cohesion for nonabused students is between 0.7 and 4.9 higher than the population mean family cohesion for sexually abused students. (b) With $P = 0.01$ (rounded) for each two-sided analysis, there is strong evidence of a difference between the population mean family cohesion for nonabused students and the population mean family cohesion for sexually abused students. If the null hypothesis of identical population means were true, it would be very unusual to get a sample mean difference of 2.8 or even larger.

7.41. (a) The following is a bar chart:


Relative Frequency
0.3
0.2
0.1
0
1
2
3
4
5
6
7
pl
pa
g
r

Summary statistics for political ideology:

pa	n	Mean	Std. Dev.	Std. Err.	Median	Range	Min	Max	Q1	Q3
Dem	21	2	0.84	0.18	2	3	1	4	2	2
Rep	15	5.13	1.36	0.35	5	5	2	7	5	6

(b) Software gives us the following for a hypothesis test:

Hypothesis test results:

μ_1 : mean of Dem
μ_2 : mean of Rep
$H_0 : \mu_1 - \mu_2 = 0$
$H_A : \mu_1 - \mu_2 \neq 0$
(without pooled variances)

Difference	Sample Mean	Std. Err.	T-Stat	P-value
$\mu_1 - \mu_2$	-3.13	0.39	-7.94	<0.0001

Software gives us the following for a confidence interval:

95% confidence interval results:

μ_1 : mean of Dem
μ_2 : mean of Rep
(without pooled variances)

Difference	Sample Mean	Std. Err.	L. Limit	U. Limit
$\mu_1 - \mu_2$	-3.13	0.39	-3.95	-2.31

The hypothesis test tells us that there is strong evidence that the population mean political ideology differs for students identifying with the Democratic party and with the Republican party. We would reject the null hypothesis of equal population means at the usual significance levels such as 0.05. The confidence interval tells us that we are 95% confident that the population mean political ideology for students identifying with the Democratic party is between 2.3 and 4.0 below (more liberal) than the population mean political ideology for students identifying with the Republican party.

7.49. (a) (i) The interpretation suggests that the 95% confidence interval is (46-42) ± 3.4, or (0.6, 7.4). (ii) The interpretation suggests that the P-value is 0.02 for a two-sided test.

7.50. One possible approach compares mean numbers of dates for different levels of attractiveness within each gender. For example, a 95% confidence interval for More – Less for Men is: $(9.7-9.9) \pm 1.96\sqrt{\frac{10^2}{35}+\frac{12.6^2}{36}}$ = –5.5 to 5.1, for which it is plausible that there is no difference

between the population mean number of dates for the two levels of attractiveness. By contrast, a 95% confidence interval for More – Less for Women is: $(17.8-10.4)\pm 1.96\sqrt{\frac{14.2^2}{33}+\frac{16.6^2}{27}}$ = –0.52 to 15.3. Here, 0 is barely in the interval, and it is plausible that the population mean number of dates is very much larger for the more attractive group. One could alternatively compare population means between genders within each level of attractiveness.

7.55. The *se* value for estimating a difference between two means is larger than the *se* value for estimating a single mean. The analysis in Chapter 6 was for a single mean, for which the *se* value is smaller. So, a difference of 3.0 pounds in the Chapter 6 analysis is not necessarily less significant that a difference of 3.46 pounds in the analysis in Example 7.7 because the latter difference compared two sample means and had a larger *se*.

7.59. (a) False: The confidence interval addresses the difference between the population proportions for Hispanic and white youths, not the population proportion of white youths alone. (b) False: The samples are independent, because a youth is either a female or a male (i.e., the two groups had separate individuals).

7.60. False: We can conclude that μ_2 is greater than μ_1 but the confidence interval does not indicate plausible values for the separate population means.

7.61. True, from the formula on p. 185 you could figure out the standard error of the difference between the sample means.

7.62. (b)

7.63. (a), (c), and (d)

Chapter 8

8.1. (a)

	Allow unrestricted abortion	
	Yes	No
Male	0.40	0.60
Female	0.40	0.60

(b) Since the proportions do not depend on gender, statistical independence does seem plausible.

8.3. (a) Dependent. (b) One possible answer is

	Legalize Marijuana	
	Yes	No
Male	0.40	0.60

Female	0.40	0.60

8.4. The response variable is whether a German has a favorable opinion of the U.S. The explanatory variable is year. The conditional distribution for 2000 is (0.78, 0.22) and for 2006 is (0.37, 0.63).

	Favorable opinion of U.S.	
	Yes	No
2000	0.78	0.22
2006	0.37	0.63

8.5. (a) The conditional distribution is (86%, 14%) for the categories (Positive, Negative) for those who have breast cancer, and (12%, 88%) for those who do not have breast cancer. The mammogram does seem to be a good diagnostic tool. (b) (6.8%, 93.2%) for categories (Yes, No) in the "positive" column of the table. Of those with a positive test result, only 6.8% truly have breast cancer, so when the test says that a woman may have breast cancer, there is actually a small chance that she does. When a disease is not common, the number of positive diagnoses for people not having the disease (even though a small percentage of diagnoses for people not having the disease) can be much larger than the number of positive diagnoses for people having the disease.

8.9. (a) H_0: Sex and opinion about cuts in the standard or living to help the environment are independent. H_a: Sex and opinion about cuts in the standard or living to help the environment are not independent. (b) $df = 4$. (c) Based on the chi-squared table with $df = 4$, the P-value is between 0.05 and 0.10. (i) Do not reject H_0; There is not enough evidence to conclude that sex and opinion about cuts in the standard or living to help the environment are dependent. (ii) Reject H_0; There is enough evidence to conclude that sex and opinion about cuts in the standard or living to help the environment are dependent.

8.10. (a) (74%, 26%) for those who have used alcohol, (14%, 86%) for those who have not; there is a strong association, whereby those who have used alcohol are much more likely to have smoked cigarettes than are those who have not used alcohol. (b) H_0: Cigarette use and alcohol use are independent. H_a: Cigarette use and alcohol use are not independent. $c^2 = 451.4$, $df = 1$, $P < 0.001$. There is extremely strong evidence of an association between cigarette use and alcohol use.

8.11. (a) H_0: Happiness and belief in life after death are independent. H_a: Happiness and belief in life after death are dependent. (b) It is (32.6%, 54.4%, 13.0%) for the categories (Very happy, Pretty happy, Not too happy) for those who believe in life after death, and (25.8%, 60.4%, 13.8%) for those who do not believe in life after death. (c) $c^2 = 7.98$, $df = 2$, $P = 0.02$. There is strong evidence to suggest an association between happiness and belief in life after death. (d) The standardized residuals are (2.81, –2.32, –0.45) for the categories (Very happy, Pretty happy, Not too happy) for those who believe in life after death, and (–2.81, 2.32, 0.45) for those who do not believe in life after death. For those who believe in life after death, there are more people in the "very happy" category than we would expect if the two variables were independent. For those who do not believe in life after death, there are fewer people in the "very happy" category than we would expect if the two variables were independent.

8.14. (a) The expected frequency is 129.1, which is (373)(890)/2571. (b) H_0: Party ID and race are independent. H_a: Party ID and race are dependent. c^2 = 243.73, df = 2, P = 0.0000. There is strong evidence of an association between party ID and race. (c) Blacks tend to be Democrats more often than we would expect under independence, while whites tend to be Republicans more often than we would expect under independence.

8.16. (a) With df = 8, P < 0.001; there is strong evidence of an association between happiness and marital status. (b) There are (i) more married people and (ii) fewer widowed, divorced, separated, and never married people in the *very happy* category than if the variables were independent. (c) 42.5% of married people are in the *very happy* category, while 19.2% of divorced people are in the *very happy* category. The difference in proportions is 0.425 – 0.192 = 0.233. The sample proportion of married people in the *very happy* category is considerably higher than the proportion of divorced people in the *very happy* category.

8.17. Political party affiliation and opinion about Bush's performance appears to be very strongly associated, with the sample difference of proportions approving of his performance being 0.82 – 0.09 = 0.73.

8.18. Race is more strongly associated with death penalty opinion than is gender, since the difference in proportions for whites and blacks (which is 0.31) who support the death penalty is much greater than the difference in proportions of males and females (which is 0.12).

8.20. (a) (2409/37,792) – (3865/30,902) = 0.064 – 0.125 = –0.061; the proportion who were injured or killed is 0.06 lower for those who wore seat belts. (b) (2409)(27,037)/[(3865)(35,383)] = 0.49; users of seat belts have odds of being injured or killed that are estimated to be 0.49 times the odds for nonusers of seat belts.

8.22. (a) (i) (1/108)/(1/1562) = 14.5; the odds that a male resident is incarcerated is 14.5 times the odds that a female resident is incarcerated. (ii) (1694/98,306)/(252/99,748) = 6.8; the odds that a black resident is incarcerated is 6.8 times the odds that a white resident is incarcerated. (b) Gender has a stronger association with whether incarcerated, since the odds ratio is greater for gender than for race.

8.29. (a) Computer output for relating opinion about abortion and political affiliation follows:

PolAffiliation * Abortion Crosstabulation

Count

		Abortion		Total
		n	y	n
PolAffiliation	d	1	20	21
	i	3	21	24
	r	9	6	15
Total		13	47	60

The sample conditional distributions computed within rows would show that the proportion who supported legalized abortion in this sample is much higher for Democrats and Independents than for Republicans.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	17.711	2	0.000

The P-value of 0.000 for testing that opinion about abortion is independent of political party affiliation gives very strong evidence against the null hypothesis and in favor of an association. A couple of the expected frequencies were small, so the Pearson test is on somewhat shaky grounds. Standardized residuals should also be reported, which back up the descriptive analysis given above.

Chapter 9

9.1. (a) y = college GPA. (b) y = number of children. (c) y = annual income. (d) y = assessed value of home.

9.3. (a) The y-intercept is 61.4, and the slope is 2.4. For each additional centimeter in length of the femur, predicted height increases by 2.4 centimeters. (b) $\hat{y} = 61.4 + 2.4(50) = 181.4$ cm.

9.7. (a) (i) $\hat{y} = 1.26 + 0.346(0.8) = 1.54$; (ii) $\hat{y} = 1.26 + 0.346(34.3) = 13.1$. (b) $y - \hat{y} = 19.7 - 13.1 = 6.6$; the U.S. is producing 6.6 metric tons per capita more in CO_2 emissions than predicted by the regression line. (c) $\hat{y} = 1.26 + 0.346(28.1) = 11.0$; $y - \hat{y} = 5.7 - 11.0 = -5.3$; Switzerland is producing 5.3 metric tons per capita less in CO_2 emissions than predicted by the regression line.

9.10. (a) The point for Palm Beach county appears to be an outlier, with more votes for Buchanan than we would expect based on the number of votes for Perot. (b) (Note: the sign of the prediction equation in the textbook should be positive, not negative.) $\hat{y} = 45.7 + 0.02414(30{,}739) = 788$; $y - \hat{y} = 3407 - 788 = 2619$; the number of Buchanan votes in Palm Beach county is 2619 higher than we would predict with the regression equation. (c) The two rightmost points follow the pattern of the regression line for the rest of the data, whereas the top point is quite far from the trend that the rest of the data follow.

9.11. (a) (i) (20, 85); (ii) (34, 45). (b) $\hat{y} = -0.13 + 2.62(34.3) = 89.7$; $y - \hat{y} = 45.1 - 89.7 = -44.6$; the percent of people using cell phones in the U.S. is 44.6% lower than would be predicted by the regression line. (c) The correlation is positive; higher values of percent of people using cell phones are associated with higher values of GDP, and lower values of percent of people using cell phones are associated with lower values of GDP.

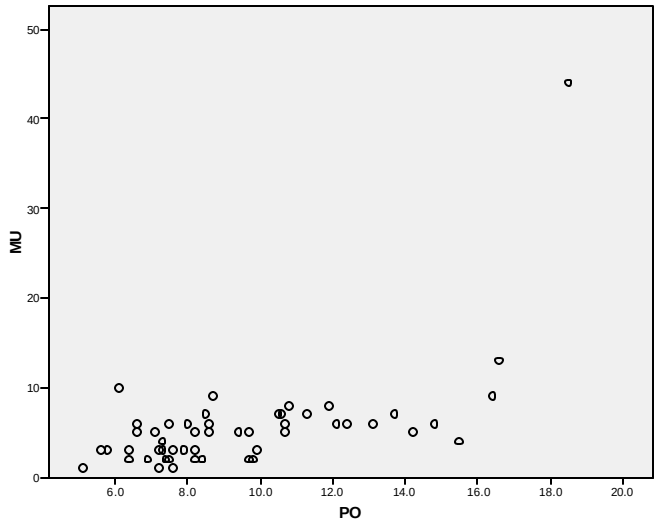
9.12. (a) (i) higher percents of people using the Internet are associated with higher per capita GDP; (ii) higher percents of people using the Internet are associated with lower fertility rates. (b) (i) Per capita GDP has the strongest linear association with Internet use, because it has the largest correlation (in absolute value). (ii) Fertility rate has the weakest linear association with Internet use.

9.13. $b = r\left(\frac{s_y}{s_x}\right) = 0.60\left(\frac{120}{80}\right) = 0.90$; $a = \bar{y} - b\bar{x}$ = 500 – 0.90(480) = 68; $\hat{y} = a + bx$ = 68 + 0.90x.

9.18. (a) (i) $\hat{y}$ = 30 + 0.60(100) = 90; (ii) $\hat{y}$ = 30 + 0.60(50) = 60. (b) $r = b\left(s_x/s_y\right)$ = 0.60(10/10) = 0.60. (c) b = 1, so $r = b\left(s_x/s_y\right)$ = 1(10/10) = 1.0. (d) $r = b\left(s_x/s_y\right)$ = 0(10/10) = 0.

9.20. (a) $\hat{y}$ = –0.105 + 0.546x; for each \$1000 increase in GDP per person, the predicted annual oil consumption per person increases by 0.55 barrels. (b) r = 0.847; there is quite a strong positive association between GDP per person and annual oil consumption per person. (c) For Canada, $\hat{y}$ = –0.105 + 0.546(34) = 18.5; $y - \hat{y}$ = 26 – 18.5 = 7.5; the predicted annual oil consumption per person is 7.5 barrels higher than would be predicted by the regression line.

9.25. (a) There appears to be a positive relationship between poverty rate and murder rate. One point (D.C.) does appear to fall outside of the pattern of the rest of the data.



(b) $\hat{y}$ = –5.176 + 1.130x. For D.C.: $\hat{y}$ = 15.7; the residual is 28.3. The murder rate for D.C. is 28.3 higher than we would predict with the regression equation. (c) D.C. is definitely a regression outlier, because it falls far from the least squares line that would fit through the rest of the data. That estimated regression line without D.C. is $\hat{y}$ = 0.197 + 0.494x. The slope is less than half of what it was when the point for D.C. was included.

9.29. (a) H_0: b = 0 and H_a: b ? 0, $t = b/se$ = 0.294/0.0149 = 19.7, P < 0.001. There is extremely strong evidence that number of years of mother's education has a positive effect on number of years of education. (b) $b \pm t_{0.025}(se)$ = 0.294 ± 1.96(0.0149) = 0.265 to 0.323. We can be 95% confident that b falls between 0.265 and 0.323. In the population, the mean number of years of education increases by between 0.27 and 0.32 for each additional year of the mother's education. (c) Since r = 0.37 is less than 1, number of years of education is predicted to be fewer standard deviations from its mean than number of years of mother's education is from its mean.

9.32. (a) $P = 0.0015$. There is strong evidence of an association between number of hours spent in the home on religious activity and political ideology, and the estimated slope shows that it is positive. In this case, more hours spent in the home on religious activity is associated with being more conservative. (b) While these results are statistically significant, the slope of 0.0064 is so close to 0 that it has no practical significance. For example, an increase in x of 10 hours a month corresponds to a predicted change on the political ideology scale of 0.064, which is extremely small because political ideology is on a scale from 1 to 7.

9.38. A correlation matrix follows. Quality of food was most strongly correlated with the rating for service. The correlation was positive, higher rated quality of food tending to occur in restaurants with higher rated service. The other correlations with food rating (for décor and for cost) were positive, but not as strong.

Correlations

		Food	Decor	Service	Cost
Food	Pearson Correlation	1	.293	.617	.411
	Sig. (2-tailed)		.019	.000	.001
	N	64	64	64	64
Decor	Pearson Correlation	.293	1	.682	.824
	Sig. (2-tailed)	.019		.000	.000
	N	64	64	64	64
Service	Pearson Correlation	.617	.682	1	.718
	Sig. (2-tailed)	.000	.000		.000
	N	64	64	64	64
Cost	Pearson Correlation	.411	.824	.718	1
	Sig. (2-tailed)	.001	.000	.000	
	N	64	64	64	64

9.42. The parameters are the b values for the prediction of diet cost by amount of fats and sweets eaten and diet cost by amount of fruit and vegetables eaten. The statistical inference that was performed was a set of two confidence intervals for the slopes. In their interpretation, the study should have replaced "diet costs" by "expected diet costs" or "population mean diet costs".

9.50. No, an increase would tend to occur for the poorest students simply because of regression toward the mean. If the correlation were 0.50 and if the standard deviations were the same for each exam, for example, then a student who is 20 units below the mean on the midterm would be predicted to be 10 units below the mean on the final exam.

9.51. Because of regression toward the mean, we would expect the heaviest readers prewar to show less reading (on the average) during the war, and we would expect the lightest readers prewar to show more reading (on the average) during the war.

9.52. Individuals' income will tend to vary much more than does median income at the county level. For instance, in Table 9.16, the median incomes vary between 15.4 thousand and 35.6 thousand dollars, whereas individuals' incomes might vary between close to 0 to several million dollars. This additional variability should result in a serious diminishing of the correlation.

9.53. In using only students from Yale University, the values of x would very likely be highly restricted to a very narrow range, leading to a smaller correlation than one would get with the broader range of x values that occurs with a more diverse student body (e.g., the student body of the University of Bridgeport, CT).

9.54. In using only adults having a college degree, the range of values of x is restricted, leading to a smaller correlation (in absolute value) than one would get with the broader range of x values that occurs with a random sample of all adults.

9.55. (a) The standard deviation of y scores in the sample. (b) The standard deviation of x scores in the sample. (c) The estimated standard deviation for the conditional distribution of y at each fixed value of x. (d) The estimated standard error of the sample slope b.

9.58. (a) True, that correlation is largest in absolute value. (b) False, $\left|r_{yx_1}\right| < \left|r_{x_1x_2}\right|$. (c) True, because $r_{x_1x_2} <$ 0. (d) True because r_{yx_1} = 0.30. (e) False, $r^2_{yx_1}$ = 0.09 is the proportional reduction in error. (f) True, because the coefficient of 0.40 for X_2 corresponds to \$400. (g) True, since $\left|r_{yx_1}\right| < \left|r_{yx_2}\right|$, their squares have the same order, and because larger r-squared values occur with smaller SSE values (for a given total sum of squares, TSS), it follows that $SSE_1 > SSE_2$. (h) True, a 10-unit increase in x_2 is estimated to correspond to a 0.40(10) = 4.0 (i.e., \$4000) increase in mean income. (i) False, $\hat{y}$ = 10 + 1.0(10) = 20. If s = 8, then an income of \$70,000 is (70 – 20)/8 = 6.25 standard deviations above the predicted value, which would be very unusual. (j) True, since then $r_{yx_1} = b\left(s_{x_1} / s_y\right)$ = 1.0(3.6/12) = 0.30. (k) False. At x_1 = 13, $\hat{y}$ = 10 + 1.0(13) = 23. Since the least squares line must pass through the point $\left(\overline{x}, \overline{y}\right)$, this would imply that $\overline{y}$ = 23 rather than 20.

9.59. (b)

9.60. (b), (d), and (g)

9.61. (c), (f), and (g)

9.66. (a) Since $a = \overline{y} - b\overline{x}$, it follows that $a + b\overline{x} = \overline{y}$. Thus, $\hat{y} = a + b\overline{x} = \overline{y}$; the predicted value at $x = \overline{x}$ is $\hat{y} = \overline{y}$, and the line passes through $\left(\overline{x}, \overline{y}\right)$. (b) Since $\hat{y} = a + bx$ and $a = \overline{y} - b\overline{x}$, $\hat{y} = \overline{y} - b\overline{x} + bx$. This equation can be expressed as $\hat{y} - \overline{y} = b\left(x - \overline{x}\right)$. (c) When $s_x = s_y$, $b = r\left(s_y / s_x\right) = r\left(1\right) = r$. Thus, we can substitute 0.70 into the equation from part (b) to get $\hat{y} - \overline{y} = 0.70\left(x - \overline{x}\right)$.

9.67. (a) Interchange x and y in the formula, and one gets the same value. (b) If the units of measurement change, the z-score does not. For instance, if the values are doubled, then the deviation of an observation

from the mean doubles, but so does the standard deviation, and the ratio of the deviation to the standard deviation does not change.

Chapter 10

10.3. (a) No. Without as many firefighters, the damage could potentially be even worse. (b) Larger fires tend to have more firefighters at the fire and have more costly damage.

10.5. (a) Suppose there is a negative correlation. Heavier use of marijuana might cause a student to spend less time studying and get a lower GPA. Or, students who do more poorly in school might tend to get dejected and spend more time doing other things, such as using marijuana. (b) Suppose the correlation is positive. A third variable dealing with a factor such as the subject's natural curiosity or inquisitiveness could be positively associated with both variables. Students who tend to be higher in this characteristic might tend to have higher GPAs and to be more likely to experiment with marijuana.

10.7. (a) The positive correlation between shoe size and number of books read is explained by age, which is strongly positively associated with each of these. You should draw a scatterplot with points identified by the value of age, such that at a fixed age there is no association between shoe size and number of books read, but such that when you ignore the age values you see a positive association between shoe size and number of books read. (b) Draw a scatterplot with points identified by gender, such that for each gender, as height goes up there is no tendency for annual income to go up or go down. However, if the male points tend to be at higher height values and (on the average) at a higher income value, then higher height values tend to be associated with higher incomes (when you ignore the gender labeling).

10.11. An arrow should go from "race to "whether poor," with another arrow from "whether poor" to "whether arrested". To support this explanation, the association between race and arrest rates would have to disappear after you control for income.

10.14. (a)

Victim = White	DEATH PENALTY		Victim = Black	DEATH PENALTY	
DEFENDANT	Yes	No	DEFENDANT	Yes	No
White	19	132	White	0	9
Black	11	52	Black	6	97

For white victims, 12.6% of white defendants received the death penalty, whereas 17.5% of black defendants received the death penalty. For black victims, 0% of white defendants received the death penalty, whereas 5.8% of black defendants received the death penalty. In each case, the proportion receiving the death penalty was higher for black defendants than for white defendants.

(b)

DEFENDANT	Yes	No
White	19	141
Black	17	149

Ignoring the race of the victim, 11.9% of white defendants received the death penalty, whereas 10.2% of black defendants received the death penalty. (c) The proportions are higher for black defendants in each

partial table, reflecting a greater chance of the death penalty for black defendants (controlling for victim's race), but the proportion is higher for white defendants for the bivariate table, indicating the reverse association when victim's race is ignored rather than controlled. (d) Inspection of the counts in the partial tables indicates that victim's race is strongly associated with defendant's race, with whites tending to kill whites and blacks tending to kill blacks. Also, we see that killing a white person is more likely to result in the death penalty, regardless of defendant's race. So, roughly speaking, whites are tending to kill whites, and killing a white person is more likely to result in the death penalty, and these two associations combined result in, overall, a higher proportion of whites than blacks receiving the death penalty. (e) (i) is not realistic, because victim's race would be regarded as a response and defendant's race as explanatory.

10.16. (a) This is possibly a spurious relationship, whereby the percent living in metropolitan areas is positively related both to crime rate and to percent with a high school education. Even if there is no association between high school graduation rate and crime rate at each fixed level of percent living in metropolitan areas, overall there will be a positive association because of the joint positive association between percent living in metropolitan areas and each of the other variables. (b) spurious, with percent urban as an explanatory variable having an effect both on crime rate and on graduation rate.

10.32. Age may be associated with both exercising and illnesses. For example, perhaps older people are not as able to exercise and are more susceptible to illnesses.

10.33. No. A factor such as socioeconomic status (SES) may be a common cause of birth defects and of buying bottled water.

10.34. Suppose that at a fixed age, there is no difference in rates between now and the beginning of the century. Because women tend to live longer now and because the cancer rate is higher at higher ages, the overall rate of breast cancer would be higher now than at the beginning of the century.

10.39. Here, we are studying the association between whether a compulsive buyer and total credit card balances. There appeared to be no association, the mean total credit card balance being similar for compulsive buyers and non-compulsive buyers. However, income may have masked any potential association between compulsive buying and mean total credit card balances, because lower income people were more likely to be compulsive buyers and more likely to have higher credit card balances. If we controlled for income, we might find that mean total credit card balances are higher for compulsive buyers than for noncompulsive buyers.

10.40. False. Association does not imply causation. Also, there is variability. If you consider an increase of a standard deviation in saturated fat intake, the predicted depression increases by 0.68 standard deviations, but individual subjects may change by more or less than that average change.

10.42. (b)

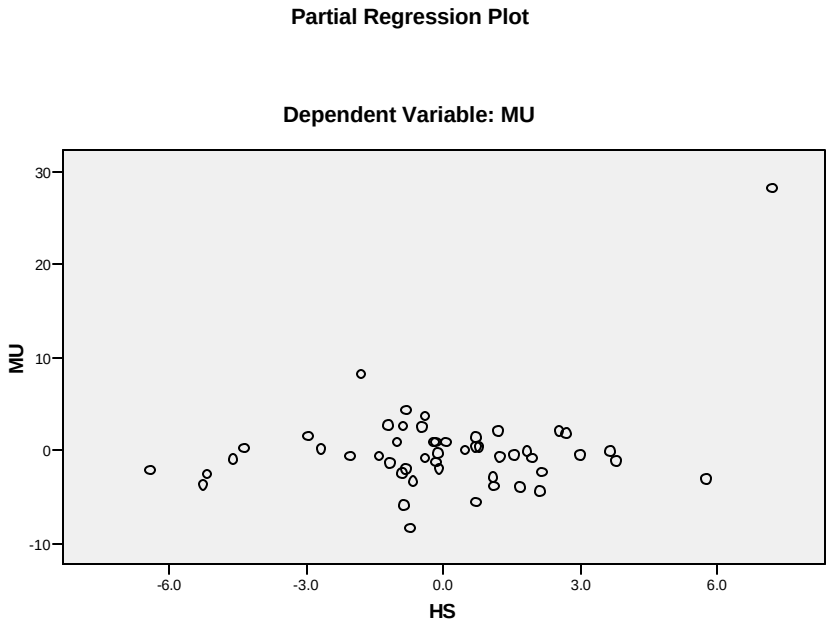
10.43. (b)

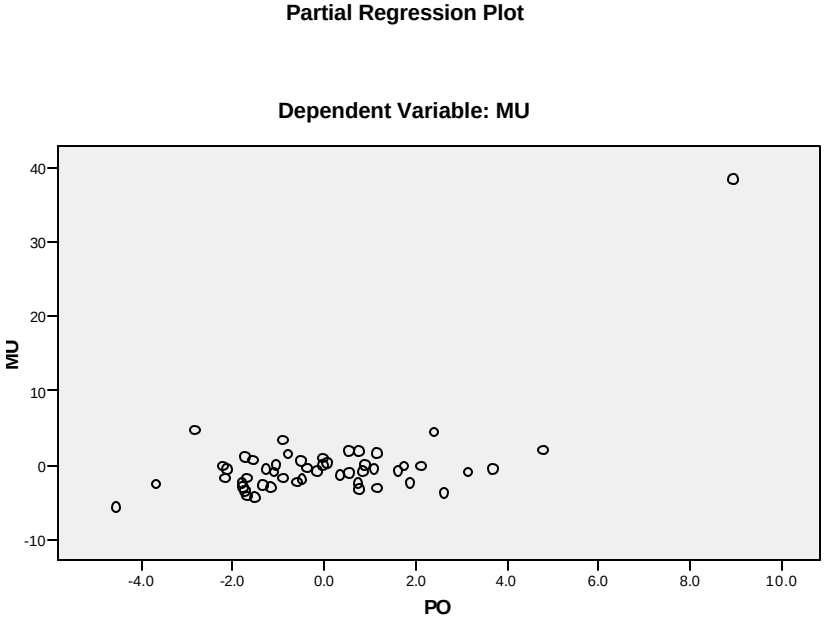
10.44. (a)

Chapter 11

11.1. (a) (i) $E(y) = 0.20 + 0.50(4.0) + 0.002(800) = 3.8$; (ii) $E(y) = 0.20 + 0.50(3.0) + 0.002(300) = 2.3$. (b) $E(y) = 0.20 + 0.50x_1 + 0.002(500) = 1.20 + 0.50x_1$. (c) $E(y) = 0.20 + 0.50x_1 + 0.002(600) = 1.40 + 0.50x_1$. (d) For instance, consider $x_1 = 3$ for which $E(y) = 1.70 + 0.002x_2$. By contrast, when $x_1 = 2$, $E(y) = 1.20 + 0.002x_2$, which has a different y-intercept but the same slope of 0.002.

11.4. (a) D.C. appears to be an outlier in each of the partial regression plots.





(b) $\hat{y} = -60.498 + 0.588x_1 + 1.605x_2$. For fixed poverty rate, the murder rate is predicted to increase by 0.588 for each additional percent increase of high school graduates. For fixed percentage of graduates, the murder rate is predicted to increase by 1.65 for each additional percent increase in the poverty rate.

Model Summary(b)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	0.636(a)	0.405	0.380	4.764

a Predictors: (Constant), PO, HS
b Dependent Variable: MU

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta	B	Std. Error
1	(Constant)	-60.498	24.615		-2.458	0.018
	HS	0.588	0.260	0.351	2.256	0.029
	PO	1.605	0.301	0.830	5.332	0.000

a Dependent Variable: MU

(c) With D.C. removed, the predicted effect of poverty rate is reduced from 1.605 to 0.304, less than a fifth as large. In addition, note that the estimated effect of percent of high school graduates now has a negative partial coefficient rather than a positive one. So, outliers can be highly influential in a regression analysis.

Model Summary(b)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	0.582(a)	0.338	0.310	2.136

a Predictors: (Constant), PO_noDC, HS_noDC
b Dependent Variable: MU_noDC

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta	B	Std. Error
1	(Constant)	18.912	12.437		1.521	0.135
	HS_noDC	-0.196	0.130	-0.278	-1.510	0.138
	PO_noDC	0.304	0.164	0.340	1.846	0.071

a Dependent Variable: MU_noDC

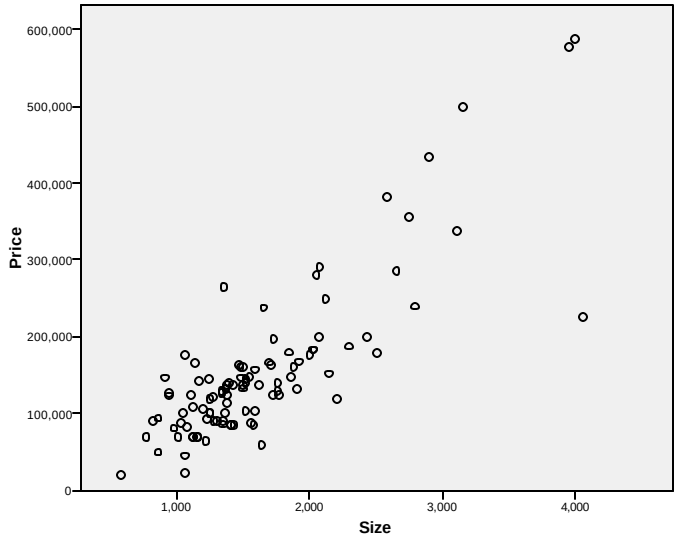
11.5. (a) $\hat{y} = -3.601 + 1.2799x_1 + 0.1021x_2$. (b) $\hat{y} = -3.601 + 1.2799(10) + 0.1021(50) = 14.3$. (c) (i) $\hat{y} = -3.601 + 1.2799x_1 + 0.1021(0) = -3.601 + 1.2799x_1$; (ii) $\hat{y} = -3.601 + 1.2799x_1 + 0.1021(100) = 6.609 + 1.2799x_1$; On average, for each increase of \$1000 in GDP, the percentage of people who use the Internet increases by 1.28. (d) Since the slope for GDP is the same for each of the two models in part (c), the lines are parallel, and there is no interaction between cell-phone use and GDP in their effects on the response variable.

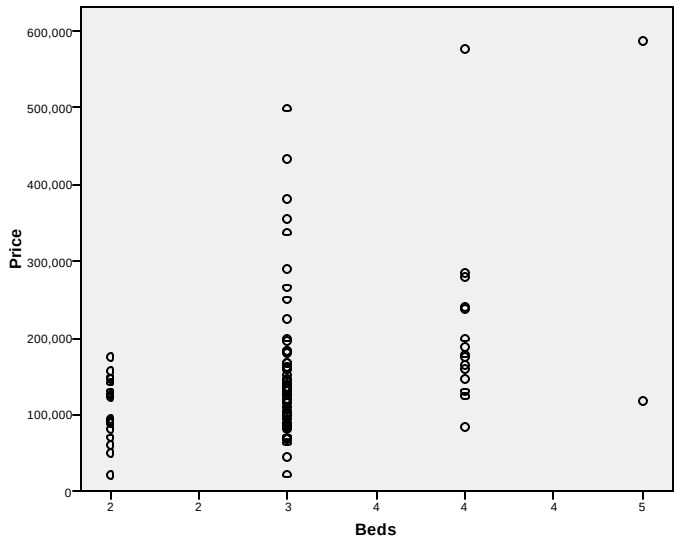
11.6. (a) R-squared = (TSS – TSE)/TSS = (12,959.3 – 2642.5)/12,959.3 = 10,316.8/12,959.3 = 0.796. There is about 80% less error when we predict cell-phone use with the prediction equation with these two predictors than when we predict it using the sample mean. (b) Since cell-phone use and GDP are strongly positively correlated, adding cell-phone use to the model will not improve the model much beyond only using GDP as a predictor of Internet use.

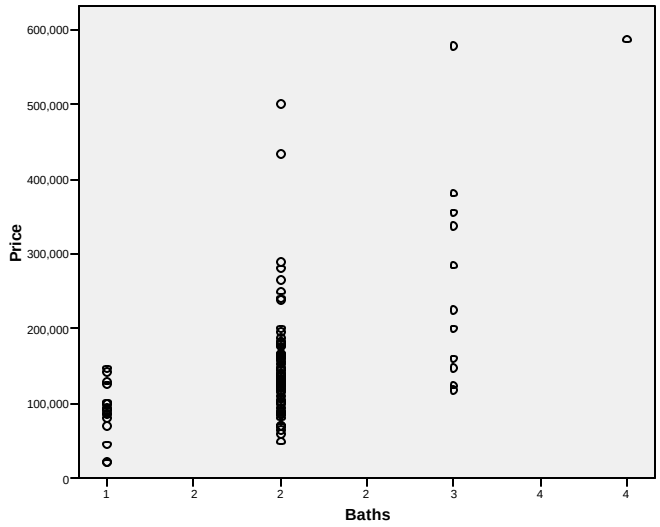
11.7. (a) Positive, since crime appears to increase as income increases. (b) Negative, since controlling for percent urban, the trend appears to be negative. (c) $\hat{y} = -11.5 + 2.6x_1$; the predicted crime rate increases

by 2.6 (per 1000 residents) for every \$1000 increase in median income. (d) $\hat{y} = 40.3 - 0.81x_1 + 0.65x_2$; the predicted crime rate decreases by 0.8 (per 1000 residents) for each \$1000 increase in median income, controlling for level of urbanization. Compared to (c), the effect is weaker and has a different direction. (e) Urbanization is highly positively correlated both with income and with crime rate. This makes the overall bivariate association between income and crime rate more positive than the partial association. Counties with high levels of urbanization tend to have high median incomes and high crime rates, whereas counties with low levels of urbanization tend to have low median incomes and low crime rates, and these effects induce the overall positive bivariate association between crime rate and income. (f) (i) $\hat{y} = 40.3 - 0.81x_1$; (ii) $\hat{y} = 73 - 0.81x_1$; (iii) $\hat{y} = 105 - 0.81x_1$. The slope stays constant, but at a fixed level of x_1, the crime rates are higher at higher levels of x_2.

11.19. (a) The scatterplots are:







There is a moderately strong positive linear relationship between the selling price of a home and its size (in square feet). In the scatterplots of selling price versus either number of bedrooms or number of bathrooms, we see stacking at the integer values for the predictors, which are highly discrete.

(b) $\hat{y} = -27{,}290 + 130x_1 - 14{,}466x_2 + 6890x_3$; \$130 = change in predicted selling price of a home for 1 square foot increase in size, controlling for number of bedrooms and number of bathrooms.

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta	B	Std. Error
1	(Constant)	-27290.075	28240.514		-0.966	0.336
	Size	130.434	11.951	0.859	10.914	0.000
	Beds	-14465.770	10583.489	-0.093	-1.367	0.175
	Baths	6890.267	13539.977	0.039	0.509	0.612

a Dependent Variable: Price

(c) (i) The size of the house has the strongest correlation with the selling price of the house. (ii) The number of bedrooms has the weakest correlation with the selling price of the house.

Correlations

		Price	Size	Beds	Baths
Price	Pearson Correlation	1	0.834(**)	0.394(**)	0.558(**)
	Sig. (2-tailed)		0.000	0.000	0.000
	N	100	100	100	100
Size	Pearson Correlation	0.834(**)	1	0.545(**)	0.658(**)
	Sig. (2-tailed)	0.000		0.000	0.000
	N	100	100	100	100
Beds	Pearson Correlation	0.394(**)	0.545(**)	1	0.492(**)
	Sig. (2-tailed)	0.000	0.000		0.000
	N	100	100	100	100
Baths	Pearson Correlation	0.558(**)	0.658(**)	0.492(**)	1
	Sig. (2-tailed)	0.000	0.000	0.000	
	N	100	100	100	100

** Correlation is significant at the 0.01 level (2-tailed).

(d) $R^2 = 0.701$ for the full model; $r^2 = 0.695$ for the simpler model using x_1 alone as the predictor. Once x_1 is in the model, x_2 and x_3 do not add much more information for predicting selling price.

11.25. (a) (i) –0.612, (ii) –0.819, (iii) 0.757, (iv) 2411.4, (v) 585.4, (vi) 29.27, (vii) 5.41, (viii) 10.47, (ix) 0.064, (x) –2.676, (xi) 0.0145, (xii) 0.007, (xiii) 31.19, (xiv) 0.0001. (b) $\hat{y} = 61.71 - 0.171x_1 - 0.404x_2$; 61.71 = predicted birth rate at ECON = 0 and LITERACY = 0 (may not be useful), –0.171 = change in predicted birth rate for 1 unit change in ECON, controlling for LITERACY, –0.404 = change in predicted birth rate for 1 unit change in LITERACY, controlling for ECON. (c) –0.612; there is a moderate negative association between birth rate and ECON;
–0.819; there is a strong negative association between birth rate and LITERACY. (d) $R^2 = (2411.4 - 585.4)/2411.4 = 0.76$; there is a 76% reduction in error in using these two variables (instead of $\overline{y}$) to predict birth rate. (e) $R = \sqrt{0.76} = 0.87$ = the correlation between observed y values and the predicted y values, which is quite strong. (f) $F = [0.757/2]/[(1 - 0.757)/20] = 31.2$, $df_1 = 2$, $df_2 = 20$, $P = 0.0001$; at least one of ECON and LITERACY has a significant effect. (g) $t = -0.171/0.064 =$
–2.676, $df = 20$, $P = 0.0145$; there is strong evidence of a relationship between birth rate and ECON, controlling for LITERACY.

11.44. Since there is essentially no effect for political conservatives and a considerably positive effect for political liberals, there is an interaction between number of years of education and political ideology. One would add an interaction term to the multiple regression model to allow the partial slope to be positive for liberals and close to 0 for conservatives. One of the predictors would be political ideology, for example on the scale of 1 to 7 that the GSS uses, ranging from very conservative to very liberal.

11.46. False. We could make this conclusion if these were standardized coefficients, but not as unstandardized coefficients. We cannot compare the sizes of partial slopes when the different predictor variables have different units of measurement.

11.48. (a) R cannot be smaller than the absolute value of any of the bivariate correlations. (b) SSE cannot increase when we add predictors, (h) R^2 cannot exceed 1. (i) This is the multiple correlation, which cannot be negative. (j) No, this is only true when $df_1 = 1$, in which case $F = t^2$. (k) We need to compare the absolute values of standardized coefficients to determine which explanatory variable has a stronger effect. We cannot compare the sizes of partial slopes when the different predictor variables have different units of measurement. (n) The product takes value over a very wide range, and a 1-unit change in the product is a trivial amount, so this coefficient is then small even if the effect is strong in practical terms.

11.49. (b), since (20 – 10)3 = 30.

11.66. (a) $\hat{y} = -26.1 + 72.6x_1 + 19.6x_2$; Older homes: $\hat{y} = -26.1 + 72.6x_1$; new homes: $\hat{y} = -6.5 + 72.6x_1$. (b) When $x_2 = 1$, the y-intercept is 19.6 higher than when $x_2 = 0$. This is the difference of estimated mean selling prices between new and older homes, controlling for house size.

11.67. (a) Older homes: $\hat{y} = -16.6 + 66.6x_1$; New homes: $\hat{y} = 15.2 + 96x_1$. New homes cost more, on average, than older homes, for each fixed value of size. In addition, the selling price for a new home increases at a higher rate as the size increases than does the selling price for an older home. (b) Older homes: $\hat{y} = -16.6 + 66.6x_1$; New homes: $\hat{y} = -7.6 + 71.6x_1$. With the outlier removed, there is not as much difference in selling price between new and older homes, and, as the size increases, the selling price

for a new home increases at a much slower rate than the model with the outlier. The outlier results in rather misleading conclusions about the difference between new and old homes in the effect of size.

Chapter 12

12.1. The response variable is the number of firefights reported. The explanatory variable was branch of service and deployment venue combination. The null hypothesis is that the population means are the same for each branch of service and deployment venue combination. The alternative hypothesis is that at least two of the population means differ.

12.2. (a) $H_0: \mu_1 = \mu_2 = \cdots = \mu_{12}$; H_a: at least two of the population means are unequal. (b) If the null hypothesis were true, we would expect F to equals about 1, with relatively large values above 1 giving strong evidence against the null. Since $F = 0.61$ does not differ from 1 by much, it does not appear that we have strong evidence against H_0. (c) If all 12 population means were equal (e.g., the mean number of good friends was not associated with astrological sign), we would see results at least as extreme as those observed with probability 0.82. There is not enough evidence to show an association between the mean number of good friends and astrological sign.

12.5. (a) (i) $H_0: \mu_1 = \mu_2 = \mu_3$; H_a: at least two of the population means are unequal. (ii) $F = 3.03$; (iii) $P = 0.049$; (iv) We reject the null hypothesis at the 0.05 significance level. (b) Yes. The standard deviations are all much larger than the means, suggesting that the three distributions are very highly skewed to the right, rather than normal.

12.6. (a) $H_0: \mu_1 = \mu_2 = \mu_3$; H_a: at least two of the population means are unequal. $F = 2.50$, $P = 0.18$. There is not enough evidence to conclude that the population mean quiz scores differ for the three groups. (b) The F statistic will be smaller, because the *between groups* variation would not be as large, so the numerator of the F statistic would be smaller. (c) The F statistic will be larger, since the *within groups* variation is smaller. (d) The F statistic would be larger, since larger sample sizes provide stronger evidence, for a given size of effect. (e) The P-values would be larger for part (b), smaller for part (c), and smaller for part (d).

12.10. (a) $F = 13.00/0.47 = 27.6$; $df_1 = 2$; $df_2 = 297$; $P = 0.000$. There is sufficient evidence to conclude that the population mean customer satisfaction ratings differ among the cities. (b) Since all of the sample sizes are the same, the margin of error for separate 95% confidence intervals will be the same for comparing the population means for each pair of cities. $1.972\sqrt{0.47\left(\frac{1}{100}+\frac{1}{100}\right)} = 0.19$. We are 95% confident that the interval –0.39 to –0.01 contains the true difference in population mean customer satisfaction ratings between San Jose and Toronto. We are 95% confident that the interval 0.31 to 0.69 contains the true difference in population mean customer satisfaction ratings between San Jose and Bangalore. We are 95% confident that the interval 0.51 to 0.89 contains the true difference in mean population customer satisfaction ratings between Toronto and Bangalore. (c) The Bonferroni method uses error probability 0.05/3 = 0.0167 for each interval, leading to a larger margin of error than for the individual confidence intervals. The Tukey method is similar in theory. The advantage of this approach is that the overall error probability is 0.05 instead of 0.05 for each interval. (d) For San Jose, $z_1 = 1$ and $z_2 =$

0, so $\hat{y}$ = 7.6. For Toronto, z_1 = 0 and z_2 = 1, so $\hat{y}$ = 7.8. For Bangalore, z_1 = 0 and z_2 = 0, so $\hat{y}$ = 7.1. Note that in the prediction equation, 7.1 is the sample mean for Bangalore, 0.5 is the difference between the sample means for San Jose and Bangalore, and 0.7 is the difference between the sample means for Toronto and Bangalore.

12.14. $H_0: \mu_1 = \mu_2 = \mu_3$; H_a: at least two of the population means are unequal. Between-groups mean square = 36.0; within-groups mean square = 45.333, F = 0.79, P = 0.48. There is not enough evidence to conclude that the population mean amount of time of REM sleep differs for the three groups. The minimum significant difference is $2.93\sqrt{45.333\left(\frac{1}{4}+\frac{1}{4}\right)} = 13.95$, which is the margin of error for each pairwise 95% confidence interval using the Bonferroni method. The means are all grouped together, in the sense that no pair of them show evidence of being different.

12.15. (a) $E(y) = \alpha + \beta_1 z_1 + \beta_2 z_2$, with z_1 = 1 for group 1 and 0 otherwise, z_2 = 1 for group 2 and 0 otherwise. $H_0: \mu_1 = \mu_2 = \mu_3$ is equivalent to $H_0: \beta_1 = \beta_2 = 0$. (b) For group 1, $\hat{y}$ = 18 – 6(1) – 3(0) = 12.0. For group 2, $\hat{y}$ = 18 – 6(0) – 3(1) = 15.0. For group 3, $\hat{y}$ = 18 – 6(0) – 3(0) = 18.0. These are the means listed at the bottom of the SAS printout. Note that in the prediction equation, 18 is the sample mean for group 3, -6 is the difference between the sample means for group 1 and group 3, and -3 is the difference between the sample means for group2 and group 3.

12.17. Sex does not appear to be a significant factor in the comparison of mean number of hours a day watching TV (P = 0.55, and the means for each sex are comparable within race). Race does appear to be a significant factor in the comparison of mean number of hours a day watching TV (P = 0.000, and the means for each race are quite different, with blacks having a higher mean than whites).

12.22. (a) The response variable is hourly wage. The two factors are sex and whether the job is classified as white-collar, blue-collar, or service jobs. (b) The population mean hourly wages are

	White-collar	Blue-collar	Service
Males	\$22	\$14	\$11
Females	\$15	\$10	\$8

(c) (i) For white-collar jobs, the mean hourly wage is \$7 more for males than for females. (ii) For blue-collar jobs, the mean hourly wage is \$4 more for males than for females. Since this difference is not the same for both job categories, there is evidence of interaction. The disparity between mean hourly wages for males and females increases from \$3 for service jobs to \$4 for blue-collar jobs to \$7 for white-collar jobs.

12.24. "No interaction" means that the differences in mean political ideology are consistent across religion and also between the sexes. The large difference in sample mean political ideology between Jewish women and Jewish men (compared to the other groups, for which the means are relatively similar) suggests interaction is present.

12.40. The P-value had to be less than 0.05/35 = 0.0014 for a given test to be significant.

12.41. For each word type (abstract or concrete), we can conclude that in the population, women have a larger verbal memory, in the sense that the mean for all women is higher than the mean for all men. The lack of interaction suggests that the difference between the population means for women and for men is similar for each word type.

12.42. (a) With the single-comparison approach, in the long run 95% of the intervals will contain the true differences in means. With the multiple comparisons method, in the long run 95% of the time the entire set of comparisons is correct, and any one comparison has confidence greater than 95%. (b) Suppose the sample means equal 8 for A, 15 for B, and 22 for C.

12.45.

a.	10	10	b.	10	20	c.	10	20	d.	10	10
	20	20		30	40		30	60		10	10

12.48. (a)

	Science	Humanities
Women	80,000 (5)	66,000 (25)
Men	79,000 (30)	65,000 (20)

The overall means (weighted averages) are $68,333 for women and $73,400 for men. The mean is higher overall for men even though it is lower in each division. (b) Controlling for division, women have the higher mean, whereas for a one-way analysis ignoring division, men have the higher mean.

12.51. (a), (b), (c), and (d)

12.52. (c)

12.57. (a) This is the sample variance of the g sample mean values, and we know the theoretical variance of the sampling distribution of the sample mean equals the population variance divided by the sample size. (b) Multiply both sides by n in part (a).