PART II REVIEW

Data analysis is the art of describing data using graphs and numerical summaries. The purpose of data analysis is to help us see and understand the most important features of a set of data. Chapter 10 commented on basic graphs, especially pie charts, bar graphs, and line graphs. Chapters 11, 12, and 13 showed how data analysis works by presenting statistical ideas and tools for describing the distribution of one variable. Figure II.1 organizes the big ideas. We plot our data, then describe their center and spread using either the mean and standard deviation or the five-number summary. The last step, which makes sense only for some data, is to summarize the data in compact form by using a Normal curve as a model for the overall pattern. The question marks at the last two stages remind us that the usefulness of numerical summaries and Normal distributions depends on what we find when we examine graphs of our data. No short summary does justice to irregular shapes or to data with several distinct clusters.

Chapters 14 and 15 applied the same ideas to relationships between two quantitative variables. Figure II.2 retraces the big ideas from Figure II.1, with details that fit the new setting. We always begin by making graphs of our data. In the case of a scatterplot, we have learned a numerical summary only for data that show a roughly straight-line pattern on the scatterplot. The summary is then the means and standard deviations of the two variables and their correlation. A regression line drawn on the plot

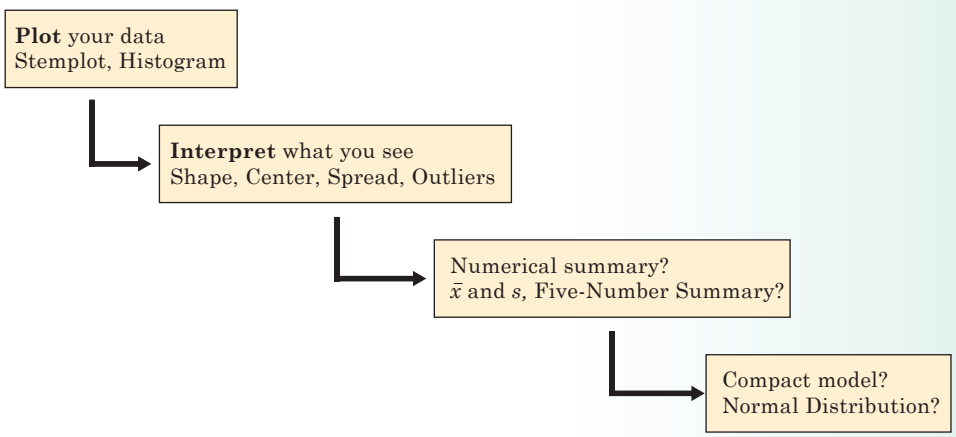

Figure II.1

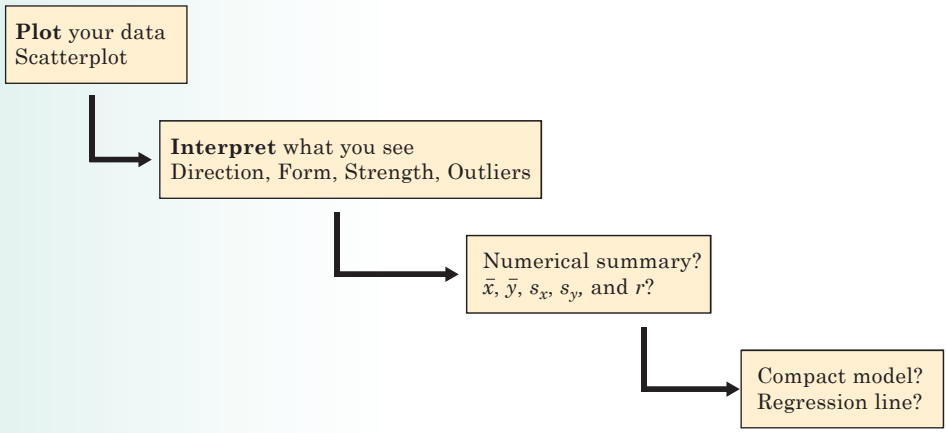

Figure II.2

gives us a compact model of the overall pattern that we can use for prediction. Once again there are question marks at the last two stages to remind us that correlation and regression describe only straight-line relationships.

Relationships often raise the question of causation. We know that evidence from randomized comparative experiments is the "gold standard" for deciding that one variable causes changes in another variable. Chapter 15 reminded us in more detail that strong associations can appear in data even when there is no direct causation. We must always think about the possible effects of variables lurking in the background. In Chapter 16, we met a new kind of description, index numbers, with the Consumer Price Index as the leading example. Chapter 16 also discussed government statistical offices, a quiet but important part of the statistical world.

PART II SUMMARY

Here are the most important skills you should have acquired after reading Chapters 10 through 16.

A. DISPLAYING DISTRIBUTIONS

1. Recognize categorical and quantitative variables.

2. Recognize when a pie chart can and cannot be used.

3. Make a bar graph of the distribution of a categorical variable, or in general to compare related quantities.

4. Interpret pie charts and bar graphs.

5. Make a line graph of a quantitative variable over time.

6. Recognize patterns such as trends and seasonal variation in line graphs.

7. Be aware of graphical abuses, especially pictograms and distorted scales in line graphs.

8. Make a histogram of the distribution of a quantitative variable.

9. Make a stemplot of the distribution of a small set of observations. Round data as needed to make an effective stemplot.

B. DESCRIBING DISTRIBUTIONS (QUANTITATIVE VARIABLE)

1. Look for the overall pattern of a histogram or stemplot and for major deviations from the pattern.

2. Assess from a histogram or stemplot whether the shape of a distribution is roughly symmetric, distinctly skewed, or neither. Assess whether the distribution has one or more major peaks.

3. Describe the overall pattern by giving numerical measures of center and spread in addition to a verbal description of shape.

4. Decide which measures of center and spread are more appropriate: the mean and standard deviation (especially for symmetric distributions) or the five-number summary (especially for skewed distributions).

5. Recognize outliers and give plausible explanations for them.

C. NUMERICAL SUMMARIES OF DISTRIBUTIONS

1. Find the median M and the quartiles Q_1 and Q_3 for a set of observations.

2. Give the five-number summary and draw a boxplot; assess center, spread, symmetry, and skewness from a boxplot.

3. Find the mean $\bar{x}$ and (using a calculator) the standard deviation s for a small set of observations.

4. Understand that the median is less affected by extreme observations than the mean. Recognize that skewness in a distribution moves the mean away from the median toward the long tail.

5. Know the basic properties of the standard deviation: $s \geq 0$ always; $s = 0$ only when all observations are identical and increases as the spread increases; s has the same units as the original measurements; s is greatly increased by outliers or skewness.

D. NORMAL DISTRIBUTIONS

1. Interpret a density curve as a description of the distribution of a quantitative variable.

2. Recognize the shape of Normal curves, and estimate by eye both the mean and the standard deviation from such a curve.

3. Use the 68–95–99.7 rule and symmetry to state what percentage of the observations from a Normal distribution fall between two points when the

points lie at the mean or one, two, or three standard deviations on either side of the mean.

4. Find and interpret the standard score of an observation.

5. (Optional) Use Table B to find the percentile of a value from any Normal distribution and the value that corresponds to a given percentile.

E. SCATTERPLOTS AND CORRELATION

1. Make a scatterplot to display the relationship between two quantitative variables measured on the same subjects. Place the explanatory variable (if any) on the horizontal scale of the plot.

2. Describe the direction, form, and strength of the overall pattern of a scatterplot. In particular, recognize positive or negative association and straight-line patterns. Recognize outliers in a scatterplot.

3. Judge whether it is appropriate to use correlation to describe the relationship between two quantitative variables. Use a calculator to find the correlation r.

4. Know the basic properties of correlation: r measures the strength and direction of only straight-line relationships; r is always a number between -1 and 1; $r = \pm 1$ only for perfect straight-line relations; r moves away from 0 toward ± 1 as the straight-line relation gets stronger.

F. REGRESSION LINES

1. Explain what the slope b and the intercept a mean in the equation $y = a + bx$ of a straight line.

2. Draw a graph of the straight line when you are given its equation.

3. Use a regression line, given on a graph or as an equation, to predict y for a given x. Recognize the danger of prediction outside the range of the available data.

4. Use r^2, the square of the correlation, to describe how much of the variation in one variable can be accounted for by a straight-line relationship with another variable.

G. STATISTICS AND CAUSATION

1. Understand that an observed association can be due to direct causation, common response, or confounding.

2. Give plausible explanations for an observed association between two variables: direct cause and effect, the influence of lurking variables, or both.

3. Assess the strength of statistical evidence for a claim of causation, especially when experiments are not possible.

H. THE CONSUMER PRICE INDEX AND RELATED TOPICS

1. Calculate and interpret index numbers.

2. Calculate a fixed market basket price index for a small market basket.

3. Use the CPI to compare the buying power of dollar amounts from different years. Explain phrases such as "real income."

PART II REVIEW EXERCISES

Review exercises are short and straightforward exercises that help you solidify the basic ideas and skills in each part of this book. We have provided "hints" that indicate where you can find the relevant material for the odd-numbered problems.

II.1. Poverty in the states. Table II.1 gives the percentages of people living below the poverty line in the 26 states east of the Mississippi River. Make a stemplot of these data. Is the distribution roughly symmetric, skewed to the right, or skewed to the left? Which states (if any) are outliers? (*Hint*: See page 253.)

II.2. Quarterbacks. Table II.2 gives the total passing yards for National Football League starting quarterbacks during the 2014 season. (These are the quarterbacks with the most passing yards on each team.) Make a histogram of these data. Does the distribution have a clear shape: roughly symmetric, clearly skewed to the left, clearly skewed to the right, or none of these? Which quarterbacks (if any) are outliers?

II.3. Poverty in the states. Give the five-number summary for the data on poverty from Table II.1. (*Hint*: See page 272.)

TABLE II.1 Percentages of state residents living in poverty, 2012–2013 two-year average

State	Percent	State	Percent	State	Percent
Alabama	16.4	Connecticut	10.8	Delaware	13.7
Florida	15.1	Georgia	17.2	Illinois	12.9
Indiana	13.4	Kentucky	18.9	Maine	12.5
Maryland	10.1	Massachusetts	11.6	Michigan	14.1
Mississippi	22.2	New Hampshire	8.6	New Jersey	10.2
New York	15.9	North Carolina	17.9	Ohio	14.5
Pennsylvania	13.1	Rhode Island	13.6	South Carolina	16.3
Tennessee	18.4	Vermont	10.0	Virginia	10.5
West Virginia	17.0	Wisconsin	11.2		

Source: www.census.gov/hhes/www/poverty/data/index.html

TABLE II.2 Passing yards for NFL quarterbacks in 2014

Quarterback	Yards	Quarterback	Yards
Blake Bortels	2908	Josh McCown	2206
Tom Brady	4109	Zach Mettenberger	1412
Drew Brees	4952	Cam Newton	3127
Teddy Bridgewater	2919	Kyle Orton	3018
Derek Carr	3270	Philip Rivers	4286
Kirk Cousins	1710	Aaron Rodgers	4381
Jay Cutler	3812	Ben Roethlisberger	4952
Andy Dalton	3398	Tony Romo	3705
Austin Davis	2001	Matt Ryan	4694
Ryan Fitzpatrick	2483	Mark Sanchez	2418
Joe Flacco	3986	Alex Smith	3265
Brian Hoyer	3326	Geno Smith	2525
Colin Kaepernick	3369	Matthew Stafford	4257
Andrew Luck	4761	Drew Stanton	1711
Eli Manning	4410	Ryan Tannehill	4045
Peyton Manning	4727	Russell Wilson	3475

Source: www.pro-football-reference.com/years/2014/passing.htm.

II.4. Quarterbacks. Give the five-number summary for the data on passing yards for NFL quarterbacks from Table II.2.

II.5. Poverty in the states. Find the mean percentage of state residents living in poverty from the data in Table II.1. If we removed Mississippi from the data, would the mean increase or decrease? Why? Find the mean for the 25 remaining states to verify your answer. (*Hint*: See page 277.)

II.6. Big heads? The army reports that the distribution of head circumference among male soldiers is approximately Normal with mean 22.8 inches and standard deviation 1.1 inches. Use the 68–95–99.7 rule to answer these questions.

(a) Between what values do the middle 95% of head circumferences fall?

(b) What percentage of soldiers have head circumferences greater than 23.9 inches?

II.7. SAT scores. The scale for SAT exam scores is set so that the distribution of scores is approximately Normal with mean 500 and standard deviation 100. Answer these questions without using a table.

(a) What is the median SAT score? (*Hint*: See page ~~301~~.)

(b) You run a tutoring service for students who score between 400 and 600 and hope to attract many students. What percentage of SAT scores are between 400 and 600? (*Hint*: See page ~~302~~.)

II.8. Explaining correlation. You have data on the yearly wine consumption (liters of alcohol from drinking wine per person) and yearly deaths from cirrhosis of the liver for several developed countries.

Say as specifically as you can what the correlation r between yearly wine consumption and yearly deaths from cirrhosis of the liver measures.

II.9. Data on snakes. For a biology project, you measure the length (inches) and weight (ounces) of 12 snakes of the same variety. What units of measurement do each of the following have?

(a) The mean length of the snakes. (*Hint*: See page 277.)

(b) The first quartile of the snake lengths. (*Hint*: See page 270.)

(c) The standard deviation of the snake lengths. (*Hint*: See page 277.)

(d) The correlation between length and snake weight. (*Hint*: See page 325.)

II.10. More data on snakes. For a biology project, you measure the length (inches) and weight (ounces) of 12 snakes of the same variety.

(a) Explain why you expect the correlation between length and weight to be positive.

(b) The mean length turns out to be 20.8 inches. What is the mean length in centimeters? (There are 2.54 centimeters in an inch.)

(c) The correlation between length and weight turns out to be $r = 0.6$. If you were to measure length in centimeters instead of inches, what would be the new value of r?

Figure II.3 plots the average brain weight in grams versus average body weight in kilograms for many species of mammals. There are many small mammals whose points at the lower left overlap. Exercises II.11 through II.16 are based on this scatterplot.

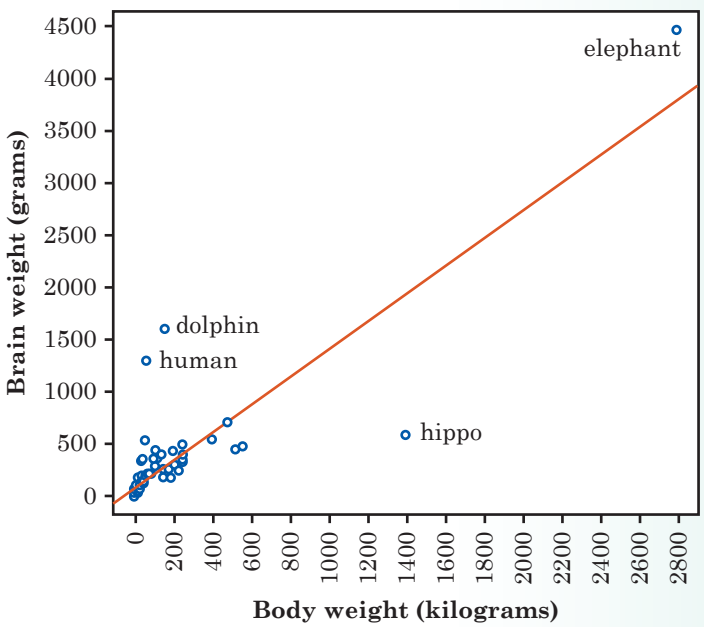

Figure II.3 Scatterplot of the average brain weight (grams) against the average body weight (kilograms) for 96 species of mammals, Exercises II.11 through II.16.

II.11. Dolphins and hippos. The points for the dolphin and hippopotamus are labeled in Figure II.3. Read from the graph the approximate body weight and brain weight for these two species. (*Hint*: See page 318.)

II.12. Dolphins and hippos. One reaction to this scatterplot is "Dolphins are smart; hippos are dumb." What feature of the plot lies behind this reaction?

II.13. Outliers. African elephants are much larger than any other mammal in the data set but lie roughly in the overall straight-line pattern. Dolphins, humans, and hippos lie outside the overall pattern. The correlation between body weight and brain weight for the entire data set is $r = 0.86$.

(a) If we removed elephants, would this correlation increase, decrease, or not change much? Explain your answer. (*Hint*: See page 326.)

(b) If we removed dolphins, hippos, and humans, would this correlation increase, decrease, or not change much? Explain your answer. (*Hint*: See page 326.)

II.14. Brain and body. The correlation between body weight and brain weight is $r = 0.86$. How well does body weight explain brain weight for mammals? Compute r^2 to answer this question, and briefly explain what r^2 tells us.

II.15. Prediction. The line on the scatterplot in Figure II.3 is the least-squares regression line for predicting brain weight from body weight. Suppose that a new mammal species with body weight 600 kilograms is discovered hidden in the rain forest. Predict the brain weight for this species. (*Hint*: See page 340.)

II.16. Slope. The line on the scatter-plot in Figure II.3 is the least-squares regression line for predicting brain weight from body weight. The slope of this line is one of the numbers below. Which number is the slope? Why?

(a) $b = 0.5$.

(b) $b = 1.3$.

(c) $b = 3.2$.

From Rex Boggs in Australia comes an unusual data set: before showering in the morning, he weighed the bar of soap in his shower stall. The weight goes down as the soap is used. The data appear in Table II.3 (weights in grams). Notice that Mr. Boggs forgot

TABLE II.3 Weight (grams) of a bar of soap used to shower

Day	Weight	Day	Weight	Day	Weight
1	124	8	84	16	27
2	121	9	78	18	16
5	103	10	71	19	12
6	96	12	58	20	8
7	90	13	50	21	6

Source: Rex Boggs.

to weigh the soap on some days. Exercises II.17, II.18, and II.19 are based on the soap data set.

II.17. Scatterplot. Plot the weight of the bar of soap against day. Is the overall pattern roughly straight-line? Based on your scatterplot, is the correlation between day and weight close to 1, positive but not close to 1, close to 0, negative but not close to -1, or close to -1? Explain your answer. (*Hint*: See page 325.)

II.18. Regression. The equation for the least-squares regression line for the data in Table II.3 is

$$\text{weight} = 133.2 - 6.31 \times \text{day}$$

(a) Explain carefully what the slope $b = -6.31$ tells us about how fast the soap lost weight.

(b) Mr. Boggs did not measure the weight of the soap on Day 4. Use the regression equation to predict that weight.

(c) Draw the regression line on your scatterplot from the previous exercise.

II.19. Prediction? Use the regression equation in the previous exercise to predict the weight of the soap after 30 days. Why is it clear that your answer makes no sense? What's wrong with using the regression line to predict weight after 30 days? (*Hint*: See page 345.)

II.20. Keeping up with the Joneses. The Jones family had a household income of $30,000 in 1980, when the average CPI (1982–84 = 100) was 82.4. The average CPI for 2014 was 236.7. How much must the Joneses earn in 2014 to have the same buying power they had in 1980?

II.21. Affording a Mercedes. A Mercedes-Benz 190 cost $24,000 in 1981, when the average CPI (1982–84 = 100) was 90.9. The average CPI for 2014 was 236.7. How many 2014 dollars must you earn to have the same buying power as $24,000 had in 1981? (*Hint*: See page 374.)

II.22. Affording a Steinway. A Steinway concert grand piano cost $13,500 in 1976. A similar Steinway cost $163,600 in August 2015. Has the cost of the piano gone up or down in real terms? Using Table 16.1 and the fact that the August 2015 CPI was 238.7, give a calculation to justify your answer.

II.23. The price of gold. Some people recommend that investors buy gold "to protect against inflation." Here are the prices of an ounce of gold at the end of the year for the years between 1985 and 2013. Using Table 16.1, make a graph that shows how the price of gold changed in real terms over this period. Would an investment in gold have protected against inflation by holding its value in real terms?

Year:	1985	1987	1989	1991	1993
Gold price:	$327	$484	$399	$353	$392

Year:	1995	1997	1999	2001	2003
Gold price:	$387	$290	$290	$277	$416

Year:	2005	2007	2009	2011	2013
Gold price:	$513	$834	$1088	$1531	$1204

(*Hint*: See page 372.)

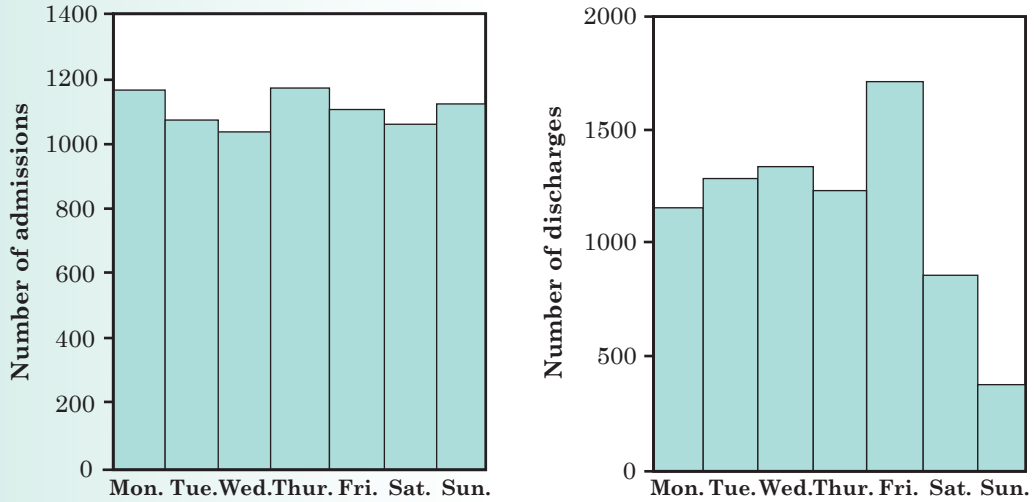

Figure II.4 Bar graphs of the number of heart attack victims admitted and discharged from hospitals in Ontario, Canada, on each day of the week, Exercise II.24.

II.24. Never on Sunday? The Canadian province of Ontario carries out statistical studies to monitor how Canada's national health care system is working in the province. The bar graphs in Figure II.4 come from a study of admissions and discharges from community hospitals in Ontario. They show the number of heart attack patients admitted and discharged on each day of the week during a two-year period.

(a) Explain why you expect the number of patients admitted with heart attacks to be roughly the same for all days of the week. Do the data show that this is true?

(b) Describe how the distribution of the day on which patients are discharged from the hospital differs from that of the day on which they are admitted. What do you think explains the difference?

II.25. Drive time. Professor Moore, who lives a few miles outside a college town, records the time he takes to drive to the college each morning. Here are the times (in minutes) for 42 consecutive weekdays, with the dates in order along the rows:

8.25	7.83	8.30	8.42	8.50	8.67	8.17
9.00	9.00	8.17	7.92	9.00	8.50	9.00
7.75	7.92	8.00	8.08	8.42	8.75	8.08
9.75	8.33	7.83	7.92	8.58	7.83	8.42
7.75	7.42	6.75	7.42	8.50	8.67	10.17
8.75	8.58	8.67	9.17	9.08	8.83	8.67

(a) Make a histogram of these drive times. Is the distribution roughly symmetric, clearly skewed, or neither? Are there any clear outliers? (*Hint*: See pages 249 and 251.)

(b) Make a line graph of the drive times. (Label the horizontal axis in days, 1 to 42.) The plot shows no clear trend, but it does show one

unusually low drive time and two unusually high drive times. Circle these observations on your plot. (*Hint*: See page 223.)

II.26. Drive time outliers. In the previous exercise, there are three outliers in Professor Moore's drive times to work. All three can be explained. The low time is the day after Thanksgiving (no traffic on campus). The two high times reflect delays due to an accident and icy roads. Remove these three observations. To summarize normal drive times, use a calculator to find the mean $\bar{x}$ and standard deviation s of the remaining 39 times.

II.27. House prices. An April 15, 2014, article in the *Los Angeles Times* reported that the median housing price in Southern California was about $400,000. Would the mean housing price be higher, about the same, or lower? Why? (*Hint*: See page 281.)

II.28. The 2012 election. Barack Obama was elected president in 2012 with 51.1% of the popular vote. His Republican opponent, Mitt Romney, received 47.2% of the vote, with minor candidates taking the remaining votes. Table II.4 gives the percentage of the popular vote won by President Obama in each state. Describe these data with a graph, a numerical summary, and a brief verbal description.

II.29. Statistics for investing. Joe's retirement plan invests in stocks through an "index fund" that follows the behavior of the

TABLE II.4 Percentage of votes for President Obama in 2012

State	Percent	State	Percent	State	Percent
Alabama	38.4	Louisiana	40.6	Ohio	50.7
Alaska	40.8	Maine	56.3	Oklahoma	33.2
Arizona	44.6	Maryland	62.0	Oregon	54.2
Arkansas	36.9	Massachusetts	60.7	Pennsylvania	52.0
California	60.2	Michigan	54.2	Rhode Island	62.7
Colorado	51.5	Minnesota	52.3	South Carolina	44.1
Connecticut	58.1	Mississippi	43.8	South Dakota	39.9
Delaware	58.6	Missouri	44.4	Tennessee	39.1
Florida	50.0	Montana	41.7	Texas	41.4
Georgia	45.5	Nebraska	38.0	Utah	24.8
Hawaii	70.6	Nevada	52.4	Vermont	66.6
Idaho	32.6	New Hampshire	52.0	Virginia	51.2
Illinois	57.6	New Jersey	58.4	Washington	56.2
Indiana	43.9	New Mexico	53.0	West Virginia	35.5
Iowa	52.0	New York	63.4	Wisconsin	52.8
Kansas	38.0	North Carolina	48.4	Wyoming	27.8
Kentucky	37.8	North Dakota	38.7		

Source: uselectionatlas.org/.

stock market as a whole, as measured by the Standard & Poor's 500 index. Joe wants to buy a mutual fund that does not track the index closely. He reads that monthly returns from Fidelity Technology Fund have correlation $r = 0.77$ with the S&P 500 index and that Fidelity Real Estate Fund has correlation $r = 0.37$ with the index.

(a) Which of these funds has the closer relationship to returns from the stock market as a whole? How do you know? (*Hint*: See page 325.)

(b) Does the information given tell Joe anything about which fund has had higher returns? (*Hint*: See page 328.)

PART II PROJECTS

Projects are longer exercises that require gathering information or producing data and that emphasize writing a short essay to describe your work. Many are suitable for teams of students.

Project 1. Statistical graphics in the press. Graphs good and bad fill the news media. Some publications, such as *USA Today*, make particularly heavy use of graphs to present data. Collect several graphs (at least five) from newspapers and magazines (not from advertisements). Include some graphs that, in your opinion, represent good style and some that represent poor style or are misleading. Use your collection as examples in a brief essay about the clarity, accuracy, and attractiveness of graphs in the press.

Project 2. Roll your own regression. Choose two quantitative variables that you think have a roughly straight-line relationship. Gather data on these variables and do a statistical analysis: make a scatterplot, find the correlation, find the regression line (use a statistical calculator or software), and draw the line on your plot. Then write a report on your work. Some examples of suitable pairs of variables are the following:

(a) The height and arm span of a group of people.

(b) The height and walking stride length of a group of people.

(c) The price per ounce and bottle size in ounces for several brands of shampoo and several bottle sizes for each brand.

Project 3. High school dropouts. Write a factual report on high school dropouts in the United States. The following are examples of questions you might address: Which states have the highest percentages of adults who did not finish high school? How do the earnings and employment rates of dropouts compare with those of other adults? Is the percentage who fail to finish high school higher among blacks and Hispanics than among whites?

The Census Bureau website will supply you with data. Go to www.census.gov/hhes/socdemo/education/.

Project 4. Association is not causation. Write a snappy, attention-getting article on the theme that "association is not causation." Use pointed but not-too-serious examples like those in Example 6 (page 348) and Exercise 15.30 (page 364) of Chapter 15, or this one: there is an association between long hair and height (because women tend to have longer hair than men but also tend to be shorter), but cutting a person's hair will not make him or her taller. Be clear, but don't be technical. Imagine that you are writing for high school students.

Project 5. Military spending. Here are data on U.S. spending for national defense for the fiscal years between 1940 and 2010 from the *Statistical Abstract*. You may want to look in the latest volume for data from the most recent year. You can also find the amounts for every year between 1940 and the present at www.whitehouse.gov/omb/budget/. See the pdf file available by clicking on *Historical Tables*. Look in Section 3 of this pdf file. The units are billions of dollars (this is serious money).

Year:	1940	1945	1950	1955	1960	1965
Military spending:	1.7	[illegible]	13.7	42.7	48.1	50.6

Year:	1970	1975	1980	1985	1990	1995
Military spending:	81.7	86.5	134.0	252.7	299.3	272.1

Year:	2000	2005	2010	2015
Military spending:	294.5	495.3	693.5	597.5

Write an essay that describes the changes in military spending in real terms during this period from just before World War II until a decade after the end of the cold war. Do the necessary calculations and write a brief description that ties military spending to the major military events of this period: World War II (1941–1945), the Korean War (1950–1953), the Vietnam War (roughly 1964–1975), the end of the cold war after the fall of the Berlin Wall in 1989, and the U.S. war with Iraq (beginning in March 2003). You may want to look at years not included in the table to help you as you write your essay.

Project 6. Your pulse rate. What is your "resting pulse rate"? Of course, even if you measure your pulse rate while resting, it may vary from day to day and with the time of day. Measure your resting pulse rate at least six times each day (spaced throughout the day) for at least four days. Write a discussion that includes a description of how you made your measurements and an analysis of your data. Based on the data, what would you say when someone asks you what your resting pulse rate is? (If several students do this project, you can discuss variation in pulse rate among a group of individuals as well.)

Project 7. The dates of coins. Coins are stamped with the year in which they were minted. Collect data from at least 50 coins of

each denomination: pennies, nickels, dimes, and quarters. Write a description of the distribution of dates on coins now in circulation, including graphs and numerical descriptions. Are there differences among the denominations? Did you find any outliers?