* FINAL EXAM REVIEW EXERCISES *
(EVERYDAY STATS. A)

In Exercises 33–36, identify which of these designs is most appropriate for the given experiment: completely randomized design, randomized block design, or matched pairs design.

33. Lunesta Lunesta is a drug designed to treat insomnia. In a clinical trial of Lunesta, amounts of sleep each night are measured before and after subjects have been treated with the drug.

34. Lipitor A clinical trial of Lipitor treatments is being planned to determine whether its effects on diastolic blood pressure are different for men and women.

35. West Nile Vaccine Currently, there is no approved vaccine for the prevention of infection by West Nile virus. A clinical trial of a possible vaccine is being planned to include subjects treated with the vaccine while other subjects are given a placebo.

36. HIV Vaccine The HIV Trials Network is conducting a study to test the effectiveness of two different experimental HIV vaccines. Subjects will consist of 80 pairs of twins. For each pair of twins, one of the subjects will be treated with the DNA vaccine and the other twin will be treated with the adenoviral vector vaccine.

37. Simple Random Sample vs. Random Sample Refer to the definition of *simple random sample* on page 28 and its accompanying definition of *random sample* enclosed within parentheses. Determine whether each of the following is a simple random sample and a random sample.

a. In Major League Baseball, there are 30 teams, each with an active roster of 25 players. The names of the teams are printed on 30 separate index cards, the cards are shuffled, and one card is drawn. The sample consists of the 25 players on the active roster of the selected team.

b. For the same Major League Baseball population described in part (a), the 750 names of the players are printed on 750 separate index cards, and the cards are shuffled. Twenty-five different cards are selected from the top. The sample consists of the 25 selected players.

c. For the same Major League Baseball population described in part (a), a sample is constructed by selecting the 25 youngest players.

1-4 Ethics in Statistics

The website www.TriolaStats.com includes a downloadable section that discusses ethical issues in statistics, including those related to data collection, analysis, and reporting.

Chapter Quick Quiz

1. Survey An example in this chapter referred to a survey of 410 human resource professionals. If those subjects are identified with numbers from 1 through 410, does it make sense to calculate the average (mean) of those numbers?

2. Survey Which of the following best describes the level of measurement of the numbers 1, 2, 3, . . . , 410 described in Exercise 1: nominal, ordinal, interval, ratio?

3. Survey In the same survey cited in Exercise 1, are the exact unrounded ages of the 410 subjects discrete data or continuous data?

4. Survey In the same survey cited in Exercise 1, are the exact unrounded ages of the 410 subjects quantitative data or categorical data?

5. Survey Which of the following best describes the level of measurement of the exact unrounded ages of the 410 survey subjects from Exercise 1: nominal, ordinal, interval, ratio?

* Do circled/boxed problems
* Skip scribbled out/x'd out problems
* Check answers using Appendix D attachment

6. Birth Weights For 100 randomly selected births from Bellevue Hospital Center, the birth weights are added and then divided by 100. The result is 3240 g. Is the value of 3240 g a statistic or a parameter?

7. Birth Weights Refer to the sample described in Exercise 6. Because Bellevue Hospital Center agreed to provide the 100 birth weights, does the sample of birth weights constitute a voluntary response sample?

8. Birth Weights Are the data described in Exercise 6 the result of an observational study or an experiment?

9. Physicians' Health Study In the Physicians' Health Study, some of the subjects were treated with aspirin while others were given a placebo. For the subjects in this experiment, what is *blinding*?

10. Sampling In a statistical study, which of the following types of samples is generally best: convenience sample, voluntary response sample, simple random sample, biased sample?

Review Exercises

1. Online Medical Info *USA Today* posted this question on its website: "How often do you seek medical information online?" Of 1072 Internet users who chose to respond, 38% of them responded with "frequently." What term is used to describe this type of survey in which the people surveyed consist of those who decided to respond? What is wrong with this type of sampling method?

2. Paying for First Dates *USA Today* posted this question on the electronic version of its newspaper: "Should guys pay for the first date?" Of the 1148 subjects who decided to respond, 85% of them said "yes."

a. What is wrong with this survey?

b. Is the value of 85% a statistic or a parameter?

c. Does the survey constitute an experiment or an observational study?

3. Sample Design Literacy In "High-Flow Oxygen for Treatment of Cluster Headache" (*Journal of the American Medical Association*, Vol. 302, No. 22), the authors explain that 150 patients were treated with oxygen, and 148 patients were given a placebo. The authors summarize the sample design as "randomized and double-blind." Describe the meaning of "randomized" and "double-blind" in the context of this study.

4. Divorces and Margarine One study showed that there is a very high correlation between the divorce rate in Maine and per capita consumption of margarine in the United States. Can we conclude that either one of those two variables is the cause of the other?

5. Sampling For each of the following, identify the term that best describes the type of sample: *systematic*, *convenience*, *stratified*, *cluster*, or *simple random sample*.

a. As Lipitor pills are being manufactured, a quality control plan is to select every 500th pill and test it to confirm that it contains 80 mg of atorvastatin.

b. To test for a gender difference in the way that men and women make online purchases, Gallup surveys 500 randomly selected men and 500 randomly selected women.

c. A list of all 1,736,997 adults in Manhattan is obtained; the list is numbered from 1 to 1,736,997; and then a computer is used to randomly generate 500 different numbers between 1 and 1,736,997. The sample consists of the adults corresponding to the selected numbers.

d. A statistics student creates a survey and presents it to fellow statistics students.

e. The Commissioner of Major League Baseball obtains a sample by randomly selecting one team from the American League and one team from the National League, and all players on the selected teams are surveyed.

6. Defense of Marriage Act Both of the following questions are essentially the same. Does the difference in wording seem as though it could affect the way that people respond?

- Are you in favor of the "Defense of Marriage Act"?
- Are you in favor of an act that for federal and state aid, only heterosexual marriages should be recognized?

7. State Populations Currently, California has the largest population with 39,776,830 residents, and Wyoming has the smallest population with 573,520 residents.

a. Are the population sizes of the different states discrete or continuous?

b. What is the level of measurement for the numbers of residents in the different states? (nominal, ordinal, interval, ratio)

c. What is wrong with surveying state residents by mailing questionnaires to 50,000 of them who are randomly selected?

d. If we randomly select 50 full-time workers in each of the 50 states, what type of sample is obtained? (random, systematic, convenience, stratified, cluster)

e. If we randomly select two states and survey all of their adult residents, what type of sample is obtained? (random, systematic, convenience, stratified, cluster)

8. Percentages

a. The labels on U-Turn protein energy bars include the statement that these bars contain "125% less fat than the leading chocolate candy brands" (based on data from *Consumer Reports* magazine). What is wrong with that claim?

b. In a Pew Research Center poll on driving, 58% of the 1182 respondents said that they like to drive. What is the actual number of respondents who said that they like to drive?

c. In a Pew Research Center poll on driving, 331 of the 1182 respondents said that driving is a chore. What percentage of respondents said that driving is a chore?

9. Types of Data In each of the following, identify the level of measurement of the sample data (nominal, ordinal, interval, ratio) and the type of sampling used to obtain the data (random, systematic, convenience, stratified, cluster).

a. At Albany Medical Center, every 10th newborn baby is selected and the body temperature is measured (degrees Fahrenheit).

b. In each of the 50 states, 50 voters are randomly selected and their political party affiliations are identified.

c. A pollster stops each person passing her office door and asks the person to rate the last movie that he or she saw (on a scale of 1 star to 4 stars).

10. Statistical Significance and Practical Significance The Genetics and IVF Institute developed a procedure designed to increase the likelihood that a baby would be a boy. In a clinical trial of their procedure, 239 boys were born among 291 births. If the method has no effect, there is less than a 1% chance that such extreme results would occur. Does the procedure appear to have statistical significance? Does the procedure appear to have practical significance?

2-4 Beyond the Basics

P-Values *In Exercises 13–16, write a statement that interprets the P-value and includes a conclusion about linear correlation.*

13. Using the data from Exercise 5 "Forecast and Actual Temperatures," the P-value is 0.166.

14. Using the data from Exercise 6 "Airport Data Speeds," the P-value is 0.003.

15. Using the data from Exercise 7 "Cigarette Tar and Nicotine," the P-value is 0.000.

16. Using the data from Exercise 8 "Pulse Rates," the P-value is 0.835.

Chapter Quick Quiz

1. Tornado Alley Refer to the accompanying frequency distribution that summarizes the number of tornadoes in Oklahoma in each year for the past several years. What is the class width? Is it possible to identify the original data values?

Annual Tornadoes in Oklahoma	Frequency (Number of Years)
0–19	3
20–39	18
40–59	7
60–79	15
80–99	6
100–119	5
120–139	0
140–159	1

2. Tornado Alley Using the same frequency distribution from Exercise 1, identify the class limits of the first class and the class boundaries of the first class.

3. Tornado Alley Using the same frequency distribution from Exercise 1, how many years are included?

4. Tornado Alley Construct the relative frequency distribution corresponding to the frequency distribution in Exercise 1.

5. Tornado Alley A stemplot of the same data summarized in Exercise 1 is created, and one of the rows of that stemplot is 3 7 0 0 3 1 4 6 6 9. Identify the values represented by that row of the stemplot.

6. Computers As a quality control manager at Texas Instruments, you find that defective calculators have various causes, including worn machinery, human error, bad supplies, and packaging mistreatment. Which of the following graphs would be best for describing the causes of defects: histogram; scatterplot; Pareto chart; dotplot; pie chart?

7. Health Test In an investigation of a relationship between systolic blood pressure and diastolic blood pressure of adult females, which of the following graphs is most helpful: histogram; pie chart; scatterplot; stemplot; dotplot?

8. Lottery In Florida's Play 4 lottery game, four digits between 0 and 9 inclusive are randomly selected each day. We normally expect that each of the 10 different digits will occur about 1/10 of the time, and an analysis of last year's results shows that this did happen. Because the results are what we normally expect, is it correct to say that the distribution of selected digits is a normal distribution?

9. Seatbelts The Reams Seatbelts company manufactures—well, you know. When a sample of seatbelts is tested for breaking point (measured in kilograms), the sample data are explored. Identify the important characteristic of data that is missing from this list: center, distribution, outliers, changing characteristics over time.

10. Normal Distribution If the following data are randomly selected, which are expected to have a normal distribution?

a. Weights of Reese's Peanut Butter Cups

b. Numbers selected in the Florida Pick 4 lottery, in which four whole numbers between 0 and 9 inclusive are randomly selected in each lottery

c. Numbers that turn up when a fair die is rolled

d. Exact volumes of Coke in 12 oz cans

e. Weights of McIntosh apples harvested from the same orchard

Review Exercises

1. Email Data Listed below are the interarrival times (minutes) of email arriving at the author's computer. Construct a frequency distribution. Use a class width of 10 minutes and begin with a lower class limit of 0 minutes. Compare the distribution of these interarrival times to the distribution of the commute times listed in the Chapter Problem (Table 2-1) and summarized in the frequency distribution of Table 2-2.

19	58	39	3	61	17	21	13	1	20	1	12	13	7	8	33	3	2	34	18
1	20	16	18	37	25	15	1	1	6	29	5	5	3	7	9	7	5	6	38
19	6	37	18	24	53	1	41	2	6	9	9	35	1	10	11	28	15	8	18

2. Histogram of Interarrival Times Construct the histogram that corresponds to the frequency distribution from Exercise 1. Use class midpoint values for the horizontal scale. Does the histogram suggest that the data are from a population having a normal distribution? Why or why not?

3. Dotplot of Interarrival Times Construct a dotplot of the interarrival times listed in Exercise 1. Which does a better job of illustrating the distribution of the data: the histogram from Exercise 2 or the dotplot?

4. Stemplot of Interarrival Times Construct a stemplot of the interarrival times listed in Exercise 1. Are there any outliers?

5. Body Temperatures Listed below are the temperatures from nine males measured at 8 AM and again at 12 AM (from Data Set 5 "Body Temperatures" in Appendix B). Construct a scatterplot. Based on the graph, does there appear to be a relationship between 8 AM temperatures and 12 AM temperatures?

8 AM	98.0	97.6	98.6	97.4	97.4	98.2	98.2	98.6	97.4
12 AM	98.0	97.6	98.8	98.0	98.8	98.8	97.6	98.6	98.6

6. Environment

a. After collecting the average (mean) global temperatures for each of the most recent 100 years, we want to construct the graph that is most appropriate for these data. Which graph is best?

b. After collecting the average (mean) global temperature and the amount of carbon monoxide emissions for the most recent 100 years, we want to construct a graph to investigate the association between those two variables. Which graph is best?

c. An investigation of carbon monoxide sources includes motor vehicles, furnaces, fires, coal-burning power plants, and tobacco smoke. If we want to construct a graph that illustrates the relative importance of these sources, which graph is best?

7. It's Like Time to Do This Exercise In a Marist survey of adults, these are the words or phrases that subjects find most annoying in conversation (along with their frequencies of response): like (127); just sayin' (81); you know (104); whatever (219); obviously (35). Construct a pie chart. Identify one disadvantage of a pie chart.

8. Whatever Use the same data from Exercise 7 to construct a Pareto chart. Which graph does a better job of illustrating the data: Pareto chart or pie chart?

Cumulative Review Exercises

In Exercises 1–5, use the data listed in the margin, which are magnitudes (Richter scale) and depths (km) of earthquakes from Data Set 24 "Earthquakes" in Appendix B.

Magnitude	Depth (km)
2.45	0.7
3.62	6.0
3.06	7.0
3.30	5.4
1.09	0.5
3.10	0.0
2.99	7.0
2.58	17.6
2.44	7.0
2.91	15.9
3.38	11.7
2.81	7.0
2.44	7.0
3.56	6.9
2.79	7.3
2.18	7.0
3.01	7.0
2.71	7.0
2.44	8.1
1.64	7.0

1. Frequency Distribution Construct a frequency distribution of the magnitudes. Use a class width of 0.50 and use a starting value of 1.00.

2. Frequency Distribution For the frequency distribution from Exercise 1, find the following.

a. Class limits of the first class

b. Class boundaries of the first class

c. Class midpoint of the first class

3. Histogram Construct the histogram corresponding to the frequency distribution from Exercise 1. For the values on the horizontal axis, use the class midpoint values. Which of the following comes closest to describing the distribution: uniform, normal, skewed left, skewed right?

4. Data Type

a. The listed earthquake *depths* (km) are all rounded to one decimal place. Before rounding, are the exact depths discrete data or continuous data?

b. For the listed earthquake depths, are the data categorical or quantitative?

c. Identify the level of measurement of the listed earthquake depths: nominal, ordinal, interval, or ratio.

d. Given that the listed earthquake depths are part of a larger collection of depths, do the data constitute a sample or a population?

5. Correlation: Between Magnitudes and Depths Using the paired magnitude/depth data, construct the graph that is helpful in determining whether there is a correlation between earthquake magnitudes and depths. Based on the result, does there appear to be a correlation?

Technology Project

Graphs It was stated in this chapter that the days of charming and primitive hand-drawn graphs are well behind us, and technology now provides us with powerful tools for generating a wide variety of different graphs. This project therefore serves as a good preparation for professional presentations that will be inevitably made in the future.

The complete data sets in Appendix B are already included in Statdisk. Also, those data sets can be downloaded from www.TriolaStats.com. They can be opened by statistical software packages, such as Minitab, Excel, SPSS, and JMP. Use a statistical software package to open Data Set 1 "Body Data." Use this software with the methods of this chapter to explore and compare the pulse rates of females and the pulse rates of males.

- Obtain a printed copy of the two histograms. In both cases, use a class width of 10 beats per minute and use 30 beats per minute as the lower class limit of the first class.
- Describe the natures of the two distributions (uniform, normal, skewed left, skewed right), and identify possible outliers.

continued

3-3 Beyond the Basics

37. Outliers and Modified Boxplots Repeat Exercise 33 "Pulse Rates" using modified boxplots. Identify any outliers as defined in Part 2 of this section.

38. Outliers and Modified Boxplots Repeat Exercise 34 "Ages of Oscar Winners" using modified boxplots. Identify any outliers as defined in Part 2 of this section.

Chapter Quick Quiz

1. Mean of Roller Coaster Speeds Listed below are maximum speeds (km/h) of randomly selected roller coasters in the United States. Find the mean.

70 76 97 81 57 151 194 65 117 65 45 105

2. Median of Roller Coaster Speeds What is the median of the sample values listed in Exercise 1?

3. Mode of Roller Coaster Speeds What is the mode of the sample values listed in Exercise 1?

4. Variance of Roller Coaster Speeds The standard deviation of the sample values in Exercise 1 is 43.1 km/h. What is the variance (including units)?

5. Roller Coaster Speed Outlier Identify any outliers among the data listed for Exercise 1.

6. Roller Coaster *z* Score A larger sample of 92 roller coaster maximum speeds has a mean of 85.9 km/h and a standard deviation of 28.7 km/h. What is the *z* score for a speed of 34 km/h? Does the *z* score suggest that the speed of 34 km/h is *significantly low*?

7. Q_3 for Roller Coaster Speeds For the sample of 92 roller coaster maximum speeds, approximately how many of those speeds are less than Q_3?

8. Roller Coaster Speed 5-Number Summary For the sample of 92 roller coaster maximum speeds, give the *names* of the values that constitute the 5-number summary. (The actual values can't be identified; just give the *names* of those values.)

9. Estimating s The sample of 92 roller coaster maximum speeds includes values ranging from a low of 10 km/h to a high of 194 km/h. Use the range rule of thumb to estimate the standard deviation.

10. Roller Coaster Speed Notation Consider a sample of roller coaster maximum speeds taken from the population of all roller coasters operating on our planet. Identify the symbols used for the sample mean, population mean, sample standard deviation, population standard deviation, sample variance, and the population variance.

Review Exercises

1. Reported and Measured Heights Listed below are self-reported heights of males aged 16 and over and their corresponding measured heights (based on data from the National Health and Nutrition Examination Survey). All heights are in inches. First find the differences (reported height–measured height), and then use those differences to find the (a) mean, (b) median, (c) mode, (d) midrange, (e) range, (f) standard deviation, (g) variance, (h) Q_1, (i) Q_3.

Reported	68.0	71.0	63.0	70.0	71.0	60.0	65.0	64	54.0	63.0	66	72.0
Measured	67.9	69.9	64.9	68.3	70.3	60.6	64.5	67	55.6	74.2	65	70.8

2. Outliers Identify any of the differences found from Exercise 1 that appear to be outliers. For any outliers, how much of an effect do they have on the mean, median, and standard deviation?

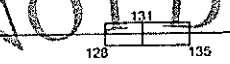
3. z Score Using the differences from Exercise 1, find the z score corresponding to the difference of -11.2 in. Is that difference significantly low, significantly high, or neither?

4. Boxplot Using the same differences from Exercise 1, construct a boxplot and include the values of the 5-number summary.

5. ER Codes In an analysis of activities that resulted in brain injuries presenting at hospital emergency rooms, the following activities were identified by the codes shown in parentheses: bicycling (12); football (14); playground (22); basketball (27); swimming (40). Find the mean of 12, 14, 22, 27, and 40. What is wrong with this result?

6. MCAT In a recent year, applicants to medical schools achieved scores on the Medical College Admission Test (MCAT) with a mean of 504.7 and a standard deviation of 9.4. Identify the MCAT scores that are significantly low or high.

7. Interpreting a Boxplot Shown below is a boxplot of a sample of 30 maximal skull widths (mm) measured from Egyptian skulls from around 4000 B.C. What do the numbers in the boxplot represent?



8. Estimating Standard Deviation Listed below are sorted weights (g) of a sample of M&M plain candies randomly selected from one bag. Use the range rule of thumb to estimate the value of the standard deviation of all 345 M&Ms in the bag. Compare the result to the standard deviation of 0.0366 g computed from all of the 345 M&Ms in the bag.

0.799 0.843 0.849 0.855 0.870 0.872 0.874 0.875 0.879 0.886
0.887 0.889 0.894 0.901 0.902 0.902 0.926 0.940 0.943 0.944

9. Percentiles Use the sorted weights of M&Ms from the preceding exercise to find the value of P_{25}. How does the result compare to the value of Q_1?

10. Comparing Birth Weights The birth weights of a sample of females have a mean of 3037.1 g and a standard deviation of 706.3 g. The birth weights of a sample of males have a mean of 3272.8 g and a standard deviation of 660.2 g (based on Data Set 6 "Births" in Appendix B). When considered among members of the same gender, which baby has the relatively larger birth weight: a female with a birth weight of 3200 g or a male with a birth weight of 3400 g? Why?

Cumulative Review Exercises

1. Sugar Listed below are measured weights (mg) of sugar in Domino packets labelled as containing 3500 mg (or 3.5 g).

a. Are the data qualitative or quantitative?

b. What is the level of measurement of the data (nominal, ordinal, interval, or ratio)?

c. Before any rounding, are the weights discrete or continuous?

d. Given that the weights are from Domino sugar packets selected from a much larger population, are the weights a sample or a population?

e. If we calculate the mean of the listed values, is the result a statistic or a parameter?

3511 3516 3521 3531 3532 3545 3583 3588 3590 3610
3621 3635 3638 3643 3645 3647 3666 3673 3678 3736

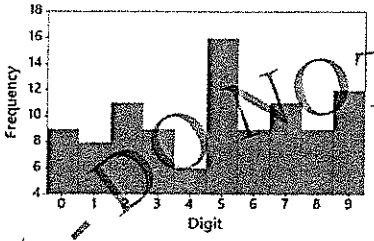
2. Frequency Distribution Using the data from Exercise 1, construct a frequency distribution using a class width of 50 mg and a first class with a lower class limit of 3500 mg.

3. Histogram Use the frequency distribution from Exercise 2 to construct a histogram. Use class midpoint values for the horizontal scale.

4. Percentile Use the weights from Exercise 1 to find the percentile for 3647 mg.

5. Descriptive Statistics Use the weights of the Domino sugar packets from Exercise 1 and find the following: (a) mean, (b) median, (c) standard deviation, (d) variance, (e) range. Include the appropriate units of measurement.

6. Histogram The accompanying histogram depicts outcomes of digits from the Florida Play 4 lottery. What is the major flaw in this histogram?



7. Normal Distribution Examine the distribution shown in the histogram from Exercise 6. Does it appear that the sample data are from a population with a normal distribution? Why or why not?

8. Correlation of Heights of Fathers and Sons Listed below are heights (in.) of fathers and their first sons (based on Data Set 10 "Family Heights" in Appendix B). Construct a scatterplot and then make a judgment to determine whether there appears to be a correlation between heights of fathers and heights of their first sons.

Father	70.5	69.0	68.7	70.0	68.0	65.0	69.0	68.7	67.0	65.0
Son	74.0	68.0	67.7	68.0	67.5	66.5	71.0	71.0	69.0	66.0

Technology Project

Words Spoken by Men and Women Refer to Data Set 14 "Word Counts" in Appendix B, which includes counts of words spoken by males and females. That data set includes 12 columns of data, but first stack all of the male word counts in one column and stack all of the female word counts in another column. Then proceed to generate histograms, any other suitable graphs, and find appropriate statistics that allow you to compare the two sets of data. Are there any outliers? Do both data sets have properties that are basically the same? Are there any significant differences? What would be a consequence of having significant differences? Write a brief report including your conclusions and supporting graphs.

10-4 Basic Skills and Concepts

Statistical Literacy and Critical Thinking

1. Response and Predictor Variables Using all of the Tour de France bicycle race results up to a recent year, we get this multiple regression equation: Speed = 29.2 − 0.00260 Distance + 0.540 Stages + 0.0570 Finishers, where Speed is the mean speed of the winner (km/h), Distance is the length of the race (km), Stages is the number of stages in the race, and Finishers is the number of bicyclists who finished the race. Identify the response and predictor variables.

2. Best Multiple Regression Equation For the multiple regression equation given in Exercise 1, the P-value is 0.000 and the adjusted R^2 value is 0.894. If we were to include an additional predictor variable of the number of bicyclists who entered the race, the P-value is 0.000 and the adjusted R^2 is again 0.894. Is it correct to reason that we should include the number of bicyclists who entered the race because the adjusted R^2 remains the same but the multiple regression equation gives us more information because it includes another variable? Explain.

3. Adjusted Coefficient of Determination For Exercise 2, why is it better to use values of adjusted R^2 instead of simply using values of R^2?

4. Interpreting R^2 For the multiple regression equation given in Exercise 1, we get $R^2 = 0.897$. What does that value tell us?

Interpreting a Computer Display. ***In Exercises 5–8, we want to consider the correlation between heights of fathers and mothers and the heights of their sons. Refer to the StatCrunch display and answer the given questions or identify the indicated items. The display is based on Data Set 10 "Family Heights" in Appendix B. (The response y variable represents heights of sons.)***

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	17.966577	6.4779134	≠ 0	131	2.7735131	0.0064
Father	0.50354896	0.067077219	≠ 0	131	7.507004	<0.0001
Mother	0.27714316	0.078318967	≠ 0	131	3.5386467	0.0006

Analysis of variance table for multiple regression model:

Source	DF	SS	MS	F-stat	P-value
Model	2	320.94662	160.47331	37.837221	<0.0001
Error	131	558.54293	4.2636865		
Total	133	879.48955			

Summary of fit:
Root MSE: 2.0643696
R-squared: 0.3649
R-squared (adjusted): 0.3552

5. Height of Son Identify the multiple regression equation that expresses the height of a son in terms of the height of his father and mother.

6. Height of Son Identify the following:

a. The P-value corresponding to the overall significance of the multiple regression equation

b. The value of the multiple coefficient of determination R^2

c. The adjusted value of R^2

7. Height of Son Should the multiple regression equation be used for predicting the height of a son based on the height of his father and mother? Why or why not?

8. Height of Son A son will be born to a father who is 70 in. tall and a mother who is 60 in. tall. Use the multiple regression equation to predict the height of the son. Is the result likely to be a good predicted value? Why or why not?

Garbage: Finding the Best Multiple Regression Equation. ***In Exercises 9–12, refer to the accompanying table, which was obtained by using the data from 62 households listed in Data Set 42 "Garbage Weight" in Appendix B. The response (y) variable is PLAS (weight of discarded plastic in pounds). The predictor (x) variables are METAL (weight of discarded metals in pounds), PAPER (weight of discarded paper in pounds), and GLASS (weight of discarded glass in pounds).***

Predictor (x) Variables	P-Value	R^2	Adjusted R^2	Regression Equation
METAL/PAPER/GLASS	0.000	0.563	0.540	PLAS = −0.170 + 0.290 METAL + 0.122 PAPER + 0.0777 GLASS
METAL/PAPER	0.000	0.514	0.498	PLAS = 0.00394 + 0.344 METAL + 0.111 PAPER
PAPER/GLASS	0.000	0.499	0.482	PLAS = 0.0647 + 0.167 PAPER + 0.0967 GLASS
METAL/GLASS	0.000	0.392	0.371	PLAS = 0.469 + 0.619 METAL + 0.0774 GLASS
METAL	0.000	0.344	0.333	PLAS = 0.641 + 0.573 METAL
PAPER	0.000	0.421	0.411	PLAS = 0.348 + 0.166 PAPER
GLASS	0.005	0.126	0.111	PLAS = 1.46 + 0.121 GLASS

9. If only one predictor (x) variable is used to predict the weight of discarded plastic, which single variable is best? Why?

10. If exactly two predictor (x) variables are to be used to predict the weight of discarded plastic, which two variables should be chosen? Why?

11. Which regression equation is best for predicting weight of discarded plastic? Why?

12. A household discards 3.00 lb of metal, 10.25 lb of paper, and 9.35 lb of plastic. What is the best predicted value for the weight of discarded plastic? Is that predicted value likely to be a good estimate? Is that predicted value likely to be very accurate?

Appendix B Data Sets. ***In Exercises 13–16, refer to the indicated data set in Appendix B and use technology to obtain results.***

13. Predicting Car Fuel Consumption Refer to Data Set 35 "Car Data" in Appendix B and use the weight, engine displacement, and highway fuel consumption (HWY) of all 48 cars. Find the best regression equation for predicting the highway fuel consumption. Why is it best? Is the best regression equation a good regression equation for predicting the highway fuel consumption? Why or why not?

14. Predicting Height Refer to Data Set 3 "ANSUR II 2012" in Appendix B and use the variables of Height, Foot_Length, and Arm_Span for all 6068 subjects. Find the best regression equation for predicting Height. Why is it best? Is the best regression equation a good regression equation for predicting Height? Why or why not?

15. Predicting IQ Score Refer to Data Set 12 "IQ and Brain Size" in Appendix B and find the best regression equation with IQ score as the response (y) variable. Use predictor variables of brain volume and/or body weight. Why is this equation best? Based on these results, can we predict someone's IQ score if we know their brain volume and body weight? Based on these results, does it appear that people with larger brains have higher IQ scores?

16. Global Warming Listed below are mean annual temperatures (°C) of the earth for each decade, beginning with the decade of the 1880s. Find the best model and then predict the value for 2090–2099. Comment on the result.

13.819	13.692	13.741	13.788	13.906	14.016	14.082
13.983	13.938	14.014	14.264	14.396	14.636	

10-5 Beyond the Basics

17. Moore's Law In 1965, Intel cofounder Gordon Moore initiated what has since become known as *Moore's law*: The number of transistors per square inch on integrated circuits will double approximately every 18 months. In the table below, the first row lists different years and the second row lists the number of transistors (in thousands) for different years.

1971	1974	1978	1982	1985	1989	1993	1997	2000	2002	2003	2007	2011	2018
2.3	5	29	120	275	1180	3100	7500	42,000	220,000	410,000	789,000	2,060,000	19,200,000

a. Ignoring the listed data and assuming that Moore's law is correct and transistors per square inch double every 18 months, which mathematical model best describes this law: linear, quadratic, logarithmic, exponential, power? What specific function describes Moore's law?

b. Which mathematical model best fits the listed sample data?

c. Compare the results from parts (a) and (b). Does Moore's law appear to be working reasonably well?

18. Sum of Squares Criterion In addition to the value of R, another measurement used to assess the quality of a model is the *sum of squares of the residuals*. Recall from Section 10-2 that a residual is $y - \hat{y}$ (the difference between an observed y value and the value predicted from the model). Better models have smaller sums of squares. Refer to the U.S. population data in Table 10-7 on page 565.

a. Find $\Sigma(y - \hat{y})^2$, the sum of squares of the residuals resulting from the linear model.

b. Find the sum of squares of residuals resulting from the quadratic model.

c. Verify that according to the sum of squares criterion, the quadratic model is better than the linear model.

Chapter Quick Quiz

Exercises 1–10 are based on the following sample data consisting of costs of dinner (dollars) and the amounts of tips (dollars) left by diners. The data were collected by students of the author.

Cost of Dinner (dollars)	46.60	33.46	50.68	87.92	98.84	63.60	107.34	49.88
Tip (dollars)	7.50	5.50	5.00	8.08	17.00	12.00	16.00	7.00

1. Scatterplot Construct a scatterplot and comment on the pattern of points.

2. Conclusion The linear correlation coefficient r is found to be 0.846, the P-value is 0.008, and the critical values for a 0.05 significance level are ± 0.707. What should you conclude?

3. Fixed Percentage If a restaurant were to change its tipping policy so that a constant tip of 20% of the bill is added to the cost of the dinner, what would be the value of the linear correlation coefficient for the paired amounts of dinners/tips?

4. Fixed Percentage If a restaurant were to change its tipping policy so that a constant tip of 20% of the bill is added to the cost of the dinner, what would be the linear regression equation relating the cost of the dinner (x) and the amount of the tip (y)?

~~**5. Switched Variables** Which of the following values change if the two variables of dinner cost and amount of tip are switched: the value of $r = 0.846$, the P-value of 0.008, the critical values of ±0.707?~~

6. Change in Scale Exercise 1 stated that for the given paired data, $r = 0.846$. How does that value change if all of the amounts of dinners are left unchanged but all of the tips are expressed in cents instead of dollars?

7. Values of r If you had computed the value of the linear correlation coefficient to be 1.200, what should you conclude?

8. Predictions The sample data result in a linear correlation coefficient of $r = 0.846$ and the regression equation $\hat{y} = -0.00777 + 0.145x$. What is the best predicted amount of tip, given that the cost of dinner was \$84.62? How was the predicted value found?

9. Predictions Repeat the preceding exercise assuming that the linear correlation coefficient is $r = 0.132$.

10. Explained Variation Given that the linear correlation coefficient r is found to be 0.846, what is the proportion of the variation in tips that is explained by the linear relationship between amounts of dinner and amounts of tips? What is the proportion of the variation that cannot be explained by that linear relationship?

Review Exercises

1. Casino Size and Revenue Listed below are sizes (in thousands of square feet) and revenue (in millions of dollars) from casinos in Atlantic City (based on data from the *New York Times*). Is there sufficient evidence to conclude that there is a linear correlation between size and revenue of casinos? Can a casino increase its revenue by enlarging its physical size?

Size	160	227	140	144	161	147	141
Revenue	189	157	140	127	123	106	101

2. Casino Size and Revenue Use the same paired data from the preceding exercise.

a. Find the linear regression equation.

b. What is the best predicted amount of revenue for a casino with a size of 200 thousand square feet? Is it likely that the best predicted amount of revenue will be accurate?

3. Time and Motion In a physics experiment at Doane College, a soccer ball was thrown upward from the bed of a moving truck. The table below lists the time (sec) that has lapsed from the throw and the corresponding height (m) of the soccer ball.

a. Find the value of the linear correlation coefficient r.

b. Based on the result from part (a), what do you conclude about a linear correlation between time and height?

c. What horrible mistake would be easy to make if the analysis is conducted without a scatterplot?

Time (sec)	0.0	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8
Height (m)	0.0	1.7	3.1	3.9	4.5	4.7	4.6	4.1	3.3	2.1

4-5 Beyond the Basics

13. Simulating the Monty Hall Problem A problem that once attracted much attention is the *Monty Hall problem*, based on the old television game show *Let's Make a Deal*, hosted by Monty Hall. Suppose you are a contestant who has selected one of three doors after being told that two of them conceal nothing, but that a new red Corvette is behind one of the three. Next, the host opens one of the doors you didn't select and shows that there is nothing behind it. He then offers you the choice of sticking with your first selection or switching to the other unopened door. Should you stick with your first choice or should you switch? Develop a simulation of this game and determine whether you should stick or switch. (According to *Chance* magazine, business schools at such institutions as Harvard and Stanford use this problem to help students deal with decision making.)

14. Simulating Birthdays

a. Develop a simulation for finding the probability that when 50 people are randomly selected, at least 2 of them have the same birth date. Describe the simulation and estimate the probability.

b. Develop a simulation for finding the probability that when 50 people are randomly selected, at least 3 of them have the same birth date. Describe the simulation and estimate the probability.

15. Genetics: Simulating Population Control A classical probability problem involves a king who wanted to increase the proportion of women by decreeing that after a mother gives birth to a son, she is prohibited from having any more children. The king reasons that some families will have just one boy, whereas other families will have a few girls and one boy, so the proportion of girls will be increased. Conduct a simulation to determine whether his reasoning is correct, and to determine whether the proportion of girls will increase.

Chapter Quick Quiz

1. ESP A psychologist tells you that in an ESP (extrasensory perception) experiment, there is a 20% chance of answering a question correctly. What is the *probability* of answering a question correctly?

2. Standard Tests Standard tests, such as the SAT or ACT or MCAT, tend to make extensive use of multiple-choice questions because they are easy to grade using software. If one such multiple choice question has possible correct answers of a, b, c, d, e, what is the probability of a wrong answer if the answer is a random guess?

3. Birthday If a day of a year (not a leap year) is randomly selected, what is the probability it is the author's birthday?

4. Online Courses Based on data from a survey sponsored by Sallie Mae, 10% of undergraduate students take online courses only. If two undergraduate students are randomly selected, what is the probability that they both take online courses only?

5. Subjective Probability Estimate the probability that the next time you watch a TV news report, it includes a story about a plane crash.

In Exercises 6–10, use the following results from tests of an experiment to test the effectiveness of an experimental vaccine for children (based on data from USA Today). Express all probabilities in decimal form.

	Developed Flu	Did Not Develop Flu
Vaccine Treatment	14	1056
Placebo	95	437

6. If 1 of the 1602 subjects is randomly selected, find the probability of getting 1 that developed flu.

7. If 1 of the 1602 subjects is randomly selected, find the probability of getting 1 who had the vaccine treatment or developed flu.

8. If 1 of the 1602 subjects is randomly selected, find the probability of getting 1 who had the vaccine treatment and developed flu.

9. Find the probability of randomly selecting 2 subjects without replacement and finding that they both developed flu.

10. Find the probability of randomly selecting 1 of the subjects and getting 1 who developed flu, given that the subject was given the vaccine treatment.

Review Exercises

In Exercises 1–10, use the data in the accompanying table and express all results in decimal form. (The data are from "The Left-Handed: Their Sinister History," by Elaine Costas, Education Resources Information Center, Paper 399519.)

	Writes with Left Hand?	
	Yes	No
Male	23	217
Female	65	455

1. Female If one of the subjects in the study is randomly selected, find the probability of getting a female. Does it appear that the proportion of females is reasonably close to the proportion of females in the general population?

2. Lefty Given Female Find the probability of randomly selecting one of the study subjects and getting someone who writes with their left hand given that the selected person is a female.

3. Female Given Lefty Find the probability of randomly selecting one of the study subjects and getting a female given that the selected person writes with their left hand.

4. Lefty or Female Find the probability of randomly selecting one of the study subjects and getting someone who writes with their left hand or is a female.

5. Lefty or Male Find the probability of randomly selecting one of the study subjects and getting someone who writes with their left hand or is a male.

6. Both Lefties If two of the study subjects are randomly selected *without replacement*, find the probability that they both write with their left hand.

7. Both Lefties If two of the study subjects are randomly selected *with replacement*, find the probability that they both write with their left hand.

8. Complement If L represents the event of randomly selecting one of the study subjects and getting someone who writes with their left hand, what does $\overline{L}$ represent? Find the value of $P(\overline{L})$.

9. Complement If M represents the event of randomly selecting one of the study subjects and getting someone who is a male, what does $\overline{M}$ represent? Find the value of $P(\overline{M})$.

10. All Three Lefties If three of the study subjects are randomly selected *without replacement*, find the probability that they all write with their left hand. If we did get three lefties when three subjects were randomly selected, would that be a significantly high number of lefties?

11. Random Seats on Ryanair When four researchers checked into a Ryanair flight from Manchester to Dublin, there were 65 seats available, and 15 of them were middle seats (based on data from "How 'Random' is Ryanair's Seating Allocation" by Jennifer Rogers, *Significance*). All four researchers were assigned middle seats. If the four researchers were assigned seats randomly, what is the probability that they are all given middle seats? What does the result suggest about Ryanair's claim that seats are randomly assigned?

12. Vision Correction About 75% of the U.S. population uses some type of vision correction (such as glasses or contact lenses).

a. If someone is randomly selected, what is the probability that he or she does not use vision correction?

b. If four different people are randomly selected, what is the probability that they all use vision correction?

c. What is the general criterion for using probability to determine whether a number of successes among n trials is *significantly high*?

d. If you randomly select four people, is a result of all four using vision correction significantly high? Why or why not?

13. National Statistics Day

a. If a person is randomly selected, find the probability that his or her birthday is October 18, which is National Statistics Day in Japan. Ignore leap years.

b. If a person is randomly selected, find the probability that his or her birthday is in October. Ignore leap years.

c. Estimate a subjective probability for the event of randomly selecting an adult American and getting someone who knows that October 18 is National Statistics Day in Japan.

d. If ten adult Americans are randomly selected and nine of them know that October 18 is National Statistics Day in Japan, is that result of nine significantly high?

14. Composite Sampling for Diabetes Currently, the rate for new cases of diabetes in a year is 4.3 per 1000 (based on data from the Centers for Disease Control and Prevention). When testing for the presence of diabetes, the Newport Diagnostics Laboratory saves money by combining blood samples for tests. The combined sample tests positive if at least one person has diabetes. If the combined sample tests positive, then the individual blood tests are performed. In a test for diabetes, blood samples from 10 randomly selected subjects are combined. Find the probability that the combined sample tests positive with at least 1 of the 10 people having diabetes. Is it likely that such combined samples test positive?

15. Texas Two Step In the Texas Two Step lottery, winning the top prize requires that you select the correct four different numbers from 1 to 35 (in any order and without replacement) and you must also select the correct additional "Bonus Ball" number between 1 and 35, which is drawn separately. The additional Bonus Ball number could be the same as one of the first four selected numbers. What is the probability of winning the top prize? (Express the answer as a fraction.) As this exercise was written, the jackpot was advertised to be $200,000; does that seem fair?

Cumulative Review Exercises

1. Cloud Seeding The "Florida Area Cumulus Experiment" was conducted by using silver iodide to seed clouds with the objective of increasing rainfall. For the purposes of this exercise, let the daily amounts of rainfall be represented by units of rnfl. (The actual rainfall amounts are in cubic meters $\times$ 10,000,000 or $m^3 \times 10^7$.)

Find the value of the following statistics and include appropriate units based on rnfl as the unit of measurement.

15.53 7.27 7.45 10.39 4.70 4.50 3.44 5.70 8.24 7.30 4.05 4.46

a. mean b. median ~~c. midrange~~ ← use graphing calc. for a)-f)

d. range e. standard deviation f. variance

2. Cloud Seeding Use the same data given in Exercise 1.

a. Identify the 5-number summary. As in Exercise 1, use rnfl to represent the units of measurement.

b. Construct a boxplot.

c. Identify any values that appear to be outliers.

3. Organ Donors *USA Today* provided information about a survey (conducted for Donate Life America) of 5100 adult Internet users. Of the respondents, 2346 said they are willing to donate organs after death. In this survey, 100 adults were surveyed in each state and the District of Columbia, and results were weighted to account for the different state population sizes.

a. What percentage of respondents said that they are willing to donate organs after death?

b. Based on the poll results, what is the probability of randomly selecting an adult who is willing to donate organs after death?

c. What term is used to describe the sampling method of randomly selecting 100 adults from each state and the District of Columbia?

4. Sampling: Eye Color Based on a study by Dr. P. Sorita Soni at Indiana University, assume that eye colors in the United States are distributed as follows: 40% brown, 35% blue, 12% green, 7% gray, 6% hazel.

a. A statistics instructor collects eye color data from her students. What is the name for this type of sample?

b. Identify one factor that might make the sample from part (a) biased and not representative of the general population of people in the United States.

c. If one person is randomly selected, what is the probability that this person will have brown or blue eyes?

d. If two people are randomly selected, what is the probability that at least one of them has brown eyes?

5. Heights of Presidents Theories have been developed about the heights of winning candidates for the U.S. presidency and the heights of candidates who were runners-up. Listed below are heights (cm) from recent presidential elections. Construct a graph suitable for exploring an association between heights of presidents and the heights of the presidential candidates who were runners-up. What does the graph suggest about that association?

Winner	182	177	185	188	188	183	188	191
Runner-Up	180	183	177	173	188	185	175	169

* REVIEW SOLUTIONS *

3. The sample appears to be a convenience sample. By e-mailing the survey to a readily available group of Internet users, it was easy to obtain results. Although there is a real potential for getting a sample group that is not representative of the population, indications of which car is used for cell phone calls and which hand is dominant do not appear to be factors that would be distorted much by a sample bias.
7. With 717 responses, the response rate is 14% which does appear to be quite low. In general, a very low response rate creates a serious potential for getting a biased sample that consists of those with a special interest in the topic.
9. Systematic 11. Random 13. Cluster
15. Stratified 17. Random 19. Convenience
21. Observational study. The sample is a convenience sample consisting of subjects who decided themselves to respond. Such voluntary response samples have a high chance of not being representative of the larger population, so the sample may well be biased. The question was posted in an electronic edition of a newspaper, so the sample is biased from the beginning.
23. Experiment. This experiment would create an *extremely* dangerous and illegal situation that has a real potential to result in injury or death. It's difficult enough to drive in New York City while being completely sober.
25. Experiment. The biased sample created by using a small sample of college students cannot be fixed by using a larger sample. The larger sample will still be a biased sample that is not representative of the population of all adults.
27. Observational study. Respondents who have been convicted of felonies are not likely to respond honestly to the second question. The survey will suffer from a "social desirability bias" because subjects will tend to respond in ways that will be viewed favorably by those conducting the survey.
29. Prospective study 31. Cross-sectional study
33. Matched pairs design 35. Completely randomized design
37. a. Not a simple random sample, but it is a random sample.
    b. Simple random sample and also a random sample.
    c. Not a simple random sample and not a random sample.

Chapter 1: Quick Quiz

1. No. The numbers do not measure or count anything.
2. Nominal 3. Continuous 4. Quantitative data
5. Ratio 6. Statistic 7. No
8. Observational study
9. The subjects did not know whether they were getting aspirin or the placebo.
10. Simple random sample

Chapter 1: Review Exercises

1. The respondents are a voluntary response sample or a self-selected sample. Because those with strong interests in the topic are more likely to respond, it is very possible that their responses do not reflect the opinions or behavior of the general population.
2. a. The sample is a voluntary response sample, so the results are questionable.
   b. Statistic c. Observational study
3. Randomized: Subjects were assigned to the different groups through a process of random selection, whereby they had the same chance of belonging to each group. Double-blind: The subjects did not know which of the two groups they were in, and the people who evaluated results did not know either.
4. No. Correlation does not imply causality.
5. a. Systematic b. Stratified
   c. Simple random sample d. Convenience
   e. Cluster
6. Yes. The two questions give the false impression that they are addressing very different issues. Most people would be in favor of defending marriage, so the first question is likely to receive a substantial number of "yes" responses. The second question better describes the issue and subjects are much more likely to have varied responses.
7. a. Discrete b. Ratio
   c. The mailed responses would be a voluntary response sample, so those with strong opinions or greater interest in the topics are more likely to respond. It is very possible that the results do not reflect the true opinions of the population of all state residents.
   d. Stratified e. Cluster
8. a. If they have no fat at all, they have 100% less than any other amount with fat, so the 125% figure cannot be correct.
   b. 686 c. 28%
9. a. Interval data; systematic sample
   b. Nominal data; stratified sample
   c. Ordinal data; convenience sample
10. Because there is less than a 1% chance of getting the results by chance, the method does appear to have statistical significance. The result of 239 boys in 291 births is a rate of 82% so it is above the 50% rate expected by chance, and it does appear to be high enough to have practical significance. The procedure appears to have both statistical significance and practical significance.

Chapter 1: Cumulative Review Exercises

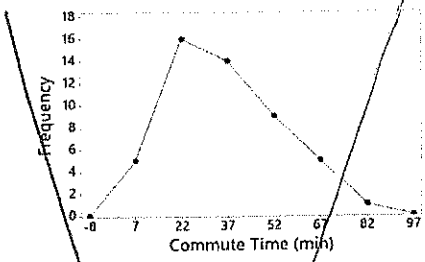
1. 133.0. The IQ score of 188 appears to be substantially higher than the other IQ scores.
2. 0.000122 3. 4.50 is a significantly high value.
4. −6.64 5. 1068
6. 20.25 7. 0.364
8. 0.20 9. 0.000729
10. 68,719,476,736 (or about 68,719,477,000)
11. 377,149,515,625 (or about 377,149,520,000)
12. 0.0000000004096

Chapter 2 Answers

Section 2-1

1. The table summarizes 1000 commute times. It is not possible to identify the exact values of all of the original data amounts.

15. The distribution does not have much skewness.



17. Because the vertical scale starts with a frequency of 200 instead of 0, the difference between the "no" and "yes" responses is greatly exaggerated. The graph makes it appear that about *five* times as many respondents said "no," when the ratio is actually a little less than 2.5 to 1.

19. The two costs are one-dimensional in nature, but the baby bottles are three-dimensional objects. The $4500 cost isn't even twice the $2600 cost, but the baby bottles make it appear that the larger cost is about five times the smaller cost.

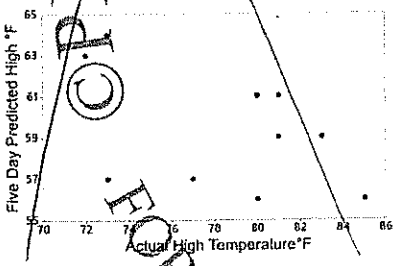
21.
```
96. | 59
97. | 0001112333444
97. | 5566666678889999
98. | 000000000000022233444444444444
98. | 55556666666666677777788888889
99. | 001244
99. | 56
```

Section 2-4

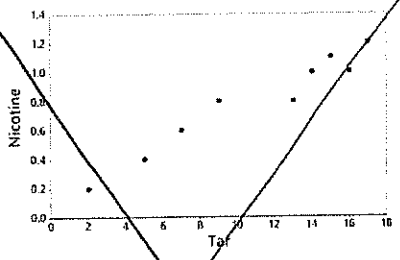
1. The term *linear* refers to a straight *line*, and *r* measures how well a scatterplot of the sample paired data fits a straight-line pattern.

3. A scatterplot is a graph of paired (x, y) quantitative data. It helps us by providing a visual image of the data plotted as points, and such an image is helpful in enabling us to see patterns in the data and to recognize that there may be a correlation between the two variables.

5. There does not appear to be a correlation. The given data suggest that five-day predicted high temperatures are not very accurate.



7. There does appear to be a linear correlation between the amounts of tar and the amounts of nicotine in cigarettes.



9. With $n = 10$ pairs of data, the critical values are ± 0.632. Because $r = -0.475$ is between -0.632 and 0.632, there is not sufficient evidence to conclude that there is a linear correlation.

11. With $n = 9$ pairs of data, the critical values are ± 0.666. Because $r = 0.971$ is in the right tail region beyond 0.666, there is sufficient evidence to conclude that there is a linear correlation.

13. Because the *P*-value is 0.166, which is not small (such as 0.05 or less), there is a high chance (16.6%) of getting the sample results when there is no correlation, so there is not sufficient evidence to conclude that there is linear correlation.

15. Because the *P*-value of 0.000 is small (such as 0.05 or less), there is a small chance of getting the sample results when there is no correlation, so there is sufficient evidence to conclude that there is a linear correlation.

Chapter 2: Quick Quiz

1. Class width: 20. It is not possible to identify the original data values.
2. Class limits: 0 and 19. Class boundaries: −0.5 and 19.5.
3. 69
4.

Annual Tornadoes in Oklahoma	Relative Frequency
0–19	4.3%
20–39	26.1%
40–59	30.4%
60–79	21.7%
80–99	8.7%
100–119	7.2%
120–139	0.0%
140–159	1.4%

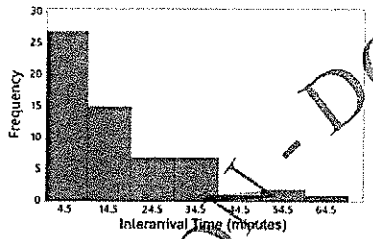
5. 30, 30, 30, 31, 34, 34, 36, 36, 39
6. Pareto chart
7. Scatterplot
8. No, the term "normal distribution" has a different meaning than the term "normal" that is used in ordinary speech. A normal distribution has a bell shape, but the randomly selected lottery digits will have a uniform or flat shape.
9. Variation
10. Parts a, d, e describe normally distributed data.

Chapter 2: Review Exercises

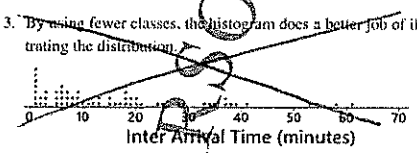
1. Both distributions are skewed to the right.

Email Interarrival Time (minutes)	Frequency
0–9	27
10–19	15
20–29	7
30–39	7
40–49	1
50–59	2
60–69	1

2. Because the histogram has a shape that is far from being bell-shaped, it suggests that the data are from a population *not* having a normal distribution.



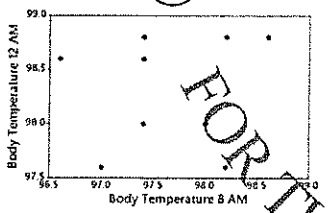
3. By using fewer classes, the histogram does a better job of illustrating the distribution.



4. There are no outliers.

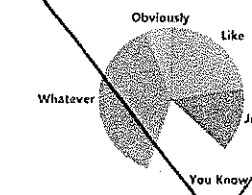
```
0 | 11111112233355566677788999
1 | 01233556788899
2 | 00 1589
3 | 3457789
4 | 1
5 | 38
6 | 1
```

5. No. There is no pattern suggesting that there is a relationship.



6. a. Time-series graph b. Scatterplot c. Pareto chart

7. A pie chart wastes ink on components that are not data; pie charts lack an appropriate scale; pie charts don't show relative sizes of different components as well as some other graphs, such as a Pareto chart.



8. The Pareto chart does a better job. It draws attention to the most annoying words or phrases and shows the relative sizes of the different categories.

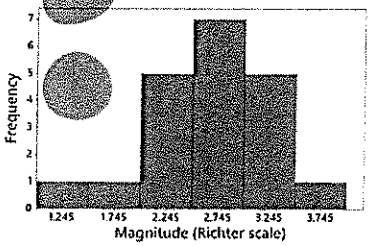
Chapter 2: Cumulative Review Exercises

1.

Magnitude	Frequency
1.00–1.49	1
1.50–1.99	1
2.00–2.49	5
2.50–2.99	7
3.00–3.49	5
3.50–3.99	1

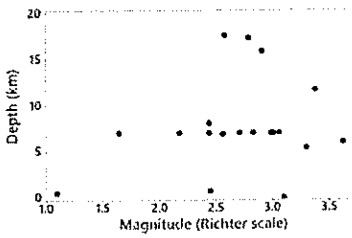
2. a. 1.00 and 1.49 b. 0.995 and 1.495 c. 1.245

3. The distribution is closer to being a normal distribution than the others. Only the single lowest value of 1.09 prevents perfect symmetry, but that one value should not be a basis for stating that the distribution is skewed left.



4. a. Continuous b. Quantitative c. Ratio d. Simple

5. The scatterplot does not show any pattern. There does not appear to be correlation between magnitude and depth.



Chapter 3 Answers

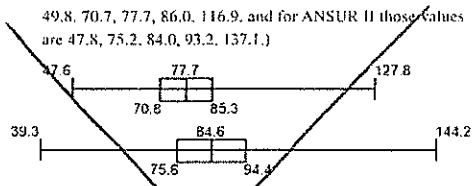
Section 3-1

1. The term *average* is not used in statistics. The term *mean* should be used for the result obtained by adding all of the sample values and dividing the total by the number of sample values.
3. They use different approaches for providing a value (or values) of the center or middle of the sorted list of data.
5. $\bar{x}$ = 47.8; median = 60.0; mode = none; midrange = 49.0. The resulting statistics are meaningless because the jersey numbers are nominal data that are just replacements for names, and they do not measure or count anything.
7. $\bar{x}$ = \$2281 million; median = \$1450 million; mode = \$1000 million; midrange = \$3225 million. Apart from the fact that all other celebrities have amounts of net worth lower than those given, nothing meaningful can be known about the population of net worth of all celebrities. The numbers all end in 0, and they appear to be rounded estimates (which is the reason for rounding to the nearest whole number).
9. $\bar{x}$ = 76.4 attacks; median = 77.5 attacks; mode = no mode; midrange = 76.0 attacks. The data are time-series data, but the measures of center do not reveal anything about a trend consisting of a pattern of change over time.
11. $\bar{x}$ = \$1[illegible]8.[illegible]; median = \$200.0; mode = \$250; midrange = \$175.0. The lowest price is a relevant statistic for someone planning to buy one of the smart thermostats.
13. $\bar{x}$ = 32.[illegible] mg; median = 39.5 mg; mode = 0 mg; midrange = 27.5 mg. Americans consume some brands much more often than others, but the 20 brands are all weighted equally in the calculations, so the statistics are not necessarily representative of the population of all cans of the same 20 brands consumed by Americans.
15. $\bar{x}$ = 1.2; median = 1.0; mode = 1; midrange = 1.5. The statistics are meaningless because the data are at the nominal level of measurement with the numbers being replacements for "right" and "left." Because the measurements were made in 1988, they are not necessarily representative of the current population of all Army women.
17. $\bar{x}$ = \$365.3; median = \$200.0; mode = \$500; midrange = \$1269.5. The amounts of \$1500 and \$2500 appear to be outliers.
19. $\bar{x}$ = 2.8 cigarettes; median = 0.0 cigarettes; mode = 0 cigarettes; midrange = 25.0 cigarettes. Because the selected subjects *report* the number of cigarettes smoked, it is very possible that the data are not at all accurate. And what about that person who smokes 50 cigarettes (or 2.5 packs) a day? What are they thinking?
21. Systolic: $\bar{x}$ = 127.6 mm Hg; median = 124.0 mm Hg. Diastolic: $\bar{x}$ = 73.6 mm Hg; median = 75.0 mm Hg. Given that systolic and diastolic blood pressures measure different characteristics, a comparison of the measures of center doesn't make sense. Because the data are matched, it would make more sense to investigate whether there is an association or *correlation* between systolic blood pressure measurements and diastolic blood pressure measurements.
23. Males: $\bar{x}$ = 69.5 beats per minute; median = 6[illegible].0 beats per minute. Females: $\bar{x}$ = 82.1 beats per minute; median = 84.0 beats per minute. The pulse rates of males appear to be lower than those of females.
25. ANSUR I 1988: $\bar{x}$ = 78.49 kg and median = 77.70 kg. ANSUR II 2012: $\bar{x}$ = 85.52 kg and median = 84.60 kg. It does appear that males have become heavier. (TI data: ANSUR I 1988: $\bar{x}$ = 78.82 kg and median = 77.[illegible]0 kg. ANSUR II 2012: $\bar{x}$ = 84.53 kg and median = 84.00 kg.)
27. $\bar{x}$ = 98.20°F; median = 98.40°F. These results suggest that the mean is less than 98.6°F.
29. $\bar{x}$ = 34.6 minutes, which is reasonably close to the mean of 31.4 minutes obtained by using the original list of values.
31. $\bar{x}$ = 55.1 years. The mean from the frequency distribution is quite close to the mean of 55.2 years obtained by using the original list of values.
33. 3.14; yes
35. a. 70 years b. $n - 1$
37. 504 lb is an outlier. Median: 285.5 lb; mean: 294.4 lb; 10% trimmed mean: 285.4 lb; 20% trimmed mean: 285.8 lb. The median, 10% trimmed mean, and 20% trimmed mean are all quite close, but the untrimmed mean of 294.4 lb differs from them because it is strongly affected by the inclusion of the outlier.
39. 0.247%
41. The median found using the given expression is 30.5 minutes. The median of the 1000 times from Data Set 31 is 30.0 minutes. The difference is 0.5 minute.

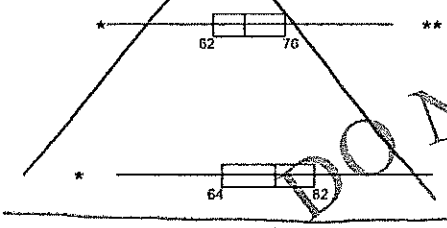
Section 3-2

1. 9.58 cm is in the general ballpark of the standard deviation of 7.10 cm calculated using the 153 heights. The range rule of thumb does not necessarily give an estimate of s that is very accurate.
3. 50.41 cm^2
5. Range = 76.0; s^2 = 755.4; s = 27.5. Because the jersey numbers are really just replacements for names, they are at the nominal level of measurement, so the results are meaningless.
7. Range = \$4550 million; s^2 = 2,825,670 (million dollars)^2; s = \$1681 million dollars. Because the data are from celebrities with the highest net worth, the measures of variation are not at all typical for all celebrities. Because all of the amounts end with 0, it appears that they are rounded to the nearest ten million dollars, so it would make sense to round the results to the nearest million dollars, as is done here.
9. Range = 44.0 attacks; s^2 = 132.7 attacks^2; s = 11.5 attacks. The measures of variation are blind to any trend for these time-series data.

49.8, 70.7, 77.7, 86.0, 116.9, and for ANSUR II those values are 47.8, 75.2, 84.0, 93.2, 137.1.)



37. Top boxplot represents males. Males appear to have slightly lower pulse rates than females. Outliers for males: 40 beats per minute, 102 beats per minute, 104 beats per minute. Outliers for females: 36 beats per minute.

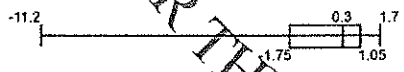


Chapter 3: Quick Quiz

1. 93.6 km/h
2. 78.5 km/h
3. 65 km/h
4. 1857.6 $(km/h)^2$
5. The speed of 194 km/h appears to be an outlier because it is substantially greater than the other speeds.
6. −1.81; no
7. About 75% of the speeds are less than Q_3.
8. Minimum, first quartile Q_1, second quartile Q_2 (or median), third quartile Q_3, maximum
9. 46.0 km/h (from range/4)
10. $\bar{x}$, μ, s, σ, s^2, σ^2

Chapter 3: Review Exercises

1. a. −1.00 in. b. 0.30 in. c. None d. 1.34 in.
   e. 12.90 $in.^2$ f. 3.52 in. g. 12.39 $in.^2$
   h. −1.75 in. (Tech: Minitab: −1.83 in.; Excel: −1.675 in.)
   i. 1.05 in. (Tech: Minitab: 1.07 in.; Excel: 1.025 in.)
2. The difference of −11.2 in. appears to be an outlier. If that outlier is excluded, the mean changes from −1.00 in. to −0.07 in., the median changes from 0.30 in. to 0.50 in., and the standard deviation changes from 3.52 in. to 1.51 in. The outlier has a strong effect on the mean and standard deviation, but very little effect on the median.
3. $z = -2.90$. The difference of −11.2 in. is significantly low (because its z score is less than or equal to −2).
4. 5-number summary: −11.2 in., −1.75 in., 0.30 in., 1.05 in., 1.70 in. (Tech: Minitab yields $Q_1 = -1.83$ in. and $Q_3 = 1.07$ in. Excel yields $Q_1 = -1.675$ in. and $Q_3 = 1.025$ in.)



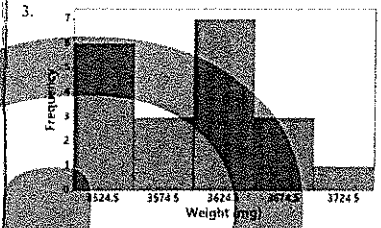
5. 23.0. The numbers don't measure or count anything. They are used as replacements for the names of the categories, so the numbers are at the nominal level of measurement, and in this case the mean is a meaningless statistic.
6. Significantly low values are less than or equal to 485.9; significantly high values are greater than or equal to 523.5.
7. The minimum is 119 mm, the first quartile Q_1 is 128 mm, the second quartile Q_2 (or median) is 131 mm, the third quartile Q_3 is 135 mm, and the maximum is 141 mm.
8. With a minimum of 0.799 g and a maximum of 0.944 g, s is estimated to be range/4 = 0.0363 g, which is very close to the standard deviation of 0.0366 g.
9. $P_{25} = 0.871$ g. The value of P_{25} is the same as the value of Q_1. (Tech: Minitab yields 0.8705 and Excel yields 0.8715.)
10. The female has the larger relative birth weight because her z score of 0.23 is larger than the z score of 0.19 for the male.

Chapter 3: Cumulative Review Exercises

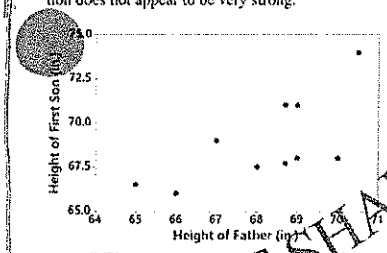
1. a. Quantitative b. Ratio level of measurement c. Continuous
   d. Sample e. Statistic
2.

Weight (mg)	Frequency
3500–3549	6
3550–3599	3
3600–3649	7
3650–3699	3
3700–3749	1

3.



4. 75th percentile
5. a. 3605.2 mg b. 3619.0 mg c. 62.4 mg
   d. 3895.3 mg^2 e. 272.0 mg
6. The vertical scale does not begin at 0, so the differences among the different outcomes are exaggerated.
7. No. A normal distribution would appear in a histogram as being bell-shaped, but the histogram is not bell-shaped.
8. Based on the scatterplot, there does appear to be a correlation between heights of fathers and heights of their first sons. Because the points are not very close to a straight-line pattern, the correlation does not appear to be very strong.



37. With $n = 703$, there are 701 degrees of freedom. From Table A-3 use the closest t value of 1.965 in the given formula to get the critical values of ± 0.074. Using a more accurate value of $t = 1.963354$ from technology leads to the same critical values of ± 0.074.

Section 10-2

1. a. $\hat{y}$ represents the predicted value of highway fuel consumption.
   b. Slope: -0.00749; y-intercept: 58.9
   c. The predictor variable is weight which is represented by x.
   d. 36.4 mi/gal
3. a. A residual is a value of $y - \hat{y}$, which is the difference between an observed value of y and a predicted value of y.
   b. The regression line has the property that the sum of squares of the residuals is the lowest possible sum.
5. With no significant linear correlation, the best predicted value is $\bar{y} = 37.3$ mi/gal.
7. With a significant linear correlation, the best predicted value is 92.0 kg.
9. $\hat{y} = 3.00 + 0.500x$. The data have a pattern that is not a straight line.
11. a. $\hat{y} = 0.264 + 0.906x$
    b. $\hat{y} = 2 + 0x$ (or $\hat{y} = 2$)
    c. The results are very different, indicating that one point can dramatically affect the regression equation.
13. $\hat{y} = 7.97 + 0.0750x$. Best predicted value: $\bar{y} = 25.6$ million tickets. The best predicted value is very different from the actual value of 90 million tickets that were sold.
15. $\hat{y} = 1.06 + 0.0452x$. Best predicted value: $\bar{y} = \$1.68$. The best predicted value is very different from the actual tip of $4.55.
17. $\hat{y} = 5.19 + 2.70x$. Best predicted value: $13.55 (or $13.50). The best predicted value is close to the actual fare of $15.30.
19. $\hat{y} = 50.0 - 0.0880x$. Best predicted value: $\bar{y} = 46.4$ years. The best predicted value isn't close to the actual value of 60 years.
21. $\hat{y} = 0.0329 + 0.969x$. Best predicted value: $3.91.
23. $\hat{y} = 350 + 5.21x$. Best predicted value: 1772 mm. The best predicted height is close to the actual height.
25. $\hat{y} = 0.923 + 0.00665x$. Best predicted value: $\bar{y} = 57$ points. The best predicted value isn't close to the actual value of 37 points.
27. $\hat{y} = 16.5 + 0.00282x$. Best predicted value: 15.1 fatalities per 100,000 population. Common sense suggests that the prediction doesn't make much sense.
29. $\hat{y} = 0.174 + 0.116x$. Best predicted value: $2.49. (Unlike Exercise 15, this larger data set results in a significant linear correlation, so the predicted value is not $\bar{y}$.) The best predicted value isn't very close to the actual tip of $4.55.
31. $\hat{y} = 5.95 + 2.86x$. Best predicted value: $14.80 (or $14.82). The best predicted value is close to the actual fare of $15.30.
33. a. 6.784, 4.802, -0.300, -1.598, -0.248, -2.420, 0.364, 1.670, -7.470
    b. 137.862
    c. Using $\hat{y} = -10.0 + 0.200x$, the sum of squares of the residuals is 535.560, which is larger than 137.862, which is the sum of squares of the residuals for the regression line.

Section 10-3

1. The value of $s_e = 16.27555$ cm is the standard error of estimate, which is a measure of the differences between the observed weights and the weights predicted from the regression equation. It is a measure of the variation of the sample points about the regression line.
3. The coefficient of determination is $r^2 = 0.155$. We know that 15.5% of the variation in weight is explained by the linear correlation between height and weight, and 84.5% of the variation in weight is explained by other factors and/or random variation.
5. $r^2 = 0.089$. 8.9% of the variation in tips is explained by the linear correlation between times and tips, and 91.1% of the variation in tips is explained by other factors and/or random variation.
7. $r^2 = 0.972$. 97.2% of the variation in fares is explained by the linear correlation between distances and fares, and 2.8% of the variation in fares is explained by other factors and/or random variation.
9. $r = -0.788$. Critical values: $r = \pm 0.576$ assuming a 0.05 significance level. There is sufficient evidence to support the claim of a linear correlation between weights of large cars and the highway fuel consumption amounts.
11. 29.0 mi/gal
13. 27.9 mi/gal $< y <$ 37.7 mi/gal
15. 24.2 mi/gal $< y <$ 36.9 mi/gal
17. a. 10,606.59
    b. 6883577
    c. 78.0°F $< y <$ 80.4°F
19. a. 352.7228
    b. 109.3722
    c. 71.09°F $< y <$ 88.71°F
21. 76.1 million tickets $< \bar{y} <$ 120 million tickets

Section 10-4

1. The response variable is Speed (the mean speed of the winner) and the predictor variables are Distance, the number of Stages, and the number of Finishers.
3. The unadjusted R^2 increases (or remains the same) as more variables are included, but the adjusted R^2 is adjusted for the number of variables and sample size. The unadjusted R^2 incorrectly suggests that the best multiple regression equation is obtained by including all of the available variables, but by taking into account the sample size and number of predictor variables, the adjusted R^2 is much more helpful in weeding out variables that should not be included.
5. Son = 18.0 + 0.504 Father+0.277 Mother
7. P-value less than 0.0001 is low, but the values of $R^2(0.3649)$ and adjusted $R^2(0.3552)$ are not high. Although the multiple regression equation fits the sample data best, it is not a good fit, so it should not be used for predicting the height of a son based on the height of his father and the height of his mother.
9. The weight of discarded paper, because it has the best combination of small P-value (0.000) and highest adjusted R^2 (0.411).
11. PLAS = $-0.170 + 0.290$ METAL + 0.122 PAPER + 0.0777 GLASS. That equation has a low P-value of 0.000 and its adjusted R^2 value of 0.540 is the largest and it is substantially higher than any of the other values of adjusted R^2.

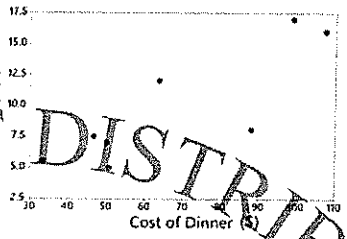
13. The best regression equation is HWY = 58.9 − 0.00749 Weight. The three different possible regression equations all have a P-value of 0.000. Given that the single predictor variable of Weight yields an adjusted R^2 of 0.787 that is only slightly less than the adjusted R^2 of 0.791 obtained by using the two predictor variables of Weight and Displacement, it is better to use the single predictor variable instead of two predictor variables. (The single predictor variable of Displacement has an adjusted R^2 of 0.506.) Because the adjusted R^2 of 0.787 isn't very close to 1, it is likely that predicted values will not be very accurate.
15. The best regression equation is $\hat{y} = 109 - 0.00670x_1$, where x_1 represents volume. It is best because it has the highest adjusted R^2 value of −0.0513 and the lowest P-value of 0.791. The three regression equations all have adjusted values of R^2 that are very close to 0, so none of them are good for predicting IQ. It does not appear that people with larger brains have higher IQ scores.
17. For H_0: $\beta_1 = 0$, the test statistic is $t = 10.814$, the P-value is less than 0.0001, so reject H_0 and conclude that the regression coefficient of $b_1 = 0.769$ should be kept. For H_0: $\beta_2 = 0$, the test statistic is $t = 29.856$, the P-value is less than 0.0001, so reject H_0 and conclude that the regression coefficient of $b_2 = 1.01$ should be kept. It appears that the regression equation should include both independent variables of height and waist circumference.
19. $\hat{y} = 3.06 + 82.4x_1 + 2.91x_2$, where x_1 represents sex and x_2 represents age. Female: 61 lb; male: 144 lb. The sex of the bear does appear to have an effect on its weight. The regression equation indicates that the predicted weight of a male bear is about 82 lb more than the predicted weight of a female bear with other characteristics being the same.

Section 10-5

1. $y = x^2$. The quadratic model describes the relationship, and $R^2 = 1$.
3. 2.7% of the variation in Super Bowl points can be explained by the exponential model that relates the variable of year and the variable of points scored. Because such a small percentage of the variation is explained by the model, the model is not very useful.
5. The quadratic and power models both yield the same result: $d = 0.8t^2$
7. Exponential: $y = 1000(1.01^x)$
9. Quadratic: $y = 0.000154x^2 + 0.0799x + 6.06$, where x is the year with 2000 coded as 1, and y is the world population in billions.
11. Logarithmic: $y = 3.22 + 0.293 \ln x$
13. Quadratic: $y = 84.0x^2 - 953x + 13,282$. (Result is based on the year 2000 coded as 1.) Using the rounded coefficients, the projected value for the last year is 25,506.0, which isn't too far from the actual value of 26,828.4. Because $R^2 = 0.925$ for the quadratic model, which is high, predicted values are likely to be reasonably accurate, but we should remember that stock market values can be dramatically affected by events that cannot be foreseen by our most creative minds.
15. Power: $y = 7.89(x^{-0.371})$, where x is the depth and y is the magnitude. The predicted magnitude is 7.82, which is far from the actual magnitude of 7.10. Because $R^2 = 0.613$ for the power model, which isn't very high, predicted values are not likely to be very accurate.
17. a. Exponential: $y = 2^{(x-1)}$ [or $y = (0.629961)(1.587401)^x$ for an initial value of 1 that doubles every 1.5 years].
    b. Exponential: $y = (1.99724152)(1.419450033)^x$, where 1971 is coded as 1.
    c. Moore's law does appear to be working reasonably well. With $R^2 = 0.991$, the model appears to be very good.

Chapter 10: Quick Quiz

1. The points appear to approximate a straight-line pattern that rises from left to right.

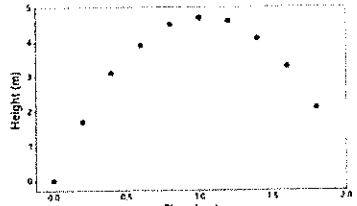


2. Conclude that there is sufficient evidence to support the claim of a linear correlation between amount of the dinner and the amount of the tip.
3. $r = 1$
4. $\hat{y} = 0 + 0.20x$ (or simply $y = 0.20x$)
5. None of the given values change when the variables are switched.
6. The value of r does not change if all values of one of the variables are multiplied by the same constant.
7. Because r must be between −1 and 1 inclusive, the value of 1.200 is the result of an error in the calculation of r.
8. The best predicted tip is $12.26. It was found by substituting $64.62 for x in the regression equation.
9. The best predicted tip is $9.76. Because there is not sufficient evidence to support the claim of a linear correlation between the cost of dinner and the tip, the best predicted tip is found by computing the mean of the eight sample tips.
10. Because $r^2 = 0.716$, it follows that 0.716 (or 71.6%) of the variation in tips is explained by the linear relationship between amounts of dinner and amounts of tips. It then follows that 0.284 (or 28.4%) of the variation in tips is not explained by the linear relationship between amounts of dinner and amounts of tips.

Chapter 10: Review Exercises

1. $r = 0.445$. P-value: 0.318 (Table: > 0.05). Critical values: ± 0.754. There is not sufficient evidence to support the claim that there is a linear correlation between size and revenue. It does not appear that a casino can increase its revenue by enlarging its size.
2. a. $\hat{y} = 63.9 + 0.443x$
   b. Best predicted value of revenue: $\bar{y} = 134.7$ million dollars. Because the predicted amount of revenue is 134.7 million dollars for *any* casino size, the prediction is not likely to be accurate.
3. a. $r = 0.450$.
   b. With P-value = 0.192 (Table: > 0.05) and critical values: $r = \pm 0.632$ (assuming a 0.05 significance level), there is not sufficient evidence to support the claim that there is a linear correlation between time and weight.

c. Although there is no *linear* correlation between time and height, the scatterplot shows a very distinct pattern revealing that time and height are associated by some function that is not linear. (The scatterplot appears to depict a parabola. The quadratic regression equation is $y = -4.44x^2 + 9.13x + 0.0482$.)



4. a. NICOTINE = −0.443 + 0.0968 TAR − 0.0262 CO, or $\hat{y} = -0.443 + 0.0968x_1 - 0.0262x_2$.
b. $R^2 = 0.936$; adjusted $R^2 = 0.910$; P-value = 0.001.
c. With high values of R^2 and adjusted R^2 and a small P-value of 0.001, it appears that the regression equation can be used to predict the amount of nicotine given the amounts of tar and carbon monoxide.
d. The predicted value is 1.39 mg or 1.4 mg rounded, which is close to the actual value of 1.3 mg of nicotine.

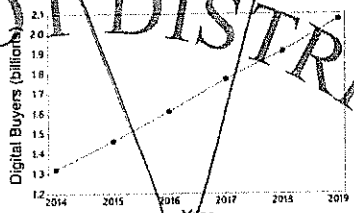
Chapter 10: Cumulative Review Exercises

1. a. Is there a difference between the mean IQ score of airline passengers and the mean IQ score of police officers?
b. Test for a difference between the means of two independent populations using the methods of Section 9-2.
c. H_0: $\mu_1 = \mu_2$. H_1: $\mu_1 \neq \mu_2$. Test statistic: $t = -1.557$. P-value = 0.1516 (Table: P-value > 0.05). Critical values (assuming a 0.05 significance level): $t = \pm 2.239$ (Table: ± 2.262). Fail to reject H_0. There is not sufficient evidence to support the claim that there is a difference between the mean IQ score of airline passengers and the mean IQ of police officers. (This 95% confidence interval could also be used: $-20.0 < \mu_1 - \mu_2 < 3.39$. Because the confidence interval includes 0, there is not sufficient evidence to support the claim that there is a difference between the mean IQ score of airline passengers and the mean IQ of police officers.)

2. a. Was the training course effective in raising the IQ scores? That is, do the "before − after" differences have a mean that is less than 0, showing that the course is effective in raising IQ scores?
b. Use the methods of Section 9-3 to test the claim that the mean of the "before − after" differences is less than 0, showing that the course is effective with larger "after" scores.
c. H_0: $\mu_d = 0$. H_1: $\mu_d < 0$. Test statistic: $t = -1.541$. P-value = 0.0789 (Table: P-value > 0.05). Critical value (assuming a 0.05 significance level): $t = -1.833$. Fail to reject H_0. There is not sufficient evidence to support the claim that the course is effective with higher "after" scores. (This 90% confidence interval could also be used: $-18.0 < \mu_d < 1.56$). Because the confidence interval includes 0, there is not sufficient evidence to support the claim that the course is effective with higher "after" scores.

3. a. For professional horse jockeys, is there a correlation between weight and number of top three race finishes?
b. Use the methods of Section 10-1 to test for a linear correlation.
c. $r = -0.060$. P-value = 0.869 (Table: > 0.05). Critical values (assuming a 0.05 significance level): $r = \pm 0.632$. There is not sufficient evidence to support the claim that there is a linear correlation between weight and the number of top three race finishes.

4. a. Because the table lists time series data, a key question is this: What is the trend of the data over time?
b. Use the methods of Section 2-3 to construct a time series graph that would reveal a trend of the data over time.
c. A time series graph clearly shows that there is a distinct trend of steadily increasing numbers of digital buyers over time. Businesses should ensure they can market and sell their goods and services online.



5. a. What is an estimate of the proportion of all adults who have wireless earbuds?
b. Use the methods of Section 7-4 to construct a confidence interval estimate of the proportion of all adults who use wireless earbuds.
c. 95% confidence interval estimate of p: $0.280 < p < 0.320$. With 95% confidence, it is estimated that between 28.0% and 32.0% of all adults have wireless earbuds. (It would also be reasonable to conduct a hypothesis test, such as a test of the claim that fewer than 50% of adults have wireless earbuds. For that test, the test statistic is $z = -17.95$ and the P-value is 0.0000 so there is sufficient evidence to support the claim that fewer than 50% of adults have wireless earbuds.)

6. a. Is Stephen Curry *significantly tall* in the population of adult males?
b. Using the methods of Section 3-3, convert Stephen Curry's height to a z score and use the range rule of thumb to determine whether his height is significantly high.
c. Converting Stephen Curry's height to a z score, we get $z = (x - \mu)/\sigma = (191 - 174.12)/7.10 = 2.38$. Stephen Curry's height is 2.38 standard deviations above the mean, so his height is significantly high.

7. a. Is the mean amount provided by the new device equal to 16 ounces? Is there anything else about the data suggesting that there is a problem with the new device?
b. Explore the sample data to see if there are any undesirable characteristics. Use the methods of Section 8-3 to test the claim that the mean of the amounts is equal to 16 ounces.

Section 4-4

1. The symbol ! is the factorial symbol, which represents the product of decreasing whole numbers, as in $5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$. Five NBA players can stand in line 120 different ways.
3. Because repetition is allowed, numbers are selected *with replacement*, so the combinations rule and the two permutation rules do not apply. The multiplication counting rule can be used to show that the number of possible outcomes is $10 \cdot 10 \cdot 10 = 1000$.
5. 1/10,000 7. 1/190
9. 50,400; 1/50,400 11. 1/5,527,200
13. 1/100,000,000. No, there are far too many different possibilities.
15. 360; 1/360 (RHYTHM)
17. 1/292,201,338 19. 1/109,000
21. Area codes: 792. Phone numbers: 6,272,640,000. Yes. (With a total population of about 400,000,000, there would be roughly 16 phone numbers for every adult and child.)
23. a. 360,360 b. 3003 c. 1/3003
25. a. 1,048,576 b. 184,756 c. 0.176
    d. With a probability of 0.176, the result is common, but it should not happen consistently.
27. 653,837,184,000
29. a. 1/302,575,350
    b. There is a *much* better chance of being struck by lightning.
    c. Probability for the old Mega Millions game: 1/258,890,850. The current Mega Millions game has a substantially lower probability of winning when compared to the old Mega Millions game.
31. There are 62 different possible characters. The alphabet requires 26 characters, and there are 10 digits, so the Morse code system is more than adequate.
33. 12
35. a. 1/177,100 b. $88,550
    c. No, because the jackpot is too small.
37. 2,095,681,645,538 (about 2 trillion)
39. 0.000000112 [from $({}_{20}C_{10})/({}_{80}C_{10})$]

Section 4-5

1. No. The generated numbers between 2 and 12 would be equally likely, but they are not equally likely with actual dice.
3. Yes, it does. Each of the 365 birthdays has the same chance of being selected, and the cards are replaced, so it is possible to select the same birthday more than once.
5. Randomly generate 50 integers, with each integer between 1 and 100. Consider the numbers 1 through 95 to be adults who recognize the brand name of McDonald's, while the numbers 96 through 100 represent adults who do not recognize McDonald's.
7. Randomly generate an integer between 1 and 1000 inclusive. Consider an outcome of 1 through 640 to be a pass that was caught and consider an outcome between 641 and 1000 to be a pass that was not caught.
9. Answers vary, but here is a typical result: Among 100 generated samples, the sample mean of $\bar{x} = 97.49°F$ or lower never occurred, so the conclusions are essentially the same as in Example 1: With the assumption that the mean body temperature is 98.6°F, we have found that the sample mean of 97.49°F is highly unlikely and is significantly low. Because we did get the sample mean of 97.49°F from Data Set 5, we have strong evidence suggesting that the assumed population mean of 98.6°F is likely to be wrong.
11. Sample statistics: $n = 15$, $\bar{x} = 62.7$ seconds, $s = 19.5$ seconds. Generate random samples from a normally distributed population with the assumed mean of 60 seconds, a standard deviation of 19.5 seconds, and a sample size of $n = 15$. Answers vary, but here is a typical result: Among 100 generated samples, the sample mean of $\bar{x} = 62.7$ seconds or higher occurred 33 times, so 62.7 seconds is not significantly high. The sample mean of 62.7 seconds could easily occur with a population mean of 60 seconds, so there isn't strong evidence against 60 seconds as the population mean.
13. With switching, $P(\text{win}) = 2/3$. With sticking, $P(\text{win}) = 1/3$.
15. The reasoning is not correct. The proportion of girls will not increase.

Chapter 4: Quick Quiz

1. 0.2 or 1/5 2. 4/5 or 0.8 3. 1/365
4. 1/100 or 0.01
5. Answer varies, but the probability should be somewhat low, such as 0.02.
6. 0.0680 7. 0.727 8. 0.00831
9. 0.00459 10. 0.0131

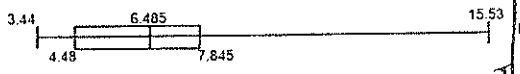
Chapter 4: Review Exercises

1. 0.684, which does not appear to be reasonably close to the proportion of females in the general population. It does not seem that the study subjects were randomly selected from the general population.
2. 0.125 3. 0.730 4. 0.714
5. 0.401 6. 0.0133 7. 0.0134
8. L is the event of randomly selecting one of the study subjects and getting someone who does not write with their left hand. $P(L) = 0.884$.
9. $\overline{M}$ is the event of randomly selecting one of the study subjects and getting someone who is not a male. $P(\overline{M}) = 0.684$.
10. 0.00151. Yes, because the probability of getting three lefties is so small.
11. $(15/65)(14/64)(13/63)(12/62) = 32,760/16,248,960 = 0.00202$. Because that probability is so low, it is very unlikely that the seats were randomly assigned.
12. a. 0.25 b. 0.316
    c. A result of x successes among n trials is a *significantly high* number of successes if the probability of x or more successes is unlikely with a probability of 0.05 or less. That is, x is a significantly high number of successes if $P(x \text{ or more}) \leq 0.05$. (The value 0.05 is not absolutely rigid. Other values, such as 0.01, could be used to distinguish between results that are significant and those that are not significant.)
    d. No, the probability of getting all four people using vision correction is 0.316, which is not unlikely with a small probability such as 0.05 or less. Because the probability of four people using vision correction is so high, that event can easily occur and it is not a significant event.

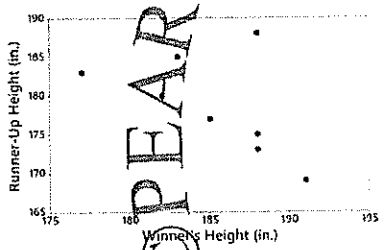
13. a. 1/365 b. 31/365
    c. Answer varies, but it is probably quite small, such as 0.01 or less.
    d. Yes
~~14. 0.0422. No.~~
15. 1/1,832,600. No, the jackpot seems disproportionately small given the probability of winning it. But then, no lotteries are fair.

Chapter 4: Cumulative Review Exercises

1. a. 6.919 rnfl b. 6.485 rnfl ~~c. 9.485 rnfl~~
   d. 12.090 rnfl e. 3.400 rnfl f. 11.558 $rnfl^2$
2. a. 3.44 rnfl, 4.480 rnfl, 6.485 rnfl, 7.845 rnfl, 15.53 rnfl
   b.



   c. The amount of 15.53 rnfl appears to be an outlier.
3. a. 46% b. 0.460 c. Stratified sample
4. a. Convenience sample
   b. If the students at the college are mostly from a surrounding region that includes a large proportion of one ethnic group, the results might not reflect the general population of the United States.
   c. 0.75 d. 0.64
5. Based on the scatterplot, it is reasonable to conclude that there is no association between heights of presidents and the heights of the presidential candidates who were runners-up. It is also reasonable to conclude that there is a very weak association with increasing heights of winners corresponding to decreasing heights of runners-up. (More objective criteria will be introduced in Chapter 10.)



Chapter 5 Answers

Section 5-1

1. The random variable is x, which is the number of unlicensed software packages. The possible values of x are 0, 1, 2, 3, and 4. The values of the random variable x are numerical.
3. $\Sigma P(x) = 0.008 + 0.076 + 0.265 + 0.412 + 0.240 = 1.001$. The sum is not exactly 1 because of the round-off error. The sum is close enough to 1 to satisfy the requirements. Also, the variable x is a numerical random variable and its values are associated with probabilities, and each of the probabilities is between 0 and 1 inclusive, as required. The table does describe a probability distribution.
5. a. Discrete random variable b. Continuous random variable
   c. Discrete random variable d. Not a random variable
   e. Discrete random variable
7. Not a probability distribution because the causes are not values of a numerical random variable.
9. Probability distribution with $\mu = 1.6$, $\sigma = 0.9$.
11. Not a probability distribution because the responses are not values of a numerical random variable.
13. Probability distribution with $\mu = 2.4$, $\sigma = 1.0$.
15. $\mu = 0.4$, $\sigma = 0.6$
17. Significantly high numbers of matches are greater than or equal to $\mu + 2\sigma$, and $\mu + 2\sigma = 0.4 + 2(0.6) = 1.6$ matches. Because 4 matches is greater than or equal to 1.6 matches, it is a significantly high number of matches.
19. a. 0.004 b. 0.004 c. Part (b)
    d. Yes, because the probability of 0.004 is low (less than or equal to 0.05).
21. $\mu = 2.1$, $\sigma = 1.1$
23. Significantly low numbers of drivers who say that they text while driving is less than or equal to $\mu - 2\sigma = 2.1 - 2(1.1) = -0.1$. Because 1 driver is not less than or equal to -0.1, 1 is not a significantly low number of drivers who say that they text while driving.
25. a. 0.344 b. 0.648 c. Part (b)
    d. No, because the probability of 2 or fewer drivers who say that they text while driving is not low (less than or equal to 0.05).
27. Because the probability of 270 or more saying that we should use biometrics is 0.0995, which is not low (less than or equal to 0.05), 270 is not significantly high. Given that 270 is not significantly greater than 50%, there is not sufficient evidence to conclude that the majority of the population says that we should replace passwords with biometric security.
29. a. 1000 b. 1/1000 c. \$499 d. −50¢
    e. The \$1 bet on the pass line in craps is better because its expected value of −1.4¢ is much greater than the expected value of −50¢ for the Florida Pick 3 lottery.
31. a. −39¢
    b. The bet on the number 27 is better because its expected value of −26¢ is greater than the expected value of −39¢ for the other bet.

Section 5-2

1. The given calculation assumes that the first two speaking characters are females and the last three are not females, but there are other arrangements consisting of two females and three males. The probabilities corresponding to those other arrangements should also be included in the result.
3. Because the 50 selections are made without replacement, they are dependent, not independent. Based on the 5% guideline for cumbersome calculations, the 50 selections can be treated as being independent. (The 50 selections constitute 3.33% of the population of 1500 speaking characters, and 3.33% is not more than 5% of the population.)
5. Not binomial. Each of the ages has more than two possible outcomes.

Formulas by Mario F. Triola

Ch. 3: Descriptive Statistics

$\bar{x} = \frac{\Sigma x}{n}$ Mean

$\bar{x} = \frac{\Sigma(f \cdot x)}{\Sigma f}$ Mean (frequency table)

$s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$ Standard deviation

$s = \sqrt{\frac{n(\Sigma x^2) - (\Sigma x)^2}{n(n-1)}}$ Standard deviation (shortcut)

$s = \sqrt{\frac{n[\Sigma(f \cdot x^2)] - [\Sigma(f \cdot x)]^2}{n(n-1)}}$ Standard deviation (frequency table)

variance $= s^2$

Ch. 4: Probability

$P(A \text{ or } B) = P(A) + P(B)$ if A, B are mutually exclusive

$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$ if A, B are not mutually exclusive

$P(A \text{ and } B) = P(A) \cdot P(B)$ if A, B are independent

$P(A \text{ and } B) = P(A) \cdot P(B|A)$ if A, B are dependent

$P(\overline{A}) = 1 - P(A)$ Rule of complements

${}_nP_r = \frac{n!}{(n-r)!}$ Permutations (no elements alike)

$\frac{n!}{n_1!\, n_2!\, \ldots\, n_k!}$ Permutations (n_1 alike, . . .)

${}_nC_r = \frac{n!}{(n-r)!\, r!}$ Combinations

Ch. 5: Probability Distributions

$\mu = \Sigma[x \cdot P(x)]$ Mean (prob. dist.)

$\sigma = \sqrt{\Sigma[x^2 \cdot P(x)] - \mu^2}$ Standard deviation (prob. dist.)

$P(x) = \frac{n!}{(n-x)!\, x!} \cdot p^x \cdot q^{n-x}$ Binomial probability

$\mu = n \cdot p$ Mean (binomial)

$\sigma^2 = n \cdot p \cdot q$ Variance (binomial)

$\sigma = \sqrt{n \cdot p \cdot q}$ Standard deviation (binomial)

$P(x) = \frac{\mu^x \cdot e^{-\mu}}{x!}$ Poisson distribution where $e = 2.71828$

Ch. 6: Normal Distribution

$z = \frac{x - \mu}{\sigma}$ or $\frac{x - \bar{x}}{s}$ Standard score

$\mu_{\bar{x}} = \mu$ Central limit theorem

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ Central limit theorem (Standard error)

Ch. 7: Confidence Intervals (one population)

$\hat{p} - E < p < \hat{p} + E$ Proportion

where $E = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$

$\bar{x} - E < \mu < \bar{x} + E$ Mean

where $E = t_{\alpha/2}\frac{s}{\sqrt{n}}$ (σ unknown)

or $E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ (σ known)

$\frac{(n-1)s^2}{\chi^2_R} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_L}$ Variance

Ch. 7: Sample Size Determination

$n = \frac{[z_{\alpha/2}]^2 0.25}{E^2}$ Proportion

$n = \frac{[z_{\alpha/2}]^2 \hat{p}\hat{q}}{E^2}$ Proportion ($\hat{p}$ and $\hat{q}$ are known)

$n = \left[\frac{z_{\alpha/2}\sigma}{E}\right]^2$ Mean

Ch. 8: Test Statistics (one population)

$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$ Proportion—one population

$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$ Mean—one population (σ unknown)

$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$ Mean—one population (σ known)

$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$ Standard deviation or variance—one population