## Exercises for Section 1.1

### SKILL BUILDER 1

For the situations described in Exercises 1.1 to 1.6:

(a) What are the cases?

(b) What is the variable and is it quantitative or categorical?

**1.1** People in a city are asked if they support a new recycling law.

**1.2** Record the percentage change in the price of a stock for 100 stocks publicly traded on Wall Street.

**1.3** Collect data from a sample of teenagers with a question that asks "Do you eat at least five servings a day of fruits and vegetables?"

**1.4** Measure the shelf life of bunches of bananas (the number of days until the bananas go bad) for a large sample.

**1.5** Estimate the bending strength of beams by bending 10 beams until they break and recording the force at which the beams broke.

**1.6** Record whether or not the literacy rate is over 75% for each country in the world.

### SKILL BUILDER 2

In Exercises 1.7 to 1.10, a relationship between two variables is described. In each case, we can think of one variable as helping to explain the other. Identify the explanatory variable and the response variable.

**1.7** Lung capacity and number of years smoking cigarettes

**1.8** Amount of fertilizer used and the yield of a crop

**1.9** Blood alcohol content (BAC) and number of alcoholic drinks consumed

**1.10** Year and the world record time in a marathon

**1.11 Student Survey Variables** Data 1.1 introduced the dataset **StudentSurvey**, and Example 1.2 identified seven of the variables in that dataset as categorical or quantitative. The remaining variables are:

| | |
|---|---|
| *Year* | FirstYear, Sophomore, Junior, Senior |
| *Height* | In inches |
| *Weight* | In pounds |
| *Siblings* | Number of siblings the person has |
| *VerbalSAT* | Score on the Verbal section of the SAT exam |
| *MathSAT* | Score on the Math section of the SAT exam |
| *SAT* | Sum of the scores on the verbal and math sections of the SAT exam |
| *HigherSAT* | Which is higher, Math SAT score or Verbal SAT score? |

(a) Indicate whether each variable is quantitative or categorical.

(b) List at least two questions we might ask about any one of these individual variables.

(c) List at least two questions we might ask about relationships between any two (or more) of these variables.

**1.12 Countries of the World** Information about the world's countries is given in **AllCountries**, introduced in Data 1.2 on page 7. You can find a description of the variables in the appendix. For the full dataset:

(a) Indicate which of the variables are quantitative and which are categorical.

(b) List at least two questions we might ask about any one of these individual variables.

(c) List at least two questions we might ask about relationships between any two (or more) of these variables.

**1.13 Diet and Retinol and Beta-Carotene Levels** The data from a study[8] examining the association between diet and plasma retinol and plasma beta-carotene levels are given in **NutritionStudy**. The data include 315 observations and 16 variables that are described in the appendix.

(a) Indicate which of the variables are quantitative and which are categorical.

(b) Discuss one possible relationship of interest in this dataset between two categorical variables. Between two quantitative variables. Between one categorical and one quantitative variable.

**1.14 Spider Sex Play** Spiders regularly engage in spider foreplay that does not culminate in mating. Male spiders mature faster than female spiders and often practice the mating routine on not-yet-mature females. Since male spiders run the risk of getting eaten by female spiders, biologists wondered why

[8]Nierenberg, D., et. al., "Determinants of plasma levels of beta-carotene and retinol", *American Journal of Epidemiology*, 1989 Sep; 130(3): 511–21.

spiders engage in this behavior. In one study,[9] some spiders were allowed to participate in these near-matings, while other maturing spiders were isolated. When the spiders were fully mature, the scientists observed real matings. They discovered that if either partner had participated at least once in mock sex, the pair reached the point of real mating significantly faster than inexperienced spiders did. (Mating faster is, apparently, a real advantage in the spider world.) Describe the variables, indicate whether each variable is quantitative or categorical, and indicate the explanatory and response variables.

**1.15 Hormones and Fish Fertility** When women take birth control pills, some of the hormones found in the pills eventually make their way into lakes and waterways. In one study, a water sample was taken from various lakes. The data indicate that as the concentration of estrogen in the lake water goes up, the fertility level of fish in the lake goes down. The estrogen level is measured in parts per trillion (ppt) and the fertility level is recorded as the percent of eggs fertilized. What are the cases in this study? What are the variables? Classify each variable as either categorical or quantitative.

**1.16 Fast-Twitch Muscles and Race** Example 1.5 studied a variant of the gene *ACTN3* which inhibits fast-twitch muscles and seems to be less prevalent in sprinters. A separate study[10] indicated ethnic differences: Approximately 20% of a sample of Caucasians, approximately 25% of a sample of Asians, and approximately 1% of a sample of Africans had the gene variant. What are the variables in this study? Classify each as categorical or quantitative.

**1.17 Trans-Generational Effects of Diet** Can experiences of parents affect future children? New studies[11] suggest that they can: Early life experiences of parents appear to cause permanent changes in sperm and eggs. In one study, some male rats were fed a high-fat diet with 43% of calories from fat (a typical American diet), while others were fed a normal healthy rat diet. Not surprisingly, the rats fed the high-fat diet were far more likely than the normal-diet rats to develop metabolic syndrome (characterized by such things as excess weight, excess fat, insulin resistance, and glucose intolerance.) What surprised the scientists was that the daughters of these rats were also far more likely to develop metabolic syndrome than the daughters of rats fed healthy diets. None of the daughters and none of the mothers ate a high-fat diet and the fathers did not have any contact with the daughters. The high-fat diet of the fathers appeared to cause negative effects for their daughters. What are the two main variables in this study? Is each categorical or quantitative? Identify the explanatory and response variables.

**1.18 Trans-Generational Effects of Environment** In Exercise 1.17, we ask whether experiences of parents can affect future children, and describe a study that suggests the answer is yes. A second study, described in the same reference, shows similar effects. Young female mice were assigned to either live for two weeks in an enriched environment or not. Matching what has been seen in other similar experiments, the adult mice who had been exposed to an enriched environment were smarter (in the sense that they learned how to navigate mazes faster) than the mice that did not have that experience. The other interesting result, however, was that the offspring of the mice exposed to the enriched environment were also smarter than the offspring of the other mice, even though none of the offspring were exposed to an enriched environment themselves. What are the two main variables in this study? Is each categorical or quantitative? Identify explanatory and response variables.

**1.19 Hookahs and Health** Hookahs are waterpipes used for smoking flavored tobacco. One study[12] of 3770 university students in North Carolina found that 40% had smoked a hookah at least once, with many claiming that the hookah smoke is safer than cigarette smoke. However, a second study observed people at a hookah bar and recorded the length of the session, the frequency of puffing, and the depth of inhalation. An average session lasted one hour and the smoke inhaled from an average session was equal to the smoke in more than 100 cigarettes. Finally, a third study measured the amount of tar, nicotine, and heavy metals in samples of hookah smoke, finding that the water in a hookah filters out only a very small percentage of these chemicals. Based on these studies and others, many states are introducing laws to ban or limit hookah bars. In each of the three studies, identify the individual cases, the variables, and whether each variable is quantitative or categorical.

[9]Pruitt, J., paper presented at the Society for Integrative and Comparative Biology Annual Meeting, January 2011, and reported in "For spiders, sex play has its pluses," *Science News*, January 29, 2011.
[10]North, K., et. al., "A common nonsense mutation results in $\alpha$-actinin-3 deficiency in the general population," *Nature Genetics*, April 1999; 21(4): 353–354.
[11]Begley, S., "Sins of the Grandfathers", *Newsweek*, November 8, 2010, pp 48 - 50.
[12]Quenqua, D., "Putting a Crimp in the Hookah", New York Times, May 31, 2011, p A1.

**1.20 Rowing Solo Across the Atlantic Ocean** On January 14, 2012, Andrew Brown of Great Britain set the world record time (40 days) for rowing solo across the northern Atlantic Ocean. On March 14, 2010, Katie Spotz of the United States became the youngest person to ever row solo across the Atlantic when she completed it in 70 days at the age of 22 years old. Table 1.3 shows times for males and females who rowed solo across the Atlantic Ocean in the last few years.[13]

**Table 1.3** *Number of days to row alone across the Atlantic Ocean*

| | |
|---|---|
| Male times: | 40, 87, 78, 106, 67 |
| Female times: | 70, 153, 81 |

(a) How many cases are there in this dataset? How many variables are there and what are they? Is each categorical or quantitative?

(b) Display the information in Table 1.3 as a dataset with cases as rows and variables as columns.

**1.21 How Are Age and Income Related?** An economist collects data from many people to determine how age and income are related. How the data is collected determines whether the variables are quantitative or categorical. Describe how the information might be recorded if we regard both variables as quantitative. Then describe a different way to record information about these two variables that would make the variables categorical.

**1.22 Psychological and Physiological Effects of Meditation** Forty-one employees of a biotechnology company participated in a study[14] that examines the immunological and psychological effects of meditation. Twenty-five of the participants, chosen at random, completed an 8-week meditation program while the other sixteen employees did no meditation. Brain wave activity across the front of the left hemisphere was measured for all participants before, immediately following, and four months after the program. (Studies have suggested that increased activity in this part of the brain is associated with decreases in negative emotions and increases in positive emotions.) All 41 people received an influenza vaccination at the end of the program and their immune response to the vaccine was measured through blood samples taken one month and two months later. All participants also completed surveys designed to measure negative and positive emotions before and after the course. The surveys produced two numerical scores (one for positive emotions and one for negative emotions) in both situations.

Meditators showed an increase in brain wave activity, a decrease in reported negative feelings, and no change in reported positive feelings. Non-meditators showed no significant change in any of these areas. Meditators had a stronger antibody response to the vaccine than the non-meditators.

(a) What are the cases in this study? How many cases are there?

(b) What are the variables? Which are categorical and which are quantitative?

(c) Which variable is the explanatory variable?

(d) How many rows and how many columns will the dataset contain if we assume that each data case is a row and each variable is a column?

**1.23 Special Shakes** A large restaurant chain (see Example 1.6) periodically offers special milk shake flavors for a limited time. Suppose that the contenders for the next special flavor are Green Mint, Orange Crush, Egg Nog, and Piña Colada. The chain plans to collect data from customers on these flavors, and there are several ways they might solicit responses. For each of the options below, state the number of variables needed to code the information in a dataset, whether the variable(s) is/are categorical or quantitative, and what sort of values should be recorded.

(a) "Which of the four flavors is most appealing to you?"

(b) "Put a check next to any of the four flavors you find appealing."

(c) "Please rank the four flavors with 1=most appealing and 4=least appealing."

(d) "Rate each of the four flavors on a 1 to 10 scale with 10=extremely appealing and 1=very unappealing."

**1.24 Political Party and Voter Turnout** Suppose that we want to investigate the question "Does voter turnout differ by political party?" How might we collect data to answer this question? What would the cases be? What would the variable(s) be?

**1.25 Wealth and Happiness** Are richer people happier? How might we collect data to answer this question? What would the cases be? What would the variable(s) be?

[13] http://www.oceanrowing.com/statistics/ocean_rowing_records2.htm.

[14] Davidson, R., et. al., "Alterations in brain and immune function produced by mindfulness meditation," *Psychosomatic Medicine*, July/August, 2003, 65: 564–570.

**1.26 Choose Your Own Question** Come up with your own question you would like to be able to answer. What is the question? How might you collect data to answer this question? What would the cases be? What would the variable(s) be?

# 1.2 SAMPLING FROM A POPULATION

While most of this textbook is devoted to analyzing data, the way in which data are *collected* is critical. Data collected well can yield powerful insights and discoveries. Data collected poorly can yield very misleading results. Being able to think critically about the method of data collection is crucial for making or interpreting data-based claims. In the rest of this chapter, we address some of the most important issues that need to be considered when collecting data.

## Samples from Populations

The US Census is conducted every 10 years and attempts to gather data about all people living in the US. For example, the census shows that, for people living in the US who are at least 25 years old, 84.6% have at least a high school degree and 27.5% have at least a college bachelor's degree.[15] The cases in the census dataset are all residents of the US, and there are many variables measured on these cases. The US census attempts to gather information from an entire *population*. In **AllCountries**, introduced as Data 1.2 on page 7, the cases are countries. This is another example of a dataset on an entire population because we have data on every country.

Usually, it is not feasible to gather data for an entire population. If we want to estimate the percent of people who wash their hands after using a public restroom, it is certainly not possible to observe all people all the time. If we want to try out a new drug (with possible side effects) to treat cancer, it is not safe to immediately give it to all patients and sit back to observe what happens. If we want to estimate what percentage of people will react positively to a new advertising campaign, it is not feasible to show the ads to everyone and then track their responses. In most circumstances, we can only work with a *sample* from what might be a very large population.

### Samples from Populations

A **population** includes all individuals or objects of interest.

Data are collected from a **sample**, which is a subset of the population.

**Example 1.10**

To estimate what percent of people in the US wash their hands after using a public restroom, researchers pretended to comb their hair while observing 6000 people in public restrooms throughout the United States. They found that 85% of the people who were observed washed their hands after going to the bathroom.[16] What is the sample in this study? What is a reasonable population to which we might generalize?

*Solution*

The sample is the 6000 people who were observed. A reasonable population to generalize to would be all people in the US. There are other reasonable answers to

[15] http://factfinder.census.gov.

[16] Zezima, K., "For many, 'Washroom' seems to be just a name," *New York Times*, September 14, 2010, p A14.

**Example 1.21**

*Illicit Drug Use*

The 2009 National Survey on Drug Use and Health[22] selected a random sample of US college students and asked them about illicit drug use, among other things. In the sample, 22.7% of the students reported using illicit drugs in the past year. Do you think this is an accurate portrayal of the percentage of all college students using illicit drugs?

*Solution*

This may be an underestimate. Even if the survey is anonymous, students may be reluctant to report illicit drug use on an official survey, and thus may not answer truthfully.

Bias in data collection can result in many other ways not discussed here. The most important message is to always think critically about the way data are collected and to recognize that not all methods of data collection lead to valid inferences. Recognizing sources of bias is often simply common sense, and you will instantly become a more statistically literate individual if, each time you are presented with a statistic, you just stop, inquire, and think about how the data were collected.

**SECTION LEARNING GOALS**

- Distinguish between a sample and a population
- Recognize when it is appropriate to use sample data to infer information about the population
- Critically examine the way a sample is selected, identifying possible sources of sampling bias
- Recognize that random sampling is a powerful way to avoid sampling bias
- Identify other potential sources of bias that may arise in studies on humans

## Exercises for Section 1.2

**SKILL BUILDER 1**

In Exercises 1.27 to 1.30, state whether the data are best described as a population or a sample.

**1.27** To estimate size of trout in a lake, an angler records the weight of 12 trout he catches over a weekend.

**1.28** A subscription-based music website tracks its total number of active users.

**1.29** The U.S. Department of Transportation announces that of the 250 million registered passenger vehicles in the US, 2.1% are electro-gas hybrids.

**1.30** A questionnaire to understand athletic participation on a college campus is emailed to 50 college students, and all of them respond.

**SKILL BUILDER 2**

In Exercises 1.31 to 1.36, describe the sample and describe a reasonable population.

**1.31** A sociologist conducting a survey at a mall interviews 120 people about their cell phone use.

[22]Substance Abuse and Mental Health Services Administration, Results from the 2009 National Survey on Drug Use and Health: Volume I. Summary of National Findings (Office of Applied Studies, NSDUH Series H-38A, HHS Publication No. SMA 10-4856Findings), Rockville, MD, 2010, https://nsduhweb.rti.org/.

**1.32** A fishing boat captain examines one day's catch of fish to see if the average weight of fish in that area is large enough to make fishing there profitable.

**1.33** Five hundred Canadian adults are asked if they are proficient on a musical instrument.

**1.34** A cell phone carrier sends a satisfaction survey to 100 randomly selected customers.

**1.35** A hungry yet diligent snacker eats an entire package of Chips Ahoy! cookies while counting and recording the number of chocolate chips in each cookie.

**1.36** The Nielsen Corporation attaches databoxes to televisions in 1000 households throughout the US to monitor what shows are being watched and produce the Nielsen Ratings for television.

### SKILL BUILDER 3

In Exercises 1.37 to 1.40, a biased sampling situation is described. In each case, give:

(a) The sample

(b) The population of interest

(c) A population we can generalize to given the sample.

**1.37** To estimate the proportion of Americans who support changing the drinking age from 21 to 18, a random sample of 100 college students are asked the question "Would you support a measure to lower the drinking age from 21 to 18?"

**1.38** To investigate growth of the canine population in New York City, 100 dogs are randomly selected from a registry of licensed pets in the city, and it is found that 78 of them have been neutered.

**1.39** To investigate interest across all residents of the US in a new type of ice skate, a random sample of 1500 people in Minnesota are asked about their interest in the product.

**1.40** To determine the height distribution of female high school students, the rosters are collected from 20 randomly selected high school girls basketball teams.

### SKILL BUILDER 4

In Exercises 1.41 to 1.46, state whether or not the sampling method described produces a random sample from the given population.

**1.41** The population is incoming students at a particular university. The name of each incoming student is thrown into a hat, the hat is mixed, and 20 names (each corresponding to a different student) are drawn from the hat.

**1.42** The population is the approximately 25,000 protein-coding genes in human DNA. Each gene is assigned a number (from 1 to 25,000), and computer software is used to randomly select 100 of these numbers yielding a sample of 100 genes.

**1.43** The population is all employees at a company. All employees are emailed a link to a survey.

**1.44** The population is adults between the ages of 18 and 22. A sample of 100 students is collected from a local university, and each student at the university had an equal chance of being selected for the sample.

**1.45** The population is all trees in a forest. We walk through the forest and pick out trees that appear to be representative of all the trees in the forest.

**1.46** The population is all people who visit the website *CNN.com*. All visitors to the website are invited to take part in the daily online poll.

### IS IT BIASED?

In Exercises 1.47 to 1.51, indicate whether we should trust the results of the study. Is the method of data collection biased? If it is, explain why.

**1.47** Ask a random sample of students at the library on a Friday night "How many hours a week do you study?" to collect data to estimate the average number of hours a week that all college students study.

**1.48** Ask a random sample of people in a given school district "Excellent teachers are essential to the well-being of children in this community, and teachers truly deserve a salary raise this year. Do you agree?" Use the results to estimate the proportion of all people in the school district who support giving teachers a raise.

**1.49** Take 10 apples off the top of a truckload of apples and measure the amount of bruising on those apples to estimate how much bruising there is, on average, in the whole truckload.

**1.50** Take a random sample of one type of printer and test each printer to see how many pages of text each will print before the ink runs out. Use the average from the sample to estimate how many pages, on average, all printers of this type will last before the ink runs out.

**1.51** Send an email to a random sample of students at a university asking them to reply to the question: "Do you think this university should fund an ultimate frisbee team?" A small number of students reply. Use the replies to estimate the proportion of all students at the university who support this use of funds.

**1.52 Do Parents Regret Having Children?** In Data 1.4 on page 24, we describe the results of a question asked by a national newspaper columnist: "If you had it to do over again, would you have children?" In addition to those results and a follow-up national survey, the *Kansas City Star* selected a random sample of parents from Kansas City and asked them the same question. In this sample, 94% said "Yes." To what population can this statistic be generalized?

**1.53 How Many People Wash Their Hands After Using the Washroom?** In Example 1.10 on page 16, we introduce a study by researchers from Harris Interactive who were interested in determining what percent of people wash their hands after using the washroom. They collected data by standing in public restrooms and pretending to comb their hair or put on make-up as they observed patrons' behavior.[23] Public restrooms were observed at Turner's Field in Atlanta, Penn Station and Grand Central Station in New York, the Museum of Science and Industry and the Shedd Aquarium in Chicago, and the Ferry Terminal Farmers Market in San Francisco. Of the over 6000 people whose behavior was observed, 85% washed their hands. Women were more likely to wash their hands: 93% of women washed, while only 77% of men did. The Museum of Science and Industry in Chicago had the highest hand-washing rate, while men at Turner's Field in Atlanta had the lowest.

(a) What are the cases? What are the variables? Classify each variable as quantitative or categorical.

(b) In a separate telephone survey of more than 1000 adults, more than 96% said they always wash their hands after using a public restroom. Why do you think there is such a discrepancy in the percent from the telephone survey compared to the percent observed?

**1.54 Teaching Ability** In a sample survey of professors at the University of Nebraska, 94% of them described themselves as "above average" teachers.[24]

(a) What is the sample? What is the population?

(b) Based on the information provided, can we conclude that the study suffers from sampling bias?

(c) Is 94% a good estimate for the percentage of above-average teachers at the University of Nebraska? If not, why not?

**1.55 Does Physical Beauty Matter?** One of the daily polls on *CNN.com* during June 2011 asked "Does Physical Beauty Matter to You?" Of 38,485 people responding, 79% said yes and 21% said no. Can we conclude that about 79% of all people think physical beauty matters? Why or why not? In making such a conclusion, what are we considering the sample? What are we considering the population? Is there any bias in the sampling method?

**1.56 Effects of Alcohol and Marijuana** In 1986 the Federal Office of Road Safety in Australia conducted an experiment to assess the effects of alcohol and marijuana on mood and performance.[25] Participants were volunteers who responded to advertisements for the study on two rock radio stations in Sydney. Each volunteer was given a randomly determined combination of the two drugs, then tested and observed. Is the sample likely representative of all Australians? Why or why not?

**1.57 What Percent of Young Adults Move Back in with Their Parents?** The Pew Research Center polled a random sample of $n = 808$ US residents between the ages of 18 and 34. Of those in the sample, 24% had moved back in with their parents for economic reasons after living on their own.[26] Do you think that this sample of 808 people is a representative sample of all US residents between the ages of 18 and 34? Why or why not?

**1.58 Do Tanning Salons Mislead Their Customers?** Investigators posing as fair-skinned teenage girls contacted 300 tanning salons nationwide, including at least three randomly selected in each state. The investigators report that 90% of the salons stated that indoor tanning did not pose a health risk and over half (51%) of the salons denied that indoor tanning would increase a fair-skinned teenager's risk of developing skin cancer. Going even further, 78% of the tanning salons even claimed that indoor tanning is beneficial to health.[27] (In fact, many

[23] Bakalar, "Study: More people washing hands after using bathroom," *Salem News*, September 14, 2010.
[24] Cross, P., "Not can, but *will* college teaching be improved?," *New Directions for Higher Education*, 1977: 17: 115.
[25] Chesher, G., Dauncey, H., Crawford, J. and Horn, K., "The Interaction between Alcohol and Marijuana: A Dose Dependent Study on the Effects on Human Moods and Performance Skills," Report No. C40, Federal Office of Road Safety, Federal Department of Transport, Australia, 1986.
[26] Parker, K., "The Boomerang Generation: Feeling OK about Living with Mom and Dad," Pew Research Center, March 15, 2012.
[27] "Congressional Report Exposes Tanning Industry's Misleading Messaging to Teens," http://www.skincancer.org/news/tanning/tanningreport, a report released by the House Committee on Energy and Commerce, February 1, 2012.

studies have shown that tanning is dangerous, especially for teenagers, and that tanning raises the risk of melanoma, the deadliest type of skin cancer, by 74%.)

(a) What is the sample?

(b) Do you think the sample is representative of all tanning salons in the US?

(c) Although the sample is random, discuss why the results do not paint an accurate picture of the dangers of tanning.

(d) Do you think the study accurately portrays the messages tanning salons give to teenage girls?

**1.59 Employment Surveys** Employment statistics in the US are often based on two nationwide monthly surveys: the Current Population Survey (CPS) and the Current Employment Statistics (CES) survey. The CPS samples approximately 60,000 US households and collects the employment status, job type, and demographic information of each resident in the household. The CES survey samples 140,000 non-farm businesses and government agencies and collects the number of payroll jobs, pay rates, and related information for each firm.

(a) What is the population in the CPS survey?

(b) What is the population in the CES survey?

(c) For each of the following statistical questions, state whether the results from the CPS or CES survey would be more relevant.

i. Do larger companies tend to have higher salaries?

ii. What percentage of Americans are self-employed?

iii. Are married men more or less likely to be employed than single men?

**1.60 National Health Statistics** The Centers for Disease Control and Prevention (CDC) administers a large number of survey programs for monitoring the status of health and health care in the US. One of these programs is the National Health and Nutrition Examination Survey (NHANES), which interviews and examines a random sample of about 5000 people in the US each year. The survey includes questions about health, nutrition, and behavior while the examination includes physical measurements and lab tests. Another program is the National Hospital Ambulatory Medical Care Survey (NHAMCS), which includes information from hospital records for a random sample of individuals treated in hospital emergency rooms around the country.

(a) To what population can we reasonably generalize findings from the NHANES?

(b) To what population can we reasonably generalize findings from the NHAMCS?

(c) For each of the questions below, indicate which survey, NHANES or NHAMCS, would probably be more appropriate to address the issue.

i. Are overweight people more likely to develop diabetes?

ii. What proportion of emergency room visits in the US involve sports-related injuries?

iii. Is there a difference in the average waiting time to be seen by an emergency room physician between male and female patients?

iv. What proportion of US residents have visited an emergency room within the past year?

**1.61 Interviewing the Film Crew on Hollywood Movies** There were 136 movies made in Hollywood in 2011. Suppose that, for a documentary about Hollywood film crews, a random sample of 5 of these movies will be selected for in-depth interviews with the crew members. Assuming the movies are numbered 1 to 136, use a random number generator or table to select a random sample of five movies by number. Indicate which numbers were selected. (If you want to know which movies you selected, check out the dataset **HollywoodMovies2011**.)

**1.62 Sampling Some Hardee's Restaurants** The Hardee's Restaurant chain has about 1900 quick-serve restaurants in 30 US states and 9 countries.[28] Suppose that a member of the Hardee's administration wishes to visit 6 of these restaurants, randomly selected, to gather some first-hand data. Suppose the restaurants are numbered 1 to 1900. Use a random-number generator or table to select the numbers for 6 of the restaurants to be in the sample.

**1.63 Strawberry Fields** A strawberry farmer has planted 100 rows of plants, each 12 inches apart, and there are about 300 plants in each row. He would like to select a random sample of 30 plants to estimate the average number and weight of berries per plant.

(a) Explain how he might choose the specific plants to include in the sample.

(b) Carry out your procedure from (a) to identify the first three plants selected for the sample.

[28] hardees.com/company/franchise.

**Example 1.32**

Does an injection of caffeine help rats learn a maze faster? Design an experiment to investigate this question. Incorporate elements of a well-designed experiment.

*Solution*

We take the rats that are available for the study and *randomly* divide them into two groups. One group will get a shot of caffeine while the other group will get a shot of saline solution (placebo). We have the rats run the maze and record their times. Don't tell the rats which group they are in! Ideally, all people who come in contact with the rats (the people giving the shots, the people recording the maze times, and so on) should not know which rats are in which group. This makes the study double-blind. Only the statistician analyzing the data will know which rats are in which group. (We describe here a randomized comparative experiment. A matched pairs experiment would also work, and in that case we would also use a placebo and blinding.)

### Realities of Randomized Experiments

Randomization should always be used in designing an experiment. Blinding and the use of a placebo treatment should be used when appropriate and possible. However, there are often ethical considerations that preclude the use of an experiment in any form. For example, imagine designing an experiment to determine whether cell phones cause cancer or whether air pollution leads to adverse health consequences. It would not be appropriate to require people to wear a cell phone on their head for large amounts of time to see if they have higher cancer rates! Similarly, it would not be appropriate to require some people to live in areas with more polluted air. In situations such as these, observational studies can at least help us determine associations.

**SECTION LEARNING GOALS**

- Recognize that not every association implies causation
- Identify potential confounding variables in a study
- Distinguish between an observational study and a randomized experiment
- Recognize that only randomized experiments can lead to claims of causation
- Explain how and why placebos and blinding are used in experiments
- Distinguish between a randomized comparative experiment and a matched pairs experiment
- Design and implement a randomized experiment

## Exercises for Section 1.3

**SKILL BUILDER 1**

In Exercises 1.64 to 1.69, we give a headline that recently appeared online or in print. State whether the claim is one of association and causation, association only, or neither association nor causation.

**1.64** Daily exercise improves mental performance.

**1.65** Among college students, no link found between number of friends on social networking websites and size of the university.

**1.66** Cell phone radiation leads to deaths in honey bees.

**1.67** Wealthy people are more likely than other folks to lie, cheat, and steal.

**1.68** Cat owners tend to be more educated than dog owners.

**1.69** Want to lose weight? Eat more fiber!

### SKILL BUILDER 2

Exercises 1.70 to 1.75 describe an association between two variables. Give a confounding variable that may help to account for this association.

**1.70** More ice cream sales have been linked to more deaths by drowning.

**1.71** The total amount of beef consumed and the total amount of pork consumed worldwide are closely related over the past 100 years.

**1.72** People who own a yacht are more likely to buy a sports car.

**1.73** Sales of toboggans tend to be higher when sales of mittens are higher.

**1.74** Air pollution is higher in places with a higher proportion of paved ground relative to grassy ground.

**1.75** People with shorter hair tend to be taller.

### SKILL BUILDER 3

In Exercises 1.76 to 1.79, we describe data collection methods to answer a question of interest. Are we describing an experiment or an observational study?

**1.76** To examine whether eating brown rice affects metabolism, we ask a random sample of people whether they eat brown rice and we also measure their metabolism rate.

**1.77** To examine whether playing music in a store increases the amount customers spend, we randomly assign some stores to play music and some to stay silent and compare the average amount spent by customers.

**1.78** To examine whether planting trees reduces air pollution, we find a sample of city blocks with similar levels of air pollution and we then plant trees in half of the blocks in the sample. After waiting an appropriate amount of time, we measure air pollution levels.

**1.79** To examine whether farm-grown salmon contain more omega-3 oils if water is more acidic, we collect samples of salmon and water from multiple fish farms to see if the two variables are related.

### REVISITING QUESTIONS FROM SECTION 1.1

Exercises 1.80 to 1.82 refer to questions of interest asked in Section 1.1 in which we describe data collection methods. Indicate whether the data come from an experiment or an observational study.

**1.80** "Is there a sprinting gene?" Introduced in Example 1.5 on page 9.

**1.81** "Do metal tags on penguins harm them?" Introduced in Data 1.3 on page 10.

**1.82** "Are there human pheromones?" Introduced on page 11. Three studies are described; indicate whether each of them is an experiment or an observational study.

**1.83 Shoveling Snow** Three situations are described at the start of this section, on page 29. In the second bullet, we describe an association between activity at a building's heating plant and more employees missing work due to back pain. A confounding variable in this case is amount of snow. Describe how snowfall meets the definition of a confounding variable by describing how it might be associated with both the variables of interest.

**1.84 Salt on Roads and Accidents** Three situations are described at the start of this section, on page 29. In the third bullet, we describe an association between the amount of salt spread on the roads and the number of accidents. Describe a possible confounding variable and explain how it fits the definition of a confounding variable.

**1.85 Height and Reading Ability** In elementary school (grades 1 to 6), there is a strong association between a child's height and the child's reading ability. Taller children tend to be able to read at a higher level. However, there is a very significant confounding variable that is influencing both height and reading ability. What is it?

**1.86 Exercise and Alzheimer's Disease** A headline at *MSNBC.com*[41] stated "One way to ward off Alzheimer's: Take a hike. Study: Walking at least one mile a day reduces risk of cognitive impairment by half." The article reports on a study[42] showing that elderly people who walked a lot tended to have more brain mass after nine years and were less likely to develop dementia that can lead to Alzheimer's disease than subjects who walked less. At the start of the study the researchers measured

[41] http://www.msnbc.msn.com/id/39657391/ns/health-alzheimer's_disease.

[42] Erickson, K., et. al., "Physical activity predicts gray matter volume in late adulthood: The Cardiovascular Health Study," *Neurology*, published online October 13,2010.

the walking habits of the elderly subjects and then followed up with measures of brain volume nine years later. Assuming that active walkers really did have more brain mass and fewer dementia symptoms, is the headline justified?

**1.87 Single-Sex Dorms and Hooking Up** The president of a large university recently announced[43] that the school would be switching to dorms that are all single-sex, because, he says, research shows that single-sex dorms reduce the number of student hook-ups for casual sex. He cites studies showing that, in universities that offer both same-sex and co-ed housing, students in co-ed dorms report hooking up for casual sex more often.

(a) What are the cases in the studies cited by the university president? What are the two variables being discussed? Identify each as categorical or quantitative.

(b) Which is the explanatory variable and which is the response variable?

(c) According to the second sentence, does there appear to be an association between the variables?

(d) Use the first sentence to determine whether the university president is assuming a causal relationship between the variables.

(e) Use the second sentence to determine whether the cited studies appear to be observational studies or experiments?

(f) Name a confounding variable that might be influencing the association. (*Hint*: Students usually request one type of dorm or the other.)

(g) Can we conclude from the information in the studies that single-sex dorms reduce the number of student hook-ups?

(h) What common mistake is the university president making?

**1.88 Music Volume and Beer Consumption** In 2008, a study[44] was conducted measuring the impact that music volume has on beer consumption. The researchers went into bars, controlled the music volume, and measured how much beer was consumed. The article states that "the sound level of the environmental music was manipulated according to a randomization scheme." It was found that louder music corresponds to more beer consumption. Does this provide evidence that louder music causes people to drink more beer? Why or why not?

**1.89 Does Red Increase Men's Attraction to Women?** A recent study[45] examined the impact of the color red on how attractive men perceive women to be. In the study, men were randomly divided into two groups and were asked to rate the attractiveness of women on a scale of 1 (not at all attractive) to 9 (extremely attractive). One group of men were shown pictures of women on a white background and the other group were shown the same pictures of women on a red background. The men who saw women on the red background rated them as more attractive. All participants and those showing the pictures and collecting the data were not aware of the purpose of the study.

(a) Is this an experiment or an observational study? Explain.

(b) What is the explanatory variable and what is the response variable? Identify each as categorical or quantitative.

(c) How was randomization used in this experiment? How was blinding used?

(d) Can we conclude that using a red background color instead of white increases men's attractiveness rating of women's pictures?

**1.90 Urban Brains and Rural Brains** A study published in 2010 showed that city dwellers have a 21% higher risk of developing anxiety disorders and a 39% higher risk of developing mood disorders than those who live in the country. A follow-up study published in 2011 used brain scans of city dwellers and country dwellers as they took a difficult math test.[46] To increase the stress of the participants, those conducting the study tried to humiliate the participants by telling them how poorly they were doing on the test. The brain scans showed very different levels of activity in stress centers of the brain, with the urban dwellers having greater brain activity than rural dwellers in areas that react to stress.

(a) Is the 2010 study an experiment or an observational study?

(b) Can we conclude from the 2010 study that living in a city increases a person's likelihood of developing an anxiety disorder or mood disorder?

[43]Stepp, L., "Single-sex dorms won't stop drinking or 'hooking-up'", www.cnn.com, June 16, 2011.
[44]Gueguen, N., Jacob, C., Le Guellec, H., Morineau, T. and Lourel, M., "Sound Level of Environmental Music and Drinking Behavior: A Field Experiment With Beer Drinkers," *Alcoholism: Clinical and Experimental Research*, 2008; 32: 1795–1798.
[45]Elliot, A. and Niesta, D., "Romantic Red: Red Enhances Men's Attraction to Women", *Journal of Personality and Social Psychology*, 2008; 95(5): 1150–1164.
[46]"A New York state of mind," *The Economist*, June 25, 2011, p. 94.

(c) Is the 2011 study an experiment or an observational study?

(d) In the 2011 study, what is the explanatory variable and what is the response variable? Indicate whether each is categorical or quantitative.

(e) Can we conclude from the 2011 study that living in a city increases activity in stress centers of the brain when a person is under stress?

**1.91 Be Sure to Get Your Beauty Sleep!** New research[47] supports the idea that people who get a good night's sleep look more attractive. In the study, 23 subjects ages 18 to 31 were photographed twice, once after a good night's sleep and once after being kept awake for 31 hours. Hair, make-up, clothing, and lighting were the same for both photographs. Observers then rated the photographs for attractiveness, and the average rating under the two conditions was compared. The researchers report in the *British Medical Journal* that "Our findings show that sleep-deprived people appear less attractive compared with when they are well rested."

(a) What is the explanatory variable? What is the response variable?

(b) Is this an experiment or an observational study? If it is an experiment, is it a randomized comparative design or a matched pairs design?

(c) Can we conclude that sleep deprivation *causes* people to look less attractive? Why or why not?

**1.92 Do Antidepressants Work?** Following the steps below, design a randomized comparative experiment to test whether fluoxetine (the active ingredient in Prozac pills) is effective at reducing depression. The participants are 50 people suffering from depression and the response variable is the change on a standard questionnaire measuring level of depression.

(a) Describe how randomization will be used in the design.

(b) Describe how a placebo will be used.

(c) Describe how to make the experiment double-blind.

[47]Stein, R., "Beauty sleep no myth, study finds," *Washington Post*, washingtonpost.com, December 15, 2010.

**1.93 Do Children Need Sleep to Grow?** About 60% of a child's growth hormone is secreted during sleep, so it is believed that a lack of sleep in children might stunt growth.[48]

(a) What is the explanatory variable and what is the response variable in this association?

(b) Describe a randomized comparative experiment to test this association.

(c) Explain why it is difficult (and unethical) to get objective verification of this possible causal relationship.

**1.94 Carbo-Loading** It is commonly accepted that athletes should "carbo-load," that is, eat lots of carbohydrates, the day before an event requiring physical endurance. Is there any truth to this? Suppose you want to design an experiment to find out for yourself: "does carbo-loading actually improve athletic performance the following day?" You recruit 50 athletes to participate in your study.

(a) How would you design a randomized comparative experiment?

(b) How would you design a matched pairs experiment?

(c) Which design do you think is better for this situation? Why?

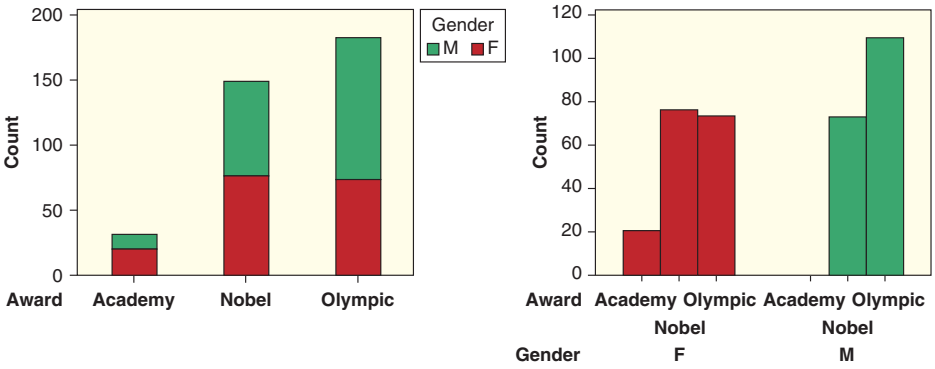
**1.95 Alcohol and Reaction Time** Does alcohol increase reaction time? Design a randomized experiment to address this question using the method described in each case. Assume the participants are 40 college seniors and the response variable is time to react to an image on a screen after drinking either alcohol or water. Be sure to explain how randomization is used in each case.

(a) A randomized comparative experiment with two groups getting two separate treatments

(b) A matched pairs experiment

**1.96 Causation and Confounding** Causation does not necessarily mean that there is no confounding variable. Give an example of an association between two variables that have a causal relationship AND have a confounding variable.

[48]Rochman, B., "Please, Please, Go to Sleep," *Time* magazine, March 26, 2012, p. 46.




(a) A segmented bar chart (b) Side-by-side bar charts

**Figure 2.2** ***Displaying a relationship between gender and preferred award***

such as a segmented bar chart or side-by-side bar charts are called *comparative plots* since they allow us to compare groups in a categorical variable.

**SECTION LEARNING GOALS**

*You should now have the understanding and skills to:*

- Display information from a categorical variable in a table or graph
- Use information about a categorical variable to find a proportion, with correct notation
- Display information about a relationship between two categorical variables in a two-way table
- Use a two-way table to find proportions
- Interpret graphs involving two categorical variables

## Exercises for Section 2.1

### SKILL BUILDER 1

Exercises 2.1 to 2.4 provide information about data in **StudentSurvey**. Find the sample proportion $\hat{p}$.

**2.1** The survey students consisted of 169 females and 193 males. Find $\hat{p}$, the proportion who are female.

**2.2** The survey included 43 students who smoke and 319 who don't. Find $\hat{p}$, the proportion who smoke.

**2.3** Of the students who identified their class year in the survey, 94 were first years, 195 were sophomores, 35 were juniors, and 36 were seniors. Find $\hat{p}$, the proportion who are upperclass students (juniors or seniors.)

**2.4** The math SAT score is higher than the verbal SAT score for 205 of the 355 students who answered the questions about SAT scores. Find $\hat{p}$, the proportion for whom the math SAT score is higher.

### SKILL BUILDER 2

In Exercises 2.5 to 2.8, give the relevant proportion using correct notation.

**2.5** In the 2010 US Census, we see that 37,342,870 people, or 12.4% of all residents, are foreign-born.[4]

[4]www.census.gov.

**2.6** A recent headline states that "73% say Woman President Likely in Next 10 Years." The report gives the results of a survey of 1000 randomly selected likely voters in the US. [5]

**2.7** A survey conducted of 1502 randomly selected US adults found that 931 of them believed the government does not provide enough support for soldiers returning from Iraq or Afghanistan.[6]

**2.8** Of all 1,547,990 members of the high school class of 2010 who took the SAT (Scholastic Aptitude Test), 1,114,273 were from a public high school.[7]

### SKILL BUILDER 3

In Exercises 2.9 and 2.10, data from the **StudentSurvey** dataset is given. Construct a relative frequency table of the data using the categories given. Give the relative frequencies rounded to three decimal places.

**2.9** Of the 362 students who answered the question about what award they would prefer, 31 preferred an Academy Award, 149 preferred a Nobel Prize, and 182 preferred an Olympic gold medal.

**2.10** Of the 361 students who answered the question about the number of piercings they had in their body, 188 had no piercings, 82 had one or two piercings, and the rest had more than two.

### SKILL BUILDER 4

In Exercises 2.11 and 2.12, cases are classified according to one variable, with categories A and B, and also classified according to a second variable with categories 1, 2, and 3. The cases are shown, with the first digit indicating the value of the first variable and the second digit indicating the value of the second variable. (So "A1" represents a case in category A for the first variable and category 1 for the second variable.) Construct a two-way table of the data.

**2.11** Twenty cases:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| A1 | A1 | A1 | A2 | A3 | A3 | A3 | A3 | A3 | A3 |
| A3 | A3 | B1 | B1 | B1 | B1 | B2 | B2 | B2 | B3 |

**2.12** Thirty cases:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| A1 | A1 | A2 | A2 | A2 | A2 | A2 | A2 | A3 | A3 |
| A3 | A3 | B1 | B1 | B1 | B1 | B1 | B2 | B2 | B3 |
| B3 | B3 | B3 | B3 | B3 | B3 | B3 | B3 | B3 | B3 |

[5]Rassmussen Reports, June 27, 2010.
[6]"Four Years After Walter Reed, Government Still Faulted for Troop Support," Pew Research Center, pewresearch.org, June 29, 2011.
[7]professionals.collegeboard.com.

**2.13 Rock-Paper-Scissors** Rock-Paper-Scissors, also called Roshambo, is a popular two-player game often used to quickly determine a winner and loser. In the game, each player puts out a fist (rock), a flat hand (paper), or a hand with two fingers extended (scissors). In the game, rock beats scissors which beats paper which beats rock. The question is: Are the three options selected equally often by players? Knowing the relative frequencies with which the options are selected would give a player a significant advantage. A study[8] observed 119 people playing Rock-Paper-Scissors. Their choices are shown in Table 2.6.

**Table 2.6** *Frequencies in Rock-Paper-Scissors*

| Option selected | Frequency |
|---|---|
| Rock | 66 |
| Paper | 39 |
| Scissors | 14 |
| Total | 119 |

(a) What is the sample in this case? What is the population? What does the variable measure?

(b) Construct a relative frequency table of the results.

(c) If we assume that the sample relative frequencies from part (b) are similar for the entire population, which option should you play if you want the odds in your favor?

(d) The same study determined that, in repeated plays, a player is more likely to repeat the option just picked than to switch to a different option. If your opponent just played paper, which option should you pick for the next round?

**2.14 Home Field Advantage in Soccer** In the book Scorecasting by Moskowitz and Wertheim,[9] we learn that "Across 43 professional soccer leagues in 24 different countries spanning Europe, South America, Asia, Africa, Australia, and the United States (covering more than 66,000 games), the home field advantage [percent of games won by the home team] in soccer worldwide is 62.4%." Is this a population or a sample? What are the cases and approximately how many are there? What is the variable and is it categorical or quantitative?

[8]Eyler, D., Shalla, Z., Doumaux, A., and McDevitt, T., "Winning at Rock-Paper-Scissors" , *College Mathematics Journal*, March 2009.
[9]Moskowitz, T. and Wertheim, L., *Scorecasting*, Crown Archetype, New York, 2011, p. 113.

What is the relevant statistic, including correct notation?

**2.15 Airborne Antibiotics** A recent study shows that antibiotics added to animal feed to accelerate growth can become airborne. Some of these drugs can be toxic if inhaled and may increase the evolution of antibiotic-resistant bacteria. Scientists[10] analyzed 20 samples of dust particles from animal farms. Tylosin, an antibiotic used in animal feed that is chemically related to erythromycin, showed up in 16 of the samples.

(a) What is the variable in this study? What are the individual cases?

(b) Display the results in a frequency table.

(c) Make a bar chart of the data.

(d) Give a relative frequency table of the data.

**2.16 What Type of Cell Phone?** A 2012 survey[11] examined cell phone ownership by US adults. The results of the survey are shown in Table 2.7.

**Table 2.7** *Frequencies in cell phone ownership*

| Cell Phone Owned | Frequency |
|---|---|
| Android smartphone | 458 |
| iPhone smartphone | 437 |
| Blackberry smartphone | 141 |
| Cell phone not smartphone | 924 |
| No cell phone | 293 |
| Total | 2253 |

(a) Make a relative frequency table of the data. Give results to three decimal places.

(b) What percent of the survey respondents do not own a cell phone? What percent own a cell phone but not a smartphone? What percent own a smartphone?

**2.17 Can Dogs Smell Cancer?** Scientists are working to train dogs to smell cancer, including early stage cancer that might not be detected with other means. In previous studies, dogs have been able to distinguish the smell of bladder cancer, lung cancer, and breast cancer. Now, it appears that a dog in Japan has been trained to smell bowel cancer.[12] Researchers collected breath and stool samples from patients with bowel cancer as well as from healthy people. The dog was given five samples in each test, one from a patient with cancer and four from healthy volunteers. The dog correctly selected the cancer sample in 33 out of 36 breath tests and in 37 out of 38 stool tests.

(a) The cases in this study are the individual tests. What are the variables?

(b) Make a two-way table displaying the results of the study. Include the totals.

(c) What proportion of the breath samples did the dog get correct? What proportion of the stool samples did the dog get correct?

(d) Of all the tests the dog got correct, what proportion were stool tests?

**2.18 Does Belief in One True Love Differ by Education Level?** In Data 2.1 on page 46, we introduce a study in which people were asked whether they agreed or disagreed with the statement that there is only one true love for each person. Is the level of a person's education related to the answer given, and if so, how? Table 2.8 gives a two-way table showing the results for these two variables. A person's education is categorized as HS (high school degree or less), Some (some college), or College (college graduate or higher).

**Table 2.8** *Education level and belief in one true love*

| | HS | Some | College |
|---|---|---|---|
| Agree | 363 | 176 | 196 |
| Disagree | 557 | 466 | 789 |
| Don't know | 20 | 26 | 32 |

(a) Create a new two-way table with row and column totals added.

(b) Find the percent who agree that there is only one true love, for each education level. Does there seem to be an association between education level and agreement with the statement? If so, in what direction?

(c) What percent of people participating in the survey have a college degree or higher?

(d) What percent of the people who disagree with the statement have a high school degree or less?

[10] Hamscher, G., et. al.,"Antibiotics in dust originating from a pig-fattening farm: A new source of health hazard for farmers?" *Environmental Health Perspectives*, October 2003; 111(13): 1590–1594.

[11] "Nearly Half of American Adults are Smartphone Owners," Pew Research Center, pewresearch.org, March 1, 2012.

[12] "Dog Detects Bowel Cancer," CNN Health Online, January 31, 2011.

**2.19 Who Smokes More: Males or Females** The **StudentSurvey** dataset includes variables on gender and on whether or not the student smokes. Who smokes more: males or females? Table 2.9 shows a two-way table of these two variables.

**Table 2.9** *Smoking habits by gender*

| | Female | Male | Total |
|---|---|---|---|
| Don't smoke | 153 | 166 | 319 |
| Smoke | 16 | 27 | 43 |
| Total | 169 | 193 | 362 |

(a) Which gender has a higher percentage of smokers: males or females?

(b) What is the proportion of smokers for the entire sample?

(c) What proportion of the smokers in the sample are female?

**2.20 Is There a Genetic Marker for Dyslexia?** A disruption of a gene called *DYXC1* on chromosome 15 for humans may be related to an increased risk of developing dyslexia. Researchers[13] studied the gene in 109 people diagnosed with dyslexia and in a control group of 195 others who had no learning disorder. The *DYXC1* break occurred in 10 of those with dyslexia and in 5 of those in the control group.

(a) Is this an experiment or an observational study? What are the variables?

(b) How many rows and how many columns will the data table have? Assume rows are the cases and columns are the variables. (There might be an extra column for identification purposes; do not count this column in your total.)

(c) Display the results of the study in a two-way table.

(d) To see if there appears to be a substantial difference between the group with dyslexia and the control group, compare the proportion of each group who have the break on the *DYXC1* gene.

(e) Does there appear to be an association between this genetic marker and dyslexia for the people in this sample? (We will see in Chapter 4 whether we can generalize this result to the entire population.)

(f) If the association appears to be strong, can we assume that the gene disruption causes dyslexia? Why or why not?

[13] *Science News*, August 30, 2003, p 131.

**2.21 Near-Death Experiences** People who have a brush with death occasionally report experiencing a near-death experience, which includes the sensation of seeing a bright light or feeling separated from one's body or sensing time speeding up or slowing down. Researchers[14] interviewed 1595 people admitted to a hospital cardiac care unit during a recent 30-month period. Patients were classified as cardiac arrest patients (in which the heart briefly stops after beating unusually quickly) or patients suffering other serious heart problems (such as heart attacks). The study found that 27 individuals reported having had a near-death experience, including 11 of the 116 cardiac arrest patients. Make a two-way table of the data. Compute the appropriate percentages to compare the rate of near-death experiences between the two groups. Describe the results.

**2.22 Painkillers and Miscarriage** A recent study[15] examined the link between miscarriage and the use of painkillers during pregnancy. Scientists interviewed 1009 women soon after they got positive results from pregnancy tests about their use of painkillers around the time of conception or in the early weeks of pregnancy. The researchers then recorded which of the pregnancies were successfully carried to term. The results are in Table 2.10.

**Table 2.10** *Does the use of painkillers increase the risk of miscarriage?*

| | Miscarriage | Total |
|---|---|---|
| Aspirin | 5 | 22 |
| Ibuprofen | 13 | 53 |
| Acetaminophen | 24 | 172 |
| No painkiller | 103 | 762 |
| Total | 145 | 1009 |

(a) What percent of the pregnancies ended in miscarriage?

(b) Compute the percent of miscarriages for each of the four groups. Discuss the results.

(c) Is this an experiment or an observational study? Describe how confounding variables might affect the results.

(d) Aspirin and ibuprofen belong to a class of medications called nonsteroidal anti-inflammatory

[14] Greyson, B., "Incidence and correlates of near-death experiences on a cardiac care unit", *General Hospital Psychiatry*, July/August 2003; 25: 269–276.

[15] Li, D-K., et. al., "Use of NSAIDs in pregnancy increases risk of miscarriage", *British Medical Journal*, August 16, 2003; 327(7411): 1.

drugs, or NSAIDs. What percent of women taking NSAIDs miscarried? Does the use of NSAIDs appear to increase the risk of miscarrying? Does the use of acetominophen appear to increase the risk? What advice would you give pregnant women?

(e) Is Table 2.10 a two-way table? If not, construct one for these data.

(f) What percent of all women who miscarried had taken no painkillers?

**2.23 Electrical Stimulation for Fresh Insight?** If we have learned to solve problems by one method, we often have difficulty bringing new insight to similar problems. However, electrical stimulation of the brain appears to help subjects come up with fresh insight. In a recent experiment[16] conducted at the University of Sydney in Australia, 40 participants were trained to solve problems in a certain way and then asked to solve an unfamiliar problem that required fresh insight. Half of the participants were randomly assigned to receive non-invasive electrical stimulation of the brain while the other half (control group) received sham stimulation as a placebo. The participants did not know which group they were in. In the control group, 20% of the participants successfully solved the problem while 60% of the participants who received brain stimulation solved the problem.

(a) Is this an experiment or an observational study? Explain.

(b) From the description, does it appear that the study is double-blind, single-blind, or not blind?

(c) What are the variables? Indicate whether each is categorical or quantitative.

(d) Make a two-way table of the data.

(e) What percent of the people who correctly solved the problem had the electrical stimulation?

(f) Give values for $\hat{p}_E$, the proportion of people in the electrical stimulation group to solve the problem, and $\hat{p}_S$, the proportion of people in the sham simulation group to solve the problem. What is the difference in proportions $\hat{p}_E - \hat{p}_S$?

(g) Does electrical stimulation of the brain appear to help insight?

**2.24 Does It Pay to Get a College Degree?** The Bureau of Labor Statistics[17] in the US tells us that, in 2010, the unemployment rate for high school graduates with no college degree is 9.7% while the unemployment rate for college graduates with a bachelor's degree is only 5.2%. Find the difference in proportions of those unemployed between these two groups and give the correct notation for the difference, with a minus sign. Since the data come from the census, you can assume that the values are from a population rather than a sample. Use the correct notation for population proportions, and use subscripts on the proportions to identify the two groups.

**2.25 Smoking and Pregnancy Rate** Studies have concluded that smoking while pregnant can have negative consequences, but could smoking also negatively affect one's ability to become pregnant? A study collected data on 678 women who had gone off birth control with the intention of becoming pregnant.[18] Smokers were defined as those who smoked at least one cigarette a day prior to pregnancy. We are interested in the pregnancy rate during the first cycle off birth control. The results are summarized in Table 2.11.

**Table 2.11** *Smoking and pregnancy rate*

| | Smoker | Non-smoker | Total |
|---|---|---|---|
| Pregnant | 38 | 206 | 244 |
| Not pregnant | 97 | 337 | 434 |
| Total | 135 | 543 | 678 |

(a) Is this an experiment or an observational study? Can we use the data to determine whether smoking influences one's ability to get pregnant? Why or why not?

(b) What is the population of interest?

(c) What is the proportion of women successfully pregnant after their first cycle ($\hat{p}$)? Proportion of smokers successful ($\hat{p}_s$)? Proportion of non-smokers successful ($\hat{p}_{ns}$)?

(d) Find and interpret ($\hat{p}_{ns} - \hat{p}_s$) the difference in proportion of success between non-smokers and smokers.

**National College Health Assessment Survey** Exercises 2.26 to 2.29 use data on college students collected from the American College Health Association–National College Health Assessment survey[19] conducted in Fall 2011. The survey was

[16]Chi, R. and Snyder, A., "Facilitate Insight by Non-Invasive Brain Stimulation", *PLoS ONE*, 2011; 6(2).
[17]Thompson, D., "What's More Expensive than College? Not Going to College", *The Atlantic*, March 27, 2012.
[18]Baird, D. and Wilcox, A., "Cigarette Smoking Associated With Delayed Conception", *Journal of the American Medical Association*, June 2011; 305(23): 2379–2484.
[19]www.acha-ncha.org/docs/ACHA-NCHA-II_ReferenceGroup_DataReport_Fall2011.pdf.

administered at 44 colleges and universities representing a broad assortment of types of schools and representing all major regions of the country. At each school, the survey was administered to either all students or a random sample of students, and more than 27,000 students participated in the survey.

**2.26 Emotionally Abusive Relationships** Students in the ACHA-NCHA survey were asked "Within the last 12 months, have you been in a relationship (meaning an intimate/coupled/partnered relationship) that was emotionally abusive?" The results are given in Table 2.12.

**Table 2.12** *Have you been in an emotionally abusive relationship?*

| | Male | Female | Total |
|---|---|---|---|
| No | 8,352 | 16,276 | 24,628 |
| Yes | 593 | 2,034 | 2,627 |
| Total | 8,945 | 18,310 | 27,255 |

(a) What percent of all respondents have been in an emotionally abusive relationship?

(b) What percent of the people who have been in an emotionally abusive relationship are male?

(c) What percent of males have been in an emotionally abusive relationship?

(d) What percent of females have been in an emotionally abusive relationship?

**2.27 Binge Drinking** Students in the ACHA-NCHA survey were asked "Within the last two weeks, how many times have you had five or more drinks of alcohol at a sitting?" The results are given in Table 2.13.

**Table 2.13** *In the last two weeks, how many times have you had five or more drinks of alcohol?*

| | Male | Female | Total |
|---|---|---|---|
| 0 | 5,402 | 13,310 | 18,712 |
| 1– 2 | 2,147 | 3,678 | 5,825 |
| 3– 4 | 912 | 966 | 1,878 |
| 5 + | 495 | 358 | 853 |
| Total | 8,956 | 18,312 | 27,268 |

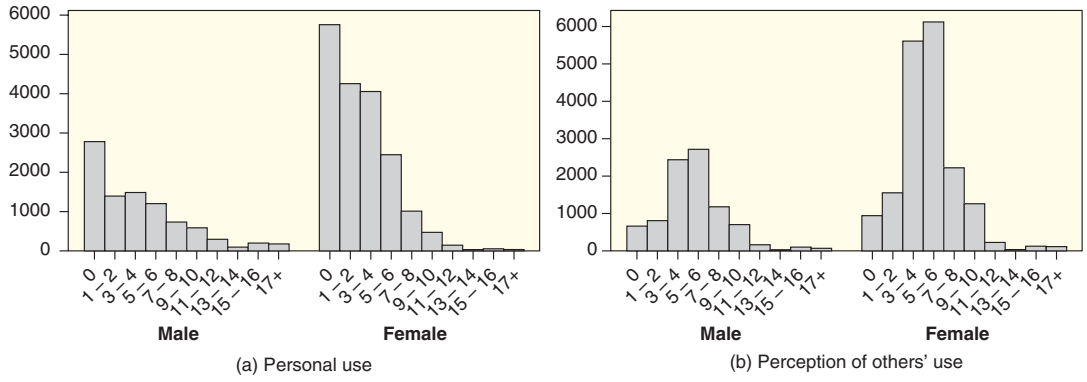
(a) What percent of all respondents answered zero?

(b) Of the students who answered five or more days, what percent are male?

(c) What percent of males report having five or more drinks at a sitting on three or more days in the last two weeks?

(d) What percent of females report having five or more drinks at a sitting on three or more days in the last two weeks?

**2.28 How Accurate are Student Perceptions?** Students in the ACHA-NCHA survey were asked two questions about alcohol use, one about their own personal consumption of alcohol and one about their perception of other students' consumption of alcohol. Figure 2.3(a) shows side-by-side bar charts for responses to the question "The last time you 'partied'/socialized, how many drinks of alcohol did you have?" while Figure 2.3(b) shows side-by-side bar charts for responses to the question "How many drinks of alcohol do you think the typical student at your school had the last time he/she 'partied'/socialized?"




(a) Personal use

(b) Perception of others' use

**Figure 2.3** *How many drinks of alcohol?*

(a) What is the most likely response for both males and females when asked about their own personal alcohol use?

(b) What is the most likely response for both males and females when asked about alcohol use of a "typical student"?

(c) Do students' perceptions of "typical" alcohol use match reality? (This phenomenon extends what we learned about the inability of students to select unbiased samples in Chapter 1. In this case, students tend to notice heavy drinkers disproportionately.)

**2.29 Does Stress Affect Academic Performance?** Students in the ACHA-NCHA survey were asked "Within the last 12 months, has stress negatively affected your academics?" Figure 2.4(a) shows a segmented bar chart for response frequencies while Figure 2.4(b) shows a segmented bar chart for response relative frequencies as percents. Possible responses were "I haven't had any stress," shown in brown, "I've had stress but it hasn't hurt my grades," shown in green, or "I've had stress and it has hurt my grades," shown in blue.

(d) Did a greater percent of males or females say that stress affected their grades, or is it approximately equal between males and females? Is graph (a) or (b) more helpful to answer this question?

**2.30 Vaccine for Malaria** In order for a vaccine to be effective, it should reduce a person's chance of acquiring a disease. Consider a hypothetical vaccine for malaria - a tropical disease that kills between 1.5 and 2.7 million people every year.[20] Suppose the vaccine is tested with 500 volunteers in a village who are malaria-free at the beginning of the trial. Two hundred of the volunteers will get the experimental vaccine and the rest will not be vaccinated. Suppose that the chance of contracting malaria is 10% for those who are not vaccinated. Construct a two-way table to show the results of the experiment if:

(a) The vaccine has no effect.

(b) The vaccine cuts the risk of contracting malaria in half.

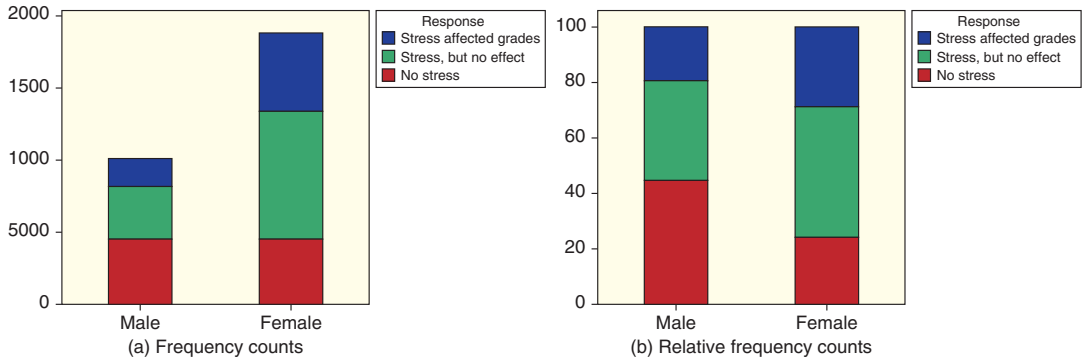



**Figure 2.4** ***Has stress hurt your grades?***

(a) Did more males or more females answer the survey, or did approximately equal numbers of males and females participate? Is graph (a) or (b) more helpful to answer this question?

(b) Did a greater *number* of males or females say they had no stress, or is it approximately equal between males and females? Is graph (a) or (b) more helpful to answer this question?

(c) Did a greater *percent* of males or females say they had no stress, or is it approximately equal between males and females? Is graph (a) or (b) more helpful to answer this question?

**2.31 Which of These Things Is Not Like the Other?** Four students were working together on a project and one of the parts involved making a graph to display the relationship in a two-way table of data with two categorical variables: college accept/reject decision and type of high school (public, private, parochial). The graphs submitted by each student are shown in Figure 2.5. Three are from the same data, but one is inconsistent with the other three. Which is the bogus graph? Explain.

[20]World Health Organization.

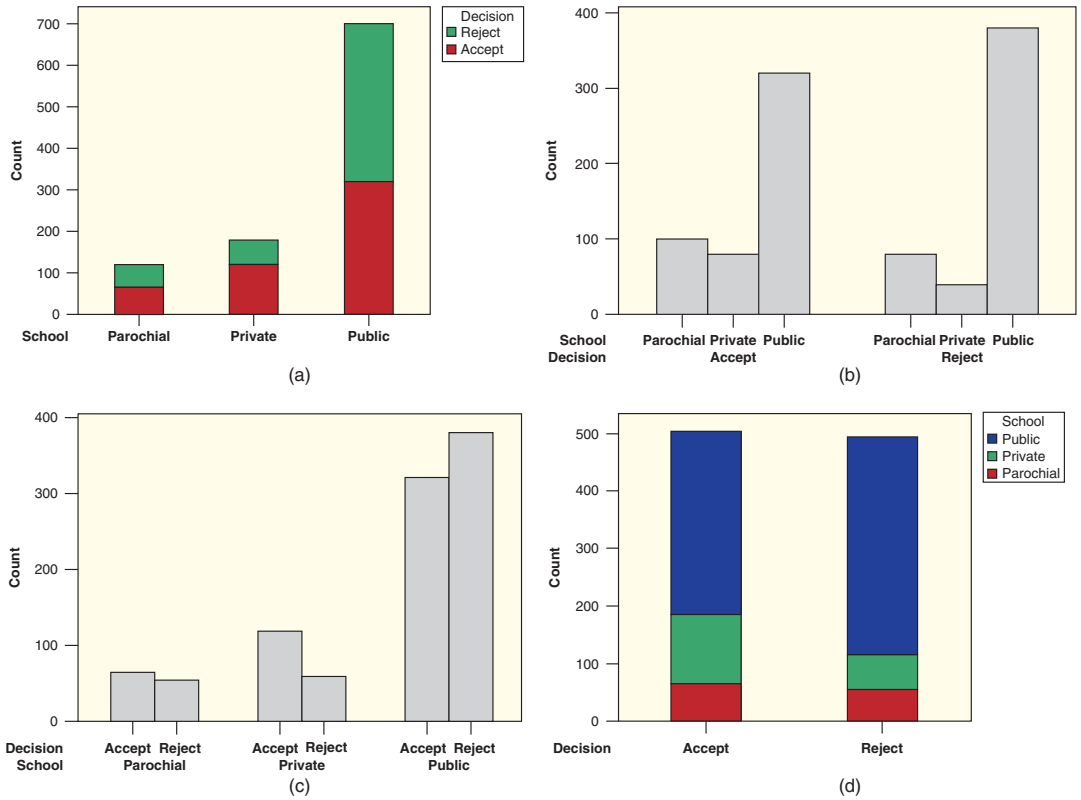



**Figure 2.5** *Views of the same two-way table – with one error*

# 2.2 ONE QUANTITATIVE VARIABLE: SHAPE AND CENTER

In Section 2.1, we see how to describe categorical variables. In this section, we begin to investigate quantitative variables. In describing a single quantitative variable, we generally consider the following three questions:

- What is the general *shape* of the data?
- Where are the data values *centered*?
- How do the data *vary*?

These are all aspects of what we call the *distribution* of the data. In this section, we focus on the first two questions and leave the third question, on variability, to Section 2.3.

## The Shape of a Distribution

We begin by looking at graphical displays as a way of understanding the shape of a distribution. A common way to visualize the shape of a moderately sized dataset is a *dotplot*. We create a dotplot by using an axis with a scale appropriate for the numbers in the dataset and placing a dot over the axis for each case in the dataset. If there are multiple data values that are the same, we stack the dots vertically. To illustrate a dotplot, we look at some data on the typical lifespan for several mammals.

**SECTION LEARNING GOALS**

*You should now have the understanding and skills to:*

- Use a dotplot or histogram to describe the shape of a distribution
- Calculate the mean and the median for a set of data values, with appropriate notation
- Identify the approximate locations of the mean and the median on a dotplot or histogram
- Explain how outliers and skewness affect the values for the mean and median

## Exercises for Section 2.2

**SKILL BUILDER 1**

Exercises 2.32 to 2.38 refer to histograms A through H in Figure 2.12.

**2.32** Which histograms are skewed to the left?

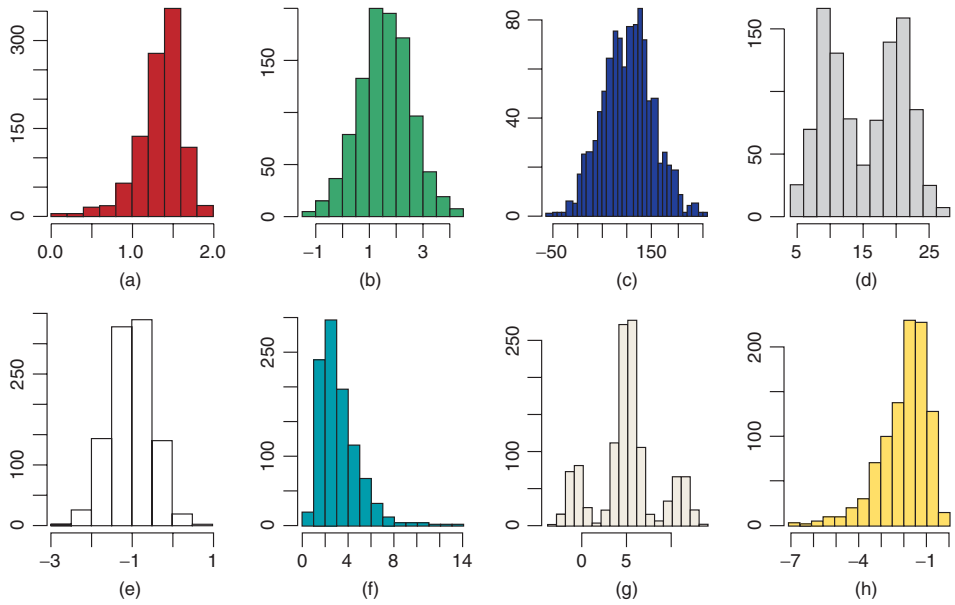



**Figure 2.12** *Eight histograms*

**2.33** Which histograms are skewed to the right?

**2.34** Which histograms are approximately symmetric?

**2.35** Which histograms are approximately symmetric and bell-shaped?

**2.36** For each of the four histograms A, B, C, and D, state whether the mean is likely to be larger than the median, smaller than the median, or approximately equal to the median.

**2.37** For each of the four histograms E, F, G, and H, state whether the mean is likely to be larger than the median, smaller than the median, or approximately equal to the median.

**2.38** Which of the distributions is likely to have the largest mean? The smallest mean?

### SKILL BUILDER 2

In Exercises 2.39 to 2.42, draw any dotplot to show a dataset that is

**2.39** Clearly skewed to the left

**2.40** Approximately symmetric and bell-shaped

**2.41** Approximately symmetric but not bell-shaped

**2.42** Clearly skewed to the right

### SKILL BUILDER 3

For each set of data in Exercises 2.43 to 2.46:

(a) Find the mean $\bar{x}$.

(b) Find the median $m$.

(c) Indicate whether there appear to be any outliers. If so, what are they?

**2.43** 8, 12, 3, 18, 15

**2.44** 41, 53, 38, 32, 115, 47, 50

**2.45** 15, 22, 12, 28, 58, 18, 25, 18

**2.46** 110, 112, 118, 119, 122, 125, 129, 135, 138, 140

### SKILL BUILDER 4

In Exercises 2.47 to 2.50, give the correct notation for the mean.

**2.47** The average number of calories eaten in one day is 2386 calories for a sample of 100 participants.

**2.48** The average number of text messages sent or received in a day was 60, in a survey of $n = 799$ teen cell phone users[26] conducted in June 2011.

**2.49** The average number of yards per punt for all punts in the National Football League is 41.5 yards.[27]

**2.50** The average number of television sets owned per household for all households in the US is 2.6.[28]

**2.51 Arsenic in Toenails** Arsenic is toxic to humans, and people can be exposed to it through contaminated drinking water, food, dust, and soil. Scientists have devised an interesting new way to measure a person's level of arsenic poisoning: by examining toenail clippings. In a recent study,[29] scientists measured the level of arsenic (in mg/kg) in toenail clippings of eight people who lived near a former arsenic mine in Great Britain. The following levels were recorded:

0.8 1.9 2.7 3.4 3.9 7.1 11.9 26.0

(a) Do you expect the mean or the median of these toenail arsenic levels to be larger? Why?

(b) Calculate the mean and the median.

**2.52 Normal Body Temperature** It is commonly believed that normal human body temperature is 98.6°F (or 37°C). In fact, "normal" temperature can vary from person to person, and for a given person it can vary over the course of a day. Table 2.16 gives a set of temperature readings of a healthy woman taken over a two-day period.

**Table 2.16** *Body temperature during the day*

| | | | | | |
|---|---|---|---|---|---|
| 97.2 | 97.6 | 98.4 | 98.5 | 98.3 | 97.7 |
| 97.3 | 97.7 | 98.5 | 98.5 | 98.4 | 97.9 |

(a) Make a dotplot of the data.

(b) Compute the mean of the data and locate it on the dotplot as the balance point.

(c) Compute the median of the data and locate it on the dotplot as the midway point.

**2.53 Population of States in the US** The dataset **USStates** has a great deal of information about the 50 states, including population. Figure 2.13 shows a histogram of the population, in millions, of the 50 states in the US.

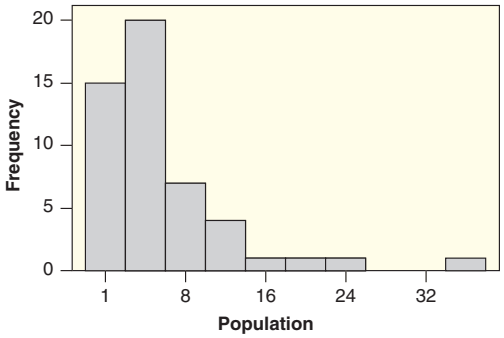



**Figure 2.13** *Population, in millions, of the 50 states*

(a) Do these values represent a population or a sample?

(b) Describe the shape of the distribution: Is it approximately symmetric, skewed to the right, skewed to the left, or none of these? Are there any outliers?

[26]"Teens, Smartphones, and Texting", Pew Research Center, pewresearch.org, March 19, 2012.

[27]Moskowitz, T. and Wertheim, L., *Scorecasting*, Crown Archetype, New York, 2011, p. 119.

[28]www.census.gov.

[29]Button, M., Jenkin, G., Harrington, C. and Watts, M., "Human Toenails as a biomarker of exposure to elevated environment arsenic", *Journal of Environmental Monitoring*, March 2009; 11(3): 610-617. Data are reproduced from summary statistics and are approximate.

(c) Estimate the median population.

(d) Estimate the mean population.

**2.54 Insect Weights** Consider a dataset giving the adult weight of species of insects. Most species of insects weigh less than 5 grams, but there are a few species that weigh a great deal, including the largest insect known: the rare and endangered Giant Weta from New Zealand, which can weigh as much as 71 grams. Describe the shape of the distribution of weights of insects. Is it symmetric or skewed? If it is skewed, is it skewed to the left or skewed to the right? Which will be larger, the mean or the median?

**2.55 Is Language Biased toward Happiness?** "Are natural languages neutrally, positively, or negatively biased?" That is the question a recent study[30] set out to answer. They found the top 5000 words used in English in each of four different places: Twitter, books on the Google Book Project, *The New York Times*, and music lyrics. The resulting complete list was 10,222 unique words in the English language. Each word was then evaluated independently by 50 different people, each giving a rating on how the word made them feel on a 1 to 9 scale where 1 = least happy, 5 = neutral, and 9 = most happy. (The highest rated word was "laughter" while the lowest was "terrorist.") The distributions of the ratings for all 10,222 words for each of the four media sources were surprisingly similar, and all had approximately the shape shown in Figure 2.14.

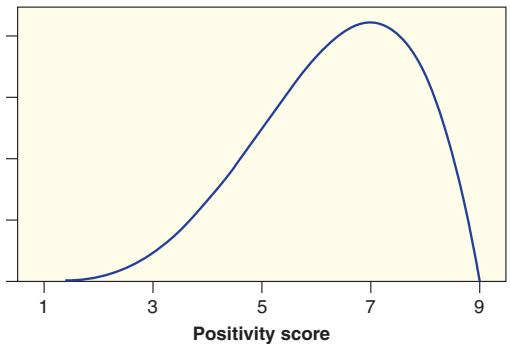



**Figure 2.14** ***Distribution of ratings of words where 9 = most positive***

(a) Describe the shape of the distribution.

(b) Which of the following values is closest to the median of the distribution:

3.5 5 6.5 7 7.5 8

(c) Will the mean be smaller or larger than the value you gave for the median in part (b)?

**2.56 Life Expectancy** Life expectancy for all the different countries in the world ranges from a low of only 43.9 years (in Afghanistan) to a high of 82.8 years (in San Marino). Life expectancies are clustered at the high end, with about half of all the countries having a life expectancy between about 72 and the maximum of 82.8. A few countries, such as Afghanistan, have a very low life expectancy. The full dataset is in **AllCountries**.

(a) What is the shape of the distribution of life expectancies for all countries?

(b) From the information given, estimate the median of the life expectancies.

(c) Will the mean be larger or smaller than the median?

**2.57 Fiber in the Diet** The number of grams of fiber eaten in one day for a sample of ten people are

10 11 11 14 15 17 21 24 28 115

(a) Find the mean and the median for these data.

(b) The value of 115 appears to be an obvious outlier. Compute the mean and the median for the nine numbers with the outlier excluded.

(c) Comment on the effect of the outlier on the mean and on the median.

**2.58 Beta-Carotene Levels in the Blood** The plasma beta-carotene level (concentration of beta-carotene in the blood), in ng/ml, was measured for a sample of $n = 315$ individuals, and the results[31] are shown in the histogram in Figure 2.15.

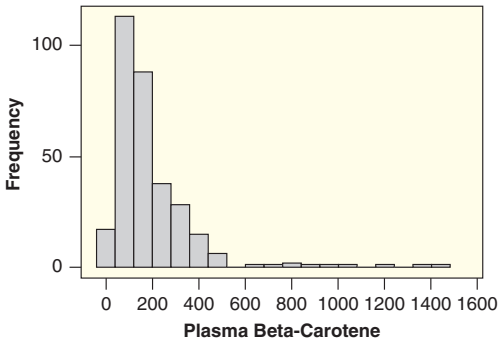



**Figure 2.15** ***Concentration of beta-carotene in the blood***

[30] Kloumann, I., Danforth, C., Harris, K., Bliss, C. and Dodds, P., "Positivity of the English Language," *PLoS ONE*, 2012; 7(1).

[31] http://lib.stat.cmu.edu/datasets/Plasma_Retinol, accessed November 24, 2003.

(a) Describe the shape of this distribution. Is it symmetric or skewed? Are there any obvious outliers?

(b) Estimate the median of this sample.

(c) Estimate the mean of this sample.

**2.59 Number of Text Messages per Day** A survey conducted in May 2010 asked 1,917 cell phone users to estimate, on average, the number of text messages sent and received per day.

(a) Do you expect the distribution of number of text messages per day to be symmetric, skewed to the right, or skewed to the left?

(b) Two measures of center for this distribution are 10 messages and 39.1 messages.[32] Which is most likely to be the mean and which is most likely to be the median? Explain your reasoning.

**2.60 Time Spent Exercising, between Males and Females** Often we are interested not just in a single mean but in a difference in means between two groups. In the **StudentSurvey** data, there are 36 seniors: 26 males and 10 females. Table 2.17 gives the number of hours per week that each said he or she spent exercising.

**Table 2.17** *Number of hours spent exercising a week*

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Females | 4 | 2 | 5 | 6 | 12 | 15 | 10 |
| | 5 | 0 | 5 | | | | |
| Males | 10 | 10 | 6 | 5 | 7 | 8 | 4 |
| | 12 | 12 | 4 | 15 | 10 | 5 | 5 |
| | 2 | 2 | 7 | 3 | 5 | 15 | 6 |
| | 6 | 5 | 0 | 8 | 5 | | |

(a) Calculate $\bar{x}_f$, the mean number of hours spent exercising by the females.

(b) Calculate $\bar{x}_m$, the mean number of hours spent exercising by the males.

(c) Compute the difference, $\bar{x}_m - \bar{x}_f$, and interpret it in context.

**2.61 Does It Pay to Get a College Degree?** In Exercise 2.24 on page 57, we saw that those with a college degree were much more likely to be employed. The same article also gives statistics on earnings in the US in 2009 by education level. The median weekly earnings for high school graduates with no college degree was $626, while the median weekly earnings for college graduates with a bachelor's degree was $1025. Give correct notation for and find the difference in medians, using the notation for a median, subscripts to identify the two groups, and a minus sign.

**2.62 Does Sexual Frustration Increase the Desire for Alcohol?** Apparently, sexual frustration increases the desire for alcohol, at least in fruit flies. Scientists[33] randomly put 24 fruit flies into one of two situations. The 12 fruit flies in the "mating" group were allowed to mate freely with many available females eager to mate. The 12 in the "rejected" group were put with females that had already mated and thus rejected any courtship advances. After four days of either freely mating or constant rejection, the fruit flies spent three days with unlimited access to both normal fruit fly food and the same food soaked in alcohol. The percent of time each fly chose the alcoholic food was measured. The fruit flies that had freely mated chose the two types of food about equally often, choosing the alcohol variety on average 47% of the time. The rejected males, however, showed a strong preference for the food soaked in alcohol, selecting it on average 73% of the time. (The study was designed to study a chemical in the brain called neuropeptide that might play a role in addiction.)

(a) Is this an experiment or an observational study?

(b) What are the cases in this study? What are the variables? Which is the explanatory variable and which is the response variable?

(c) We are interested in the difference in means, where the means measure the average percent preference for alcohol (0.47 and 0.73 in this case). Find the difference in means and give the correct notation for your answer, using the correct notation for a mean, subscripts to identify groups, and a minus sign.

(d) Can we conclude that rejection increases a male fruit fly's desire for alcohol? Explain.

**2.63 Create a Dataset** Give any set of five numbers satisfying the condition that:

(a) The mean of the numbers is substantially less than the median.

(b) The mean of the numbers is substantially more than the median.

(c) The mean and the median are equal.

**2.64 Describe a Variable** Describe one quantitative variable that you believe will give data that are skewed to the right, and explain your reasoning. Do not use a variable that has already been discussed.

[32] Lenhard, A., "Cell Phones and American Adults," Pew Research Center, pewresearch.org, September 2, 2010.

[33] Shohat-Ophir, G., Kaun, K., Azanchi, R. and Heberlein, U., "Sexual Deprivation Increases Ethanol Intake in *Drosophila*," *Science*, 16 March 2012; 335(6074): 1351-1355.

**2.65 Mean or Median** Calculate the mean and the median for the numbers

1, 1, 1, 1, 1, 1, 2, 5, 7, 12

Which do you think is a better measure of center for this set of values? Why? (There is no right answer, but think about which you would use.)

**2.66 Number of Children** Table 2.18 shows the number of women (per 1000) between 15 and 44 years of age who have been married grouped by the number of children they have had. Table 2.19 gives the same information for women who have never been married.[34]

(a) Without doing any calculations, which of the two samples appears to have the highest mean number of children? Which of the distributions appears to have the mean most different from the median? Why?

(b) Find the median for each data set.

[34]Bachu, A., Current Population Reports, P20-499, Fertility of AmericanWomen (June 1995 Update), issued October 1997, obtained from www.census.gov.

**Table 2.18** *Women who have been married*

| Number of Children | Women per 1000 |
|---|---|
| 0 | 162 |
| 1 | 190 |
| 2 | 290 |
| 3 | 289 |
| 4 | 48 |
| 5+ | 21 |

**Table 2.19**
*Women who have never been married*

| Number of Children | Women per 1000 |
|---|---|
| 0 | 791 |
| 1 | 108 |
| 2 | 53 |
| 3 | 29 |
| 4 | 12 |
| 5+ | 7 |

# 2.3 ONE QUANTITATIVE VARIABLE: MEASURES OF SPREAD

So far, we have looked at two important summary statistics for a single quantitative variable: the mean and the median. Although there are important differences between them, both of these measurements tell us something about the "middle" or "center" of a dataset. When we give a statistical summary of the values in a dataset, we are interested in not just the center of the data but also how spread out the data are. Knowing that the average high temperature in Des Moines, Iowa, in April is 62°F is helpful, but it is also helpful to know that the historical range is between 8°F and 97°F! In this section, we examine additional measures of location and measures of spread.

### Using Technology to Compute Summary Statistics

In practice, we generally use statistical software or a graphing calculator to compute the summary statistics for a dataset. For assistance in using a wide variety of different types of technology and software, see the available supplementary resources.

**Example 2.15**

*Des Moines vs San Francisco Temperatures*

Average temperature on April 14th for the 16 years ending in 2010 is given in Table 2.20 for Des Moines, Iowa, and San Francisco, California.[26] Use technology and the data in **April14Temps** to find the mean and the median temperature on April 14th for each city.

[26]www.weather.com.

have the advantage that they use all of the data values. However, they are not resistant to outliers. The median and IQR are resistant to outliers. Furthermore, if there are outliers or the data are heavily skewed, the five-number summary can give more information (such as direction of skewness) than the mean and standard deviation.

**Example 2.24**

Example 2.13 on page 67 describes salaries in the US National Football League, in which some star players are paid much more than most other players.

(a) We see in that example that players prefer to use the median ($838,000 in 2010) as a measure of center since they don't want the results heavily influenced by a few huge outlier salaries. What should they use as a measure of spread?

(b) We also see that the owners of the teams prefer to use the mean ($1.87 million in 2010) as a measure of center since they want to use a measure that includes all the salaries. What should they use as a measure of spread?

*Solution*

(a) The interquartile range (IQR) should be used with the median as a measure of spread. Both come from the five number summary, and both the median and the IQR are resistant to outliers.

(b) The standard deviation should be used with the mean as a measure of spread. Both the mean and the standard deviation use all the data values in their computation.

**SECTION LEARNING GOALS**

*You should now have the understanding and skills to:*

- Use technology to compute summary statistics for a quantitative variable
- Recognize the uses and meaning of the standard deviation
- Compute and interpret a $z$-score
- Interpret a five number summary or percentiles
- Use the range, the interquartile range, and the standard deviation as measures of spread
- Describe the advantages and disadvantages of the different measures of center and spread

## Exercises for Section 2.3

**SKILL BUILDER 1**

For the datasets in Exercises 2.67 to 2.72, use technology to find the following values:

(a) The mean and the standard deviation

(b) The five number summary

**2.67** 10, 11, 13, 14, 14, 17, 18, 20, 21, 25, 28

**2.68** 1, 3, 4, 5, 7, 10, 18, 20, 25, 31, 42

**2.69** 4, 5, 8, 4, 11, 8, 18, 12, 5, 15, 22, 7, 14, 11, 12

**2.70** 25, 72, 77, 31, 80, 80, 64, 39, 75, 58, 43, 67, 54, 71, 60

**2.71** The variable *Exercise*, number of hours spent exercising per week, in the **StudentSurvey** dataset

**2.72** The variable *TV*, number of hours spent watching television per week, in the **StudentSurvey** dataset

## SKILL BUILDER 2

In Exercises 2.73 and 2.74, match the standard deviations with the histograms.

**2.73** Match the three standard deviations $s = 1$, $s = 3$, and $s = 5$ with the three histograms in Figure 2.22.

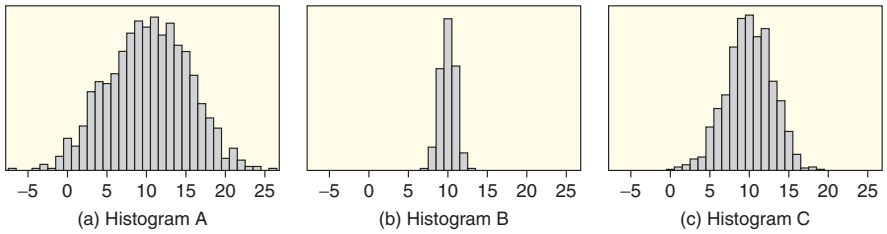



**Figure 2.22** *Three histograms for Exercise 2.73*

**2.74** Match each standard deviation with one of the histograms in Figure 2.23.

(a) $s = 0.5$
(b) $s = 10$
(c) $s = 50$
(d) $s = 1$
(e) $s = 1000$
(f) $s = 0.29$

## SKILL BUILDER 3

In Exercises 2.75 and 2.76, match each five number summary with the corresponding histogram.

**2.75** Match each five number summary with one of the histograms in Figure 2.23.

(a) (0, 0.25, 0.5, 0.75, 1)
(b) (−1.08, −0.30, 0.01, 0.35, 1.27)
(c) (0.64, 27.25, 53.16, 100, 275.7)
(d) (−3.5, −0.63, −0.11, 0.59, 2.66)
(e) (71.45, 92.77, 99.41, 106.60, 129.70)
(f) (−1296, −1005, −705, 998, 1312)

**2.76** Match each five number summary with one of the histograms in Figure 2.24. The scale is the same on all four histograms.

(a) (1, 3, 5, 7, 9)
(b) (1, 4, 5, 6, 9)
(c) (1, 5, 7, 8, 9)
(d) (1, 1, 2, 4, 9)

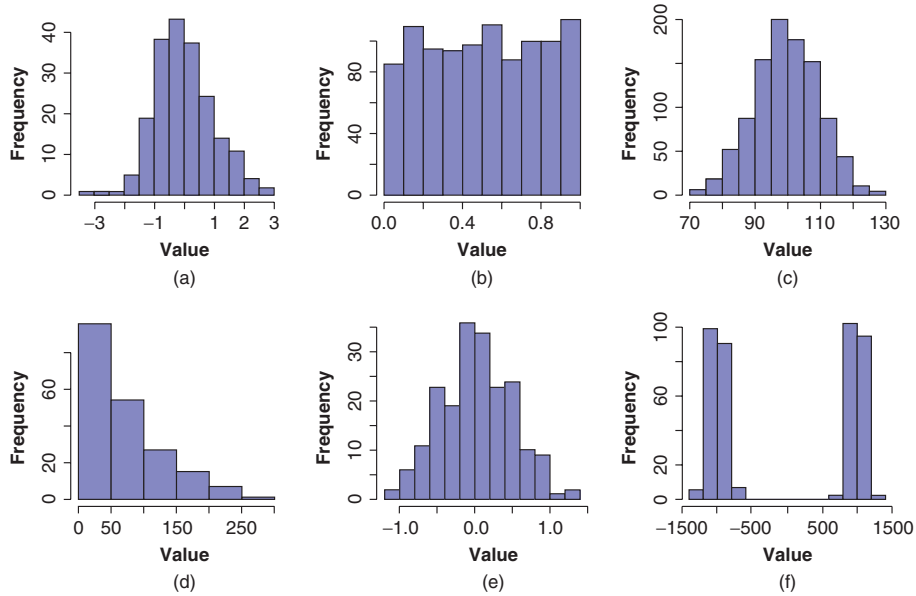



**Figure 2.23** *Histograms for Exercises 2.74 and 2.75*

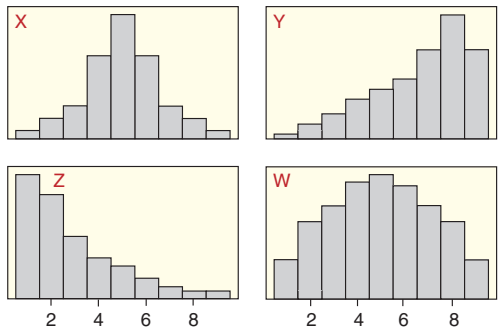



**Figure 2.24** *Match five number summaries in Exercise 2.76*

## SKILL BUILDER 4

In Exercises 2.77 to 2.82, estimate the summary statistics requested, using the histogram in Figure 2.25 for Exercises 2.77 to 2.79 and the dotplot in Figure 2.26 for Exercises 2.80 to 2.82. There are $n = 100$ data points included in the dotplot.

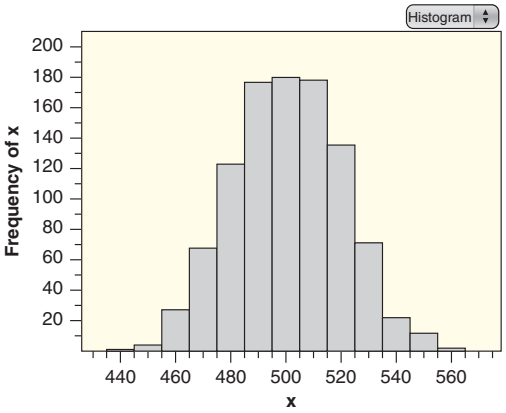



**Figure 2.25** *Histogram for Exercises 2.77 to 2.79*

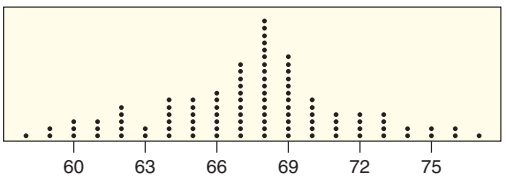



**Figure 2.26** *Dotplot with $n = 100$ for Exercises 2.80 to 2.82*

**2.77** Estimate the mean and the standard deviation for the data in the histogram in Figure 2.25.

**2.78** Estimate values at the 10th percentile and the 90th percentile for the data in Figure 2.25.

**2.79** Estimate the five number summary for the data in Figure 2.25.

**2.80** Estimate the mean and the standard deviation for the data in the dotplot in Figure 2.26.

**2.81** Estimate values at the 10th percentile and the 90th percentile for the data in Figure 2.26.

**2.82** Estimate the five number summary for the data in Figure 2.26.

## SKILL BUILDER 5

In Exercises 2.83 to 2.86, indicate whether the five number summary corresponds most likely to a distribution that is skewed to the left, skewed to the right, or symmetric.

**2.83** (15, 25, 30, 35, 45)

**2.84** (100, 110, 115, 160, 220)

**2.85** (0, 15, 22, 24, 27)

**2.86** (22.4, 30.1, 36.3, 42.5, 50.7)

## SKILL BUILDER 6: *Z*-SCORES

In Exercises 2.87 to 2.90, find and interpret the *z*-score for the data value given.

**2.87** The value 243 in a dataset with mean 200 and standard deviation 25

**2.88** The value 88 in a dataset with mean 96 and standard deviation 10

**2.89** The value 5.2 in a dataset with mean 12 and standard deviation 2.3

**2.90** The value 8.1 in a dataset with mean 5 and standard deviation 2

## SKILL BUILDER 7: THE 95% RULE

In Exercises 2.87 to 2.90, use the 95% rule and the fact that the summary statistics come from a distribution that is symmetric and bell-shaped to find an interval that is expected to contain about 95% of the data values.

**2.91** A bell-shaped distribution with mean 200 and standard deviation 25

**2.92** A bell-shaped distribution with mean 10 and standard deviation 3

**2.93** A bell-shaped distribution with mean 1000 and standard deviation 10

**2.94** A bell-shaped distribution with mean 1500 and standard deviation 300

**2.95** **Estimating Summary Statistics** For the dataset

45, 46, 48, 49, 49, 50, 50, 52, 52, 54, 57, 57, 58, 58, 60, 61

(a) Without doing any calculations, estimate which of the following numbers is closest to the mean:

60, 53, 47, 58

(b) Without doing any calculations, estimate which of the following numbers is closest to the standard deviation:

52, 5, 1, 10, 55

(c) Use statistics software on a calculator or computer to find the mean and the standard deviation for this dataset.

**2.96 Percent Obese by State** Computer output giving descriptive statistics for the percent of the population that is obese for each of the 50 US states, from the **USStates** dataset, is given in Figure 2.27. Since all 50 US states are included, this is a population not a sample.

(a) What are the mean and the standard deviation? Include appropriate notation with your answers.

(b) Calculate the $z$-score for the largest value and interpret it in terms of standard deviations. Do the same for the smallest value.

(c) This distribution is relatively symmetric and bell-shaped. Give an interval that is likely to contain about 95% of the data values.

**Descriptive Statistics: Obese**

| Variable | N | N* | Mean | SE Mean | StDev | Minimum | Q1 | Median | Q3 | Maximum |
|---|---|---|---|---|---|---|---|---|---|---|
| Obese | 50 | 0 | 24.552 | 0.431 | 3.044 | 17.800 | 22.175 | 24.400 | 26.825 | 30.900 |

**Figure 2.27** *Percent of the population that is obese by state*

**2.97 Five Number Summary for Percent Obese by State** Computer output giving descriptive statistics for the percent of the population that is obese for each of the 50 US states, from the **USStates** dataset, is given in Figure 2.27.

(a) What is the five number summary?

(b) Give the range and the IQR.

(c) What can we conclude from the five number summary about the location of the 15$^{th}$ percentile? The 60$^{th}$ percentile?

**2.98 How Many Hot Dogs Can You Eat in Ten Minutes?** Every Fourth of July, Nathan's Famous in New York City holds a hot dog eating contest, in which contestants try to eat as many hot dogs as possible in ten minutes.[40] In 2011, over 30,000 people watched the event live on Coney Island, and it was broadcast live to many more on ESPN. The winner in 2011 was Joey Chestnut, who downed 62 hot dogs (with buns), for his fifth straight title. Before Joey, the reigning hot dog eating champion was Takeru Kobayashi of Japan. (Although many people compete in the contest, every contest since 2002 has been won by one or the other of these two men.) The winning number of hot dogs along with the year is shown in Table 2.21 and is available in the dataset **HotDogs**.

**Table 2.21** *Winning number of hot dogs in the hot dog eating contest*

| Year | Hot Dogs |
|---|---|
| 2011 | 62 |
| 2010 | 54 |
| 2009 | 68 |
| 2008 | 59 |
| 2007 | 66 |
| 2006 | 54 |
| 2005 | 49 |
| 2004 | 54 |
| 2003 | 45 |
| 2002 | 50 |

(a) Use technology to find the mean and the standard deviation of the ten numbers.

(b) How many of the ten values are above the mean? How many are above the mean for the five values in the earlier five years (2002–2006)? How many are above the mean for the five values in the later five years (2007–2011)?

**2.99 The Hot Dog Eating Rivalry: Matched Pairs** In Exercise 2.98, we mention that either Joey Chestnut of California, US, or Takeru Kobayashi of Japan has won the Nathan's Famous Hot Dog Eating Contest every year from 2002 until 2011. In five of those years, both men competed and the results of the rivalry are shown in Table 2.22. (After the tie in 2008, Joey Chestnut won in an overtime.) Because the conditions of the year matter, this is a *matched pairs* situation, with the two men going against each other each year. In a matched pairs situation, we use the summary statistics of the *differences* between the pairs of values.

[40] nathansfamous.com.

**Table 2.22** *Hot dog eating rivalry*

| Year | Joey Chestnut | Takeru Kobayashi |
|---|---|---|
| 2009 | 68 | 64 |
| 2008 | 59 | 59 |
| 2007 | 66 | 63 |
| 2006 | 52 | 54 |
| 2005 | 32 | 49 |

(a) For each of the five years, find the difference in number of hot dogs eaten between Joey and Takeru. For example, in 2009, the difference is $68 - 64 = 4$. Since it is important to always subtract the same way (in this case, Joey's value minus Takeru's value), some of the differences will be negative.

(b) Use technology to find the mean and the standard deviation of the differences.

**2.100 Time in Days to Row Solo Across the Atlantic Ocean** Exercise 1.20 on page 15 gives a sample of eight times, in days, to row solo across the Atlantic Ocean. The times are

40, 87, 78, 106, 67, 70, 153, 81

(a) Use technology to find the mean and standard deviation of the eight times.

(b) Find and interpret the $z$-scores for the longest time and shortest time in the sample.

**2.101 Laptop Computers and Sperm Count** Studies have shown that heating the scrotum by just 1°C can reduce sperm count and sperm quality, so men concerned about fertility are cautioned to avoid too much time in the hot tub or sauna. A new study[41] suggests that men also keep their laptop computers off their laps. The study measured scrotal temperature in 29 healthy male volunteers as they sat with legs together and a laptop computer on the lap. Temperature increase in the left scrotum over a 60-minute session is given as $2.31 \pm 0.96$ and a note tells us that "Temperatures are given as °C; values are shown as mean ± SD." The abbreviation SD stands for standard deviation. (Men who sit with their legs together without a laptop computer do not show an increase in temperature.)

(a) If we assume that the distribution of the temperature increases for the 29 men is symmetric and bell-shaped, find an interval that we expect to contain about 95% of the temperature increases.

(b) Find and interpret the $z$-score for one of the men, who had a temperature increase of 4.9°.

**2.102 Grade Point Averages** A histogram of the $n = 345$ grade point averages reported by students in the **StudentSurvey** dataset is shown in Figure 2.28.

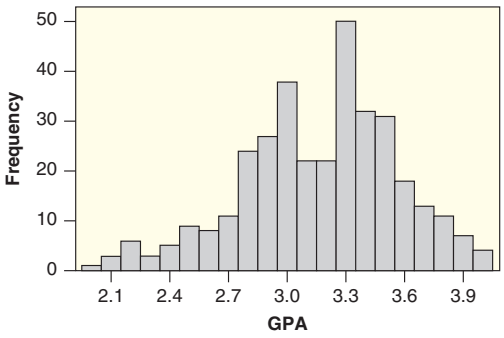



**Figure 2.28** *Estimate the* $10^{th}$ *percentile and* $75^{th}$ *percentile*

(a) Estimate and interpret the $10^{th}$ percentile and the $75^{th}$ percentile.

(b) Estimate the range.

**2.103 Arsenic in Toenails** Exercise 2.51 on page 71 discusses the use of toenail clippings to measure the level of arsenic exposure of individuals in Great Britain. A similar study was conducted in the US. Table 2.23 gives toenail arsenic concentrations (in ppm) for 19 individuals with private wells in New Hampshire, and the data are also available in **ToenailArsenic**. Such concentrations prove to be an effective indicator of ingestion of arsenic-containing water.[42]

**Table 2.23** *Arsenic concentration in toenail clippings*

| | | | | | | |
|---|---|---|---|---|---|---|
| 0.119 | 0.118 | 0.099 | 0.118 | 0.275 | 0.358 | 0.080 |
| 0.158 | 0.310 | 0.105 | 0.073 | 0.832 | 0.517 | 0.851 |
| 0.269 | 0.433 | 0.141 | 0.135 | 0.175 | | |

(a) Use technology to find the mean and standard deviation.

(b) Compute the $z$-score for the largest concentration and interpret it.

[41] Sheynkin, Y., et. al., "Protection from scrotal hyperthermia in laptop computer users", *Fertility and Sterility*, February 2011; 92(2): 647 - 651.

[42] Adapted from Karagas, M., et. al.,"Toenail samples as an indicator of drinking water arsenic exposure", *Cancer Epidemiology, Biomarkers and Prevention*, 1996; 5: 849–852.

(c) Use technology to find the five number summary.

(d) What is the range? What is the IQR?

**2.104 A Dotplot of Arsenic in Toenails** Figure 2.29 shows a dotplot of the arsenic concentrations in Table 2.23.

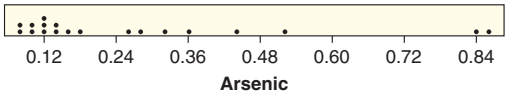



**Figure 2.29** *Dotplot of arsenic concentration in toenails*

(a) Which measures of center and spread are most appropriate for this distribution: the mean and standard deviation or the five number summary? Explain.

(b) Is it appropriate to use the general rule about having 95% of the data within two standard deviations for this distribution? Why or why not?

**STATISTICS FOR NBA PLAYERS IN 2010–2011**

Exercises 2.105 to 2.107 refer to the dataset **NBA Players2011**, which contains information on many variables for players in the NBA (National Basketball Association) during the 2010–2011 season. The dataset includes information for all players who averaged more than 24 minutes per game, and includes $n = 176$ players and 24 variables.

**2.105 Distribution of Three-Point Attempts in the NBA** In basketball, a basket is awarded three points (rather than the usual two) if it is shot from farther away. Some players attempt lots of three-point shots and quite a few attempt none, as we see in the distribution of number of three-point attempts by players in the NBA in Figure 2.30. The data are available in **NBAPlayers2011** under the variable name *FG3Attempt*. Is it appropriate to use the 95% rule with this dataset? Why or why not?

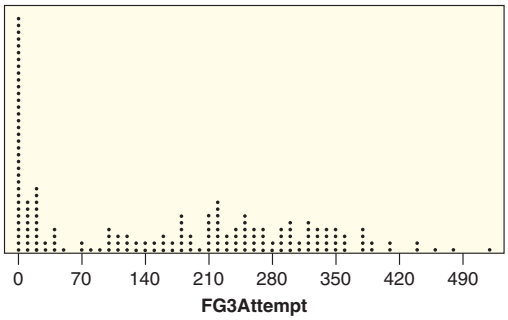



**Figure 2.30** *Number of three-point shot attempts in the NBA, by player*

**2.106 Distribution of Blocked Shots in the NBA** The variable *Blocks* in the dataset **NBAPlayers2011** includes information on the number of blocked shots during the season for each of the 176 players in the dataset.

(a) Use technology to find the mean and the standard deviation of the number of blocked shots.

(b) Use technology to find the five number summary for the same variable.

(c) Which set of summary statistics, those from part (a) or part (b), is more resistant to outliers and more appropriate if the data are heavily skewed?

(d) Use technology to create a graph of the data in *Blocks* and describe the shape of the distribution.

(e) Is it appropriate to use the 95% rule with these data? Why or why not?

**2.107 Which Accomplishment of LeBron James is Most Impressive?** Table 2.24 shows the means and standard deviations for four of the variables in the **NBAPlayers2011** dataset. *FGPct* is the field goal percentage, *Points* is total number of points scored during the season, *Assists* is total number of assists during the season, and *Steals* is total number of steals during the season. LeBron James had a field goal percentage of 0.510, scored 2111 points, had 554 assists, and had 124 steals. Find the $z$-score for each of LeBron's statistics. Use the $z$-scores to determine, relative to the other players in the NBA that season, which statistic of LeBron's is the most impressive. Which is the least impressive?

**Table 2.24** *Summary statistics on NBA players*

| Variable | Mean | Standard Deviation |
|---|---|---|
| *FGPct* | 0.464 | 0.053 |
| *Points* | 994 | 414 |
| *Assists* | 220 | 170 |
| *Steals* | 68.2 | 31.5 |

**2.108 SAT Scores** Stanley, a recent high school student, took the SAT exam in 2011 and got a 600 in all three components (Critical Reading, Math, and Writing). He was interested in how well he did compared to the rest of his peers. Table 2.25 shows the summary statistics for all students in 2011.[43]

[43] http://media.collegeboard.com/digitalServices/pdf/SAT-Percentile_Ranks_2011.pdf.

**Table 2.25** *Summary statistics for SAT scores*

| Component | Mean | Standard Deviation |
|---|---|---|
| Critical Reading | 497 | 114 |
| Math | 514 | 117 |
| Writing | 489 | 113 |

(a) Calculate $z$-scores for all three of Stanley's scores using the summary statistics in Table 2.25.

(b) Which of Stanley's three scores is the most unusual relative to his peers? Which is the least unusual?

(c) In which component did Stanley perform best relative to his peers?

**2.109 Comparing Global Internet Connections** The Nielsen Company measured connection speeds on home computers in nine different countries and wanted to determine whether connection speed affects the amount of time consumers spend online.[44] Table 2.26 shows the percent of Internet users with a "fast" connection (defined as 2Mb or faster) and the average amount of time spent online, defined as total hours connected to the web from a home computer during the month of February 2011. The data are also available in the dataset **GlobalInternet**.

(a) Use technology to find the mean and standard deviation of the nine values for percent with a fast connection.

(b) Use technology to find the mean and standard deviation of the nine values for time online.

(c) Does there seem to be a relationship between the two variables? Explain. (We examine this relationship further in Section 2.5.)

**Table 2.26** *Internet connection speed and hours online*

| Country | Percent Fast Connection | Hours Online |
|---|---|---|
| Switzerland | 88 | 20.18 |
| United States | 70 | 26.26 |
| Germany | 72 | 28.04 |
| Australia | 64 | 23.02 |
| United Kingdom | 75 | 28.48 |
| France | 70 | 27.49 |
| Spain | 69 | 26.97 |
| Italy | 64 | 23.59 |
| Brazil | 21 | 31.58 |

[44] "Swiss Lead in Speed: Comparing Global Internet Connections," NielsenWire, April 1, 2011.

**2.110 Jogging Times** Consider the jogging times from a set of 5-mile runs by two different runners in Table 2.27.

**Table 2.27** *Jogging times*

| Jogger 1 | Jogger 2 |
|---|---|
| 44 | 48 |
| 45 | 49 |
| 43 | 38 |
| 48 | 40 |
| 45 | 50 |

(a) Which runner is faster on average?

(b) What is the main difference in the jogging times of joggers 1 and 2?

**2.111 Mammal Longevities** Table 2.14 on page 61 shows longevity (typical lifespan) in years for 40 species of mammals, and the data are also available in **MammalLongevity**.

(a) Use technology to find the mean and standard deviation of the 40 values.

(b) The elephant's longevity of 40 years appears to be an outlier in the dotplot in Figure 2.6 on page 61. Find and interpret the $z$-score for the elephant.

**2.112 Daily Calorie Consumption** The five number summary for daily calorie consumption for the $n = 315$ participants in the **NutritionStudy** is (445, 1334, 1667, 2106, 6662).

(a) Give the range and the IQR.

(b) Which of the following numbers is most likely to be the mean of this dataset? Explain.

1550 1667 1796 3605

(c) Which of the following numbers is most likely to be the standard deviation of this dataset? Explain.

5.72 158 680 1897 5315

**2.113 Largest and Smallest Standard Deviation** Using only the whole numbers 1 through 9 as possible data values, create a dataset with $n = 6$ and $\bar{x} = 5$ and with:

(a) Standard deviation as small as possible

(b) Standard deviation as large as possible

### USING THE 95% RULE TO DRAW SMOOTH BELL-SHAPED CURVES

In Exercises 2.114 to 2.117, sketch a curve showing a distribution that is symmetric and bell-shaped

and has approximately the given mean and standard deviation. In each case, draw the curve on a horizontal axis with scale 0 to 10.

**2.114** Mean 3 and standard deviation 1

**2.115** Mean 7 and standard deviation 1

**2.116** Mean 5 and standard deviation 2

**2.117** Mean 5 and standard deviation 0.5

**2.118 Using the Five Number Summary to Visualize Shape of a Distribution** Draw a histogram or a smooth curve illustrating the shape of a distribution with the properties that:

(a) The range is 100 and the interquartile range is 10

(b) The range is 50 and the interquartile range is 40

**2.119 Rough Rule of Thumb for the Standard Deviation** According to the 95% rule, the largest value in a sample from a distribution which is approximately symmetric and bell-shaped should be between 2 and 3 standard deviations above the mean, while the smallest value should be between 2 and 3 standard deviations below the mean. Thus the range should be roughly 4 to 6 times the standard deviation. As a rough rule of thumb, we can get a quick estimate of the standard deviation for a bell-shaped distribution by dividing the range by 5. Check how well this quick estimate works in the following situations.

(a) Pulse rates from the **StudentSurvey** dataset discussed in Example 2.17 on page 77. The five number summary of pulse rates is (35, 62, 70, 78, 130) and the standard deviation is $s = 12.2$ bpm. Find the rough estimate using all the data, and then excluding the two outliers at 120 and 130, which leaves the maximum at 96.

(b) Number of hours a week spent exercising from the **StudentSurvey** dataset discussed in Example 2.21 on page 81. The five number summary of this dataset is $(0, 5, 8, 12, 40)$ and the standard deviation is $s = 5.741$ hours.

(c) Longevity of mammals from the **MammalLongevity** dataset discussed in Example 2.22 on page 82. The five number summary of the longevity values is $(1, 8, 12, 16, 40)$ and the standard deviation is $s = 7.24$ years.

# 2.4 OUTLIERS, BOXPLOTS, AND QUANTITATIVE/CATEGORICAL RELATIONSHIPS

In this section, we examine a relationship between a quantitative variable and a categorical variable by examining both comparative summary statistics and graphical displays. Before we get there, however, we have a bit more to do in our analysis of a single quantitative variable. We make the definition of an outlier more precise and look at one more graphical display for a single quantitative variable.

## Detection of Outliers

Consider again the data on mammal longevity in Data 2.2 on page 61. Our intuition suggests that the longevity of 40 years for the elephant is an unusually high value compared to the other lifespans in this sample. How do we determine objectively when such a value is an outlier? The criteria should depend on some measure of location for "typical" values and a measure of spread to help us judge when a data point is "far" from those typical cases. One approach uses the quartiles and interquartile range. As a rule, most data values will fall within about 1.5($IQR$)'s of the quartiles.[45]

### Detection of Outliers

As a general rule of thumb, we call a data value an **outlier** if it is

Smaller than $Q_1 - 1.5(IQR)$ or Larger than $Q_3 + 1.5(IQR)$

[45] In practice, determining outliers requires judgment and understanding of the context. We present a specific method here, but no single method is universally used for determining outliers.

**Descriptive Statistics: TV**

| Variable | Gender | N | Mean | SE Mean | StDev | Minimum | Q1 | Median | Q3 | Maximum |
|---|---|---|---|---|---|---|---|---|---|---|
| TV | F | 169 | 5.237 | 0.315 | 4.100 | 0.000 | 2.500 | 4.000 | 6.000 | 20.000 |
| | M | 192 | 7.620 | 0.464 | 6.427 | 0.000 | 3.000 | 5.000 | 10.000 | 40.000 |

**Figure 2.36** ***Output from Minitab comparing TV watching by gender***

these males had a mean of 7.620 hours spent watching TV per week with a standard deviation of 6.427. Both the mean and the standard deviation are larger for the males, which matches what we see in the graphs in Figure 2.34.

Using the notation $\bar{x}_m$ for the male mean and $\bar{x}_f$ for the female mean, the difference in means is

$$\bar{x}_m - \bar{x}_f = 7.620 - 5.237 = 2.383$$

In this sample, on average the males watched an additional 2.383 hours of television per week.

**SECTION LEARNING GOALS**

*You should now have the understanding and skills to:*

- Identify outliers in a dataset based on the $IQR$ method
- Use a boxplot to describe data for a single quantitative variable
- Use a side-by-side graph to visualize a relationship between quantitative and categorical variables
- Examine a relationship between quantitative and categorical variables using comparative summary statistics

## Exercises for Section 2.4

**SKILL BUILDER 1**

In Exercises 2.120 and 2.121, match the five number summaries with the boxplots.

**2.120** Match each five number summary with one of the boxplots in Figure 2.37.

(a) (2, 12, 14, 17, 25)

(b) (5, 15, 18, 20, 23)

(c) (10, 12, 13, 18, 25)

(d) (12, 12, 15, 20, 24)

**2.121** Match each five number summary with one of the boxplots in Figure 2.38.

(a) (1, 18, 20, 22, 25)

(b) (1, 10, 15, 20, 25)

(c) (1, 3, 5, 10, 25)

(d) (1, 1, 10, 15, 25)

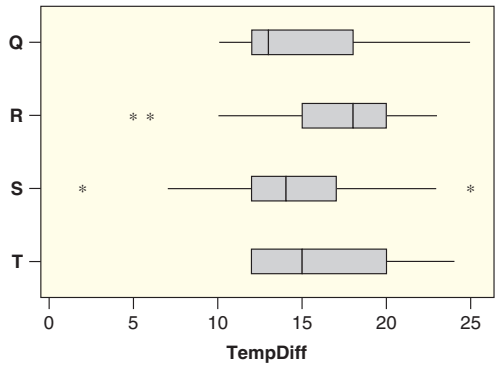



**Figure 2.37** ***Match five number summaries in Exercise 2.120***

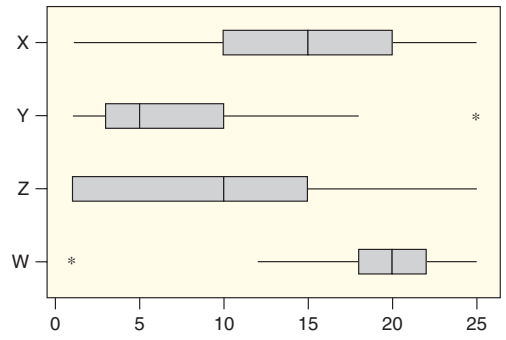



**Figure 2.38** *Match five number summaries in Exercise 2.121*

## SKILL BUILDER 2

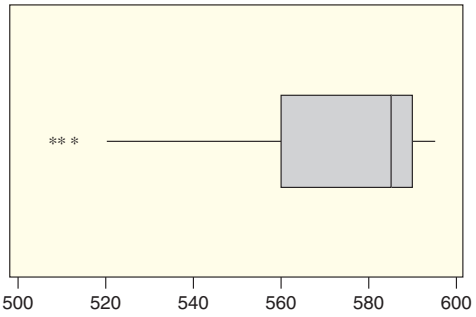
Exercises 2.122 to 2.125 show a boxplot for a set of data. In each case:

(a) Indicate whether the distribution of the data appears to be skewed to the left, skewed to the right, approximately symmetric, or none of these.

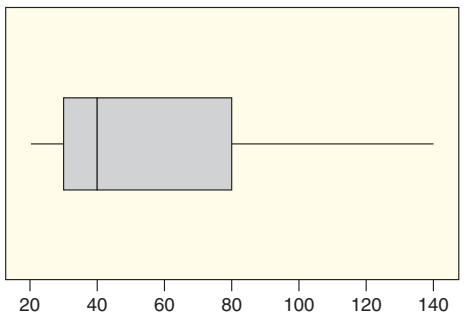
(b) Are there any outliers? If so, how many and are they high outliers or low outliers?

(c) Give a rough approximation for the mean of the dataset.

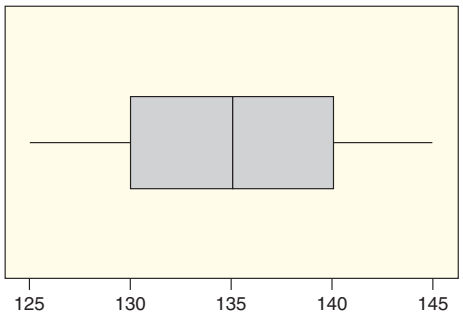
**2.122**



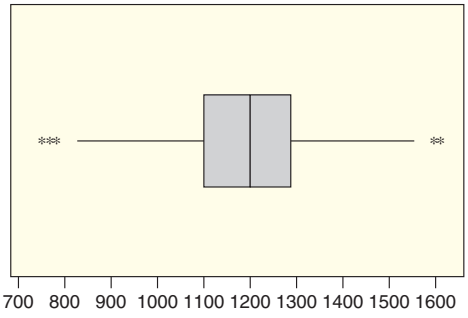

**2.123**




**2.124**




**2.125**




## SKILL BUILDER 3

Exercises 2.126 to 2.129 each describe a sample. The information given includes the five number summary, the sample size, and the largest and smallest data values in the tails of the distribution. In each case:

(a) Clearly identify any outliers.

(b) Draw a boxplot.

**2.126** Five number summary: (210, 260, 270, 300, 320); $n = 500$
Tails: 210, 215, 217, 221, 225, ..., 318, 319, 319, 319, 320, 320

**2.127** Five number summary: (15, 42, 52, 56, 71); $n = 120$
Tails: 15, 20, 28, 30, 31, ..., 64, 65, 65, 66, 71

**2.128** Five number summary: (42, 72, 78, 80, 99); $n = 120$
Tails: 42, 63, 65, 67, 68, ..., 88, 89, 95, 96, 99

**2.129** Five number summary: (5, 10, 12, 16, 30); $n = 40$
Tails: 5, 5, 6, 6, 6, ..., 22, 22, 23, 28, 30

## INVESTIGATING HOLLYWOOD MOVIES

In Exercises 2.131 to 2.133, we use data from **HollywoodMovies2011** introduced in Data 2.7 on page 93. The dataset includes information on all 136 movies to come out of Hollywood in 2011.

**2.130 How Profitable Are Hollywood Movies?** One of the variables in the **HollywoodMovies2011** dataset is *Profitability*, which measures the proportion of the budget recovered in revenue from the movie. A profitability less than 1 means the movie did not make enough money to cover the budget, while a profitability greater than 1 means it made a profit. A boxplot of the profitability ratings of all 136 movies is shown in Figure 2.39. (The largest outlier is the movie *Insidious*, which had a relatively small budget and relatively high gross revenue.)

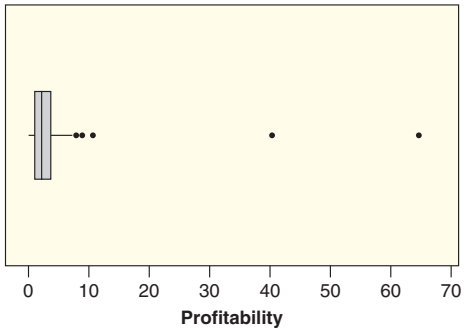



**Figure 2.39** ***Profitability of Hollywood movies***

(a) Describe the shape of the distribution.

(b) Estimate the range.

(c) Estimate the median. Interpret it in terms of profitability of movies.

(d) Do we expect the mean to be greater than or less than the median?

**2.131 Audience Scores on Rotten Tomatoes** Audience scores (on a scale from 1 to 100) on the Rotten Tomatoes website for all movies that came out of Hollywood in 2011 have a five number summary of (24, 49, 61, 77, 93). (These data are in the variable *AudienceScore* in the dataset **HollywoodMovies2011**.) Are there any outliers in these scores? How bad would an average audience score rating have to be on Rotten Tomatoes to qualify as a low outlier?

**2.132 Do Movie Budgets Differ Based on the Genre of the Movie?** The dataset **HollywoodMovies2011** includes a quantitative variable on the *Budget* of the movie, in millions of dollars, as well as a categorical variable classifying each movie by its *Genre*. Figure 2.40 shows side-by-side boxplots investigating a relationship between these two variables. (We use four of the possible categories in *Genre* for this exercise.)

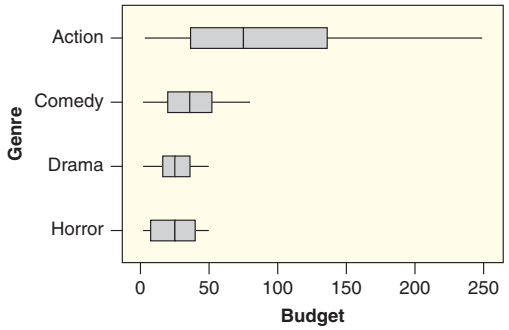



**Figure 2.40** ***Movie budgets (in millions of dollar) based on genre***

(a) Which genre appears to have the largest budgets? Which appears to have the smallest?

(b) Which genre has the biggest spread in its budgets? Which has the smallest spread?

(c) Does there appear to be an association between genre of a movie and size of the budget? Explain.

**2.133 Do Audience Ratings Differ Based on the Genre of the Movie?** The dataset **HollywoodMovies2011** includes a quantitative variable on the *AudienceScore* of the movie as well as a categorical variable classifying each movie by its *Genre*. The computer output below gives the audience rating based on genre. (We have only included the genres with at least 10 movies in that category.)

(a) Which genre has the highest mean audience score? The lowest mean audience score?

| Variable | Genre | N | Mean | StDev | Minimum | Q1 | Median | Q3 | Maximum |
|---|---|---|---|---|---|---|---|---|---|
| Audience Score | Action | 32 | 58.63 | 18.39 | 32.00 | 44.50 | 51.00 | 78.00 | 93.00 |
| | Animation | 12 | 64.08 | 13.89 | 43.00 | 50.50 | 64.00 | 78.25 | 82.00 |
| | Comedy | 27 | 59.11 | 15.68 | 31.00 | 48.00 | 58.00 | 71.00 | 93.00 |
| | Drama | 21 | 72.10 | 14.55 | 46.00 | 59.00 | 72.00 | 84.50 | 91.00 |
| | Horror | 17 | 48.65 | 15.88 | 25.00 | 34.00 | 52.00 | 60.50 | 78.00 |
| | Romance | 10 | 64.80 | 12.90 | 50.00 | 52.25 | 65.50 | 78.00 | 84.00 |
| | Thriller | 13 | 64.31 | 14.87 | 24.00 | 57.00 | 67.00 | 74.50 | 81.00 |

(b) Which genre has the highest median score? The lowest median score?

(c) In which genre is the lowest score, and what is that score? In which genre is the highest score, and what is that score?

(d) Which genre has the largest number of movies in that category?

**2.134 Physical Activity by Region of the Country in the US** The variables in **USStates** include the percent of the people in each state who say they have engaged in any physical activity in the last month as well as the region of the country in which the state is found (Midwest, Northeast, South, or West). One of these variables is quantitative and one is categorical, and Figure 2.41 allows us to visualize the relationship between the two variables.

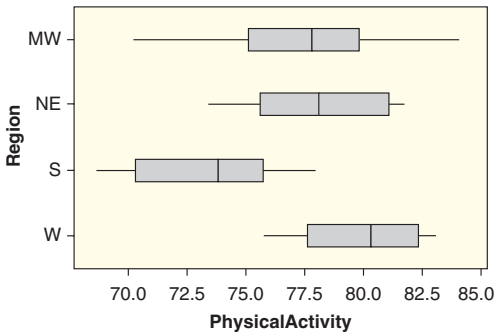



**Figure 2.41** *Physical activity in US states by region of the country*

(a) Which region shows the lowest level of physical activity? Which region shows the highest?

(b) Which region appears to have the biggest range?

(c) Are there any outliers?

(d) Does there appear to be an association between amount of physical activity and region of the country?

**2.135 Infection in Dialysis Patients** Table 2.28 gives data showing the time to infection, at the point of insertion of the catheter, for kidney patients using portable dialysis equipment. There are 38 patients, and the data give the first observation for each patient.[49] The five number summary for the data is (2, 15, 46, 149, 536).

[49]McGilchrist, C. and Aisbett, C., "Regression with frailty in survival analysis," *Biometrics*, 1991; 47: 461–466.

**Table 2.28** *Time to infection for dialysis patients*

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 2 | 5 | 6 | 7 | 7 | 8 | 12 | 13 |
| 15 | 15 | 17 | 22 | 22 | 23 | 24 | 27 |
| 30 | 34 | 39 | 53 | 54 | 63 | 96 | 113 |
| 119 | 130 | 132 | 141 | 149 | 152 | 152 | 185 |
| 190 | 292 | 402 | 447 | 511 | 536 | | |

(a) Identify any outliers in the data. Justify your answer.

(b) Draw the boxplot.

**2.136 Hits in Baseball** Major League Baseball is split into two leagues, the National League (NL) and the American League (AL). The main difference between the two leagues is that pitchers take at bats in the National League but not in the American League. Are total team hits different between the two leagues? Figure 2.42 shows side-by-side boxplots for the two leagues. The data are stored in **BaseballHits**.

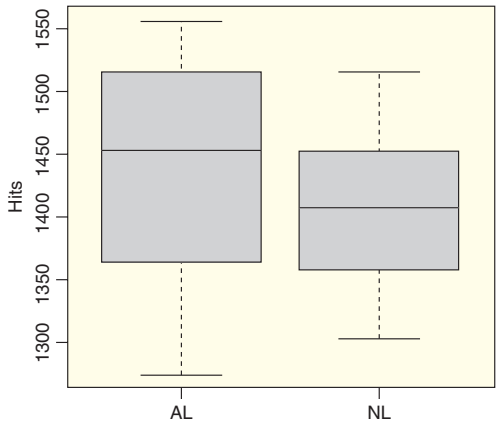



**Figure 2.42** *Side by side boxplots for hits by league*

(a) Estimate the median number of hits for each league, and estimate the difference in median hits between the two leagues. Which league appears to get more hits?

(b) What is the other obvious difference (apparent in Figure 2.42) between the two leagues?

## EFFECT OF DIET ON NUTRIENTS IN THE BLOOD

Exercises 2.137 to 2.139 use data from **NutritionStudy** on dietary variables and concentrations of micronutrients in the blood for a sample of $n = 315$ individuals.

**2.137 Daily Calorie Consumption** The five number summary for daily calorie consumption is (445, 1334, 1667, 2106, 6662).

(a) The ten largest data values are given below. Which (if any) of these is an outlier?

3185 3228 3258 3328 3450 3457
3511 3711 4374 6662

(b) Determine whether there are any low outliers. Show your work.

(c) Draw the boxplot for the calorie data.

**2.138 Daily Calories by Gender** Figure 2.43 shows side-by-side boxplots comparing calorie consumption by gender.

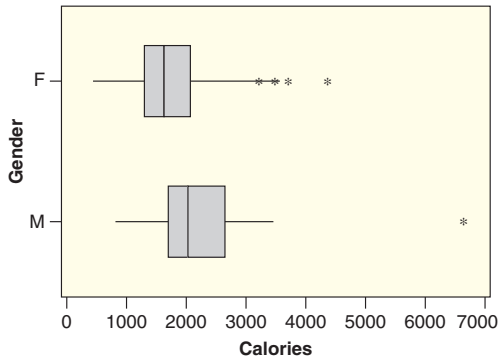



**Figure 2.43** *Calorie consumption by gender*

(a) Which gender has the largest median daily calorie consumption? Which gender has the largest outlier? Which gender has the most outliers?

(b) Does there seem to be an association between gender and calorie consumption? Explain.

**2.139 Concentration of Retinol by Vitamin Use** Figure 2.44 displays the relationship between vitamin use and the concentration of retinol (a micronutrient) in the blood. Does there seem to be an association between these two variables?

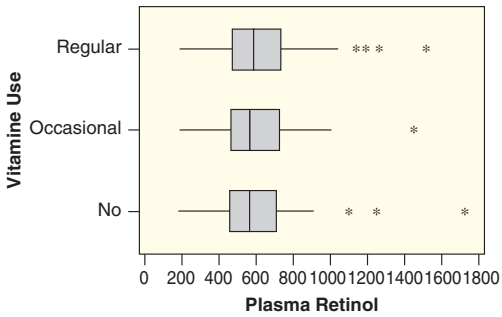



**Figure 2.44** *Concentration of retinol by vitamin use*

**2.140 Systolic Blood Pressure** Figure 2.45 shows the boxplot for the systolic blood pressures for all 200 patients in the ICU study in **ICUAdmissions**. Discuss what information this graph gives about the distribution of blood pressures in this sample of patients. What is the five-number summary?

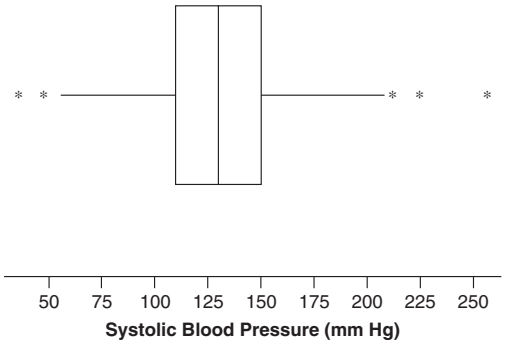



**Figure 2.45** *Systolic blood pressure of ICU patients*

**2.141 Systolic Blood Pressure and Survival** The data in **ICUAdmissions** contains a categorical variable *Status* indicating whether each patient lived (0) or died (1). Is there a relationship between the status (lived/died) and the systolic blood pressures? Use the side-by-side boxplots showing the systolic blood pressures for these two groups of patients in Figure 2.46 to discuss how the distributions compare.

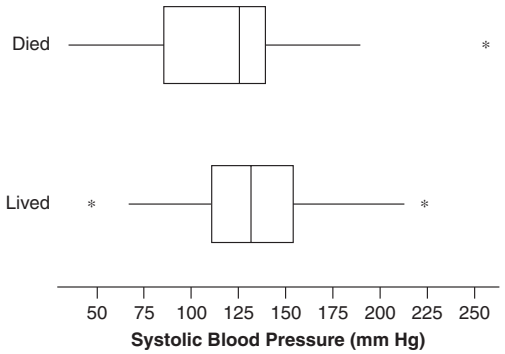



**Figure 2.46** *Systolic blood pressures of patients who lived or died*

**2.142 How Do Honeybees Communicate Quality?** When honeybees are looking for a new home, they send out scouts to explore options. When a scout

returns, she does a "waggle dance" with multiple circuit repetitions to tell the swarm about the option she found.[50] The bees then decide between the options and pick the best one. Scientists wanted to find out how honeybees decide which is the best option, so they took a swarm of honeybees to an island with only two possible options for new homes: one of very high honeybee quality and one of low quality. They then kept track of the scouts who visited each option and counted the number of waggle dance circuits each scout bee did when describing the option.[51] Comparative dotplots of number of circuits performed by the 41 bees that visited the high-quality option and the 37 bees that visited the low-quality option are shown in Figure 2.47.

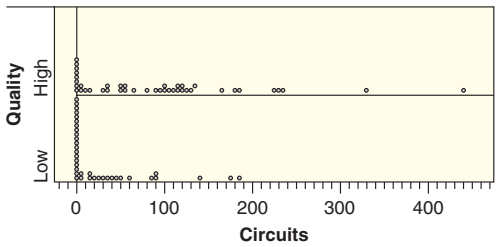



**Figure 2.47** ***Number of circuits completed in the honeybee waggle dance***

(a) Does there appear to be an association between number of circuits in the waggle dance and the quality of the site? If so, describe the association.

(b) The five number summary for the number of circuits for those describing the high-quality site is $(0, 7.5, 80, 122.5, 440)$, while the five number summary for those describing the low-quality site is $(0, 0, 0, 42.5, 185)$. Use the *IQR* method to identify any outliers in either group. Justify your answer.

(c) The mean for the high-quality group is $\bar{x}_H = 90.5$ with a standard deviation of 94.6, while the mean for the low-quality group is $\bar{x}_L = 30.0$ with a standard deviation of 49.4. What is the difference in means, $\bar{x}_H - \bar{x}_L$?

(d) Find the $z$-score for the largest value in the high-quality group and the $z$-score for the largest value in the low-quality group. Which is larger relative to its group?

(e) Is it appropriate to use the 95% rule with either set of data?

[50]Check out a honeybee waggle dance on youtube!
[51]Seeley, T., *Honeybee Democracy*, Princeton University Press, Princeton, NJ, 2010, p. 128.

**2.143 Effect of Calcium on Fish** In a study[52] to determine how the calcium level of water affects respiration rate of fish, 360 fish in a sample were randomly divided into three tanks with different levels of calcium: low, medium, and high. The respiration rate of the fish, in beats per minute, was then measured. The dataset is in **FishGills3** and the two variables are *Calcium* and *GillRate*.

(a) Use technology to create side-by-side boxplots for gill rate in the three different calcium conditions. Describe the relationship between the two variables.

(b) Use technology to obtain comparative summary statistics for gill rate in the three different calcium conditions and give the mean and the standard deviation for the gill rates in each of the three calcium conditions.

(c) Is this study an experiment or an observational study?

**2.144 Better Traffic Flow** Have you ever driven along a street where it seems that every traffic light is red when you get there? Some engineers in Dresden, Germany, are looking at ways to improve traffic flow by enabling traffic lights to communicate information about traffic flow with nearby traffic lights. The data in **TrafficFlow** show results of one experiment[53] that simulated buses moving along a street and recorded the delay time (in seconds) for both a fixed time and a flexible system of lights. The simulation was repeated under both conditions for a total of 24 trials.

(a) What is the explanatory variable? What is the response variable? Is each categorical or quantitative?

(b) Use technology to find the mean and the standard deviation for the delay times under each of the two conditions (*Timed* and *Flexible*). Does the flexible system seem to reduce delay time?

(c) The data in **TrafficFlow** are paired since we have two values, timed and flexible, for each simulation run. For paired data we generally compute the *difference* for each pair. In this example, the dataset includes a variable called *Difference* that stores the difference

[52]Thanks to Professor Brad Baldwin of St. Lawrence University for this dataset.
[53]Lammer, S. and Helbing, D., "Self-Stabilizing Decentralized Signal Control of Realistic, Saturated Network Traffic," Santa Fe Institute, Santa Fe, NM, working paper No. 10-09-019, September 2010.

*Timed – Flexible* for each simulation run. Use technology to find the mean and standard deviation of these differences.

(d) Use technology to draw a boxplot of the differences. Are there any outliers?

**DRAW THESE SIDE-BY-SIDE BOXPLOTS**

Exercises 2.145 to 2.146 examine issues of location and spread for boxplots. In each case, draw side-by-side boxplots of the datasets on the same scale. There are many possible answers.

**2.145** One dataset has median 25, interquartile range 20, and range 30. The other dataset has median 75, interquartile range 20, and range 30.

**2.146** One dataset has median 50, interquartile range 20, and range 40. A second dataset has median 50, interquartile range 50, and range 100. A third dataset has median 50, interquartile range 50, and range 60.

**2.147 Examine a Relationship in StudentSurvey** From the **StudentSurvey** dataset, select any categorical variable and select any quantitative variable. Use technology to create side-by-side boxplots to examine the relationship between the variables. State which two variables you are using and describe what you see in the boxplots. In addition, use technology to compute comparative summary statistics and compare means and standard deviations for the different groups.

**2.148 Examine a Relationship in USStates** Exercise 2.134 examined the relationship between region of the country and level of physical activity of the population of US states. From the **USStates** dataset, examine a different relationship between a categorical variable and a quantitative variable. Select one of each type of variable and use technology to create side-by-side boxplots to examine the relationship between the variables. State which two variables you are using and describe what you see in the boxplots. In addition, use technology to compute comparative summary statistics and compare means and standard deviations for the different groups.

## 2.5 TWO QUANTITATIVE VARIABLES: SCATTERPLOT AND CORRELATION

In Section 2.1 we looked at relationships between two categorical variables, and in Section 2.4 we investigated relationships between a categorical and a quantitative variable. In this section, we look at relationships between two quantitative variables.

**DATA 2.9** **Presidential Approval Ratings and Re-Election**

When a US president runs for re-election, how strong is the relationship between the president's approval rating and the outcome of the election? Table 2.29 includes all the presidential elections since 1940 in which an incumbent was running and shows the presidential approval rating at the time of the election and the margin of victory or defeat for the president in the election.[54] The data are available in **ElectionMargin**. ■

**Example 2.32**

(a) What was the highest approval rating for any of the losing presidents? What was the lowest approval rating for any of the winning presidents? Make a conjecture about the approval rating needed by a sitting president in order to win re-election.

(b) Approval rating and margin of victory are both quantitative variables. Does there seem to be an association between the two variables?

[54]Silver, N., "Approval Ratings and Re-Election Odds," fivethirtyeight.com, posted January 28, 2011. There are no results for 1944 because Gallup went on wartime hiatus.

## A Formula for Correlation

We routinely rely on technology to compute correlations, but you may be wondering how such computations are done. While computing a correlation "by hand" is tedious and often not very informative, a formula, such as the one shown below, can be helpful in understanding how the correlation works:

$$r = \frac{1}{n-1}\sum\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$$

Essentially this involves converting all values for both variables to $z$-scores, which puts the correlation on a fixed $-1$ to $+1$ scale and makes it independent of the scale of measurement. For a positive association, large values for $x$ tend to occur with large values of $y$ (both $z$-scores are positive) and small values (with negative $z$-scores) tend to occur together. In either case the products are positive which leads to a positive sum. For a negative association, the $z$-scores tend to have opposite signs (small $x$ with large $y$ and vice versa) so the products tend to be negative.

### SECTION LEARNING GOALS

*You should now have the understanding and skills to:*

- Describe an association displayed in a scatterplot
- Explain what a positive or negative association means between two variables
- Interpret a correlation
- Use technology to calculate a correlation
- Recognize that correlation does not imply cause and effect
- Recognize that you should always plot your data in addition to interpreting numerical summaries

## Exercises for Section 2.5

**SKILL BUILDER 1**

Match the scatterplots in Figure 2.54 with the correlation values in Exercises 2.149 to 2.152.

**2.149** $r = -1$

**2.150** $r = 0$

**2.151** $r = 0.8$

**2.152** $r = 1$

**SKILL BUILDER 2**

Match the scatterplots in Figure 2.55 with the correlation values in Exercises 2.153 to 2.156.

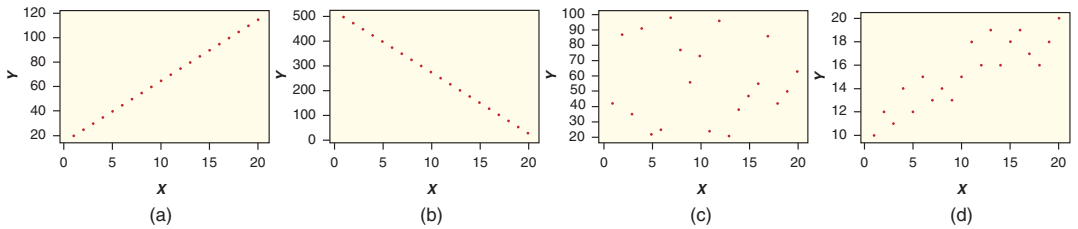



**Figure 2.54** *Match the correlations to the scatterplots*

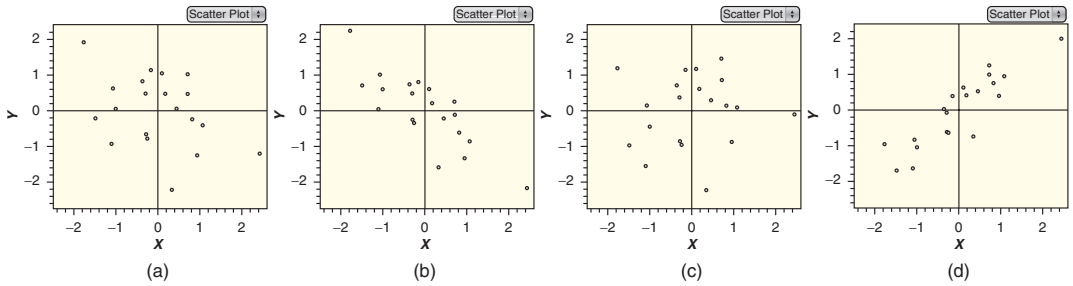



**Figure 2.55** *Match the correlations to the scatterplots*

**2.153** $r = 0.09$

**2.154** $r = -0.38$

**2.155** $r = 0.89$

**2.156** $r = -0.81$

### SKILL BUILDER 3

In Exercises 2.157 to 2.162, two quantitative variables are described. Do you expect a positive or negative association between the two variables? Explain your choice.

**2.157** Size of a house *and* Cost to heat the house

**2.158** Distance driven since the last fill-up of the gas tank *and* Amount of gas left in the tank

**2.159** Outside temperature *and* Amount of clothes worn

**2.160** Number of text messages sent on a cell phone *and* Number of text messages received on the phone

**2.161** Number of people in a square mile *and* Number of trees in the square mile

**2.162** Amount of time spent studying *and* Grade on the exam

### SKILL BUILDER 4

In Exercises 2.163 and 2.164, make a scatterplot of the data. Put the $X$ variable on the horizontal axis and the $Y$ variable on the vertical axis.

**2.163**

| $X$ | 3 | 5 | 2 | 7 | 6 |
|---|---|---|---|---|---|
| $Y$ | 1 | 2 | 1.5 | 3 | 2.5 |

**2.164**

| $X$ | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 |
|---|---|---|---|---|---|---|---|---|
| $Y$ | 532 | 466 | 478 | 320 | 303 | 349 | 275 | 221 |

### SKILL BUILDER 5

In Exercises 2.165 and 2.166, use statistical software on a computer or calculator to find the correlation for the data indicated.

**2.165** The data in Exercise 2.163

**2.166** The data in Exercise 2.164

**2.167 Presidential Approval Ratings and Re-e lection Odds** In Data 2.9 on page 103, we discuss the relationship between a president's approval rating when running for re-election and the margin of victory or defeat in the election. Table 2.29 shows the data and Figure 2.48 shows a scatterplot of the data.

(a) In how many of the 11 elections listed did the incumbent president lose? Since 1940, what percent of the time has the sitting president lost his bid for re-election?

(b) Which president had the highest approval rating? Which president had the highest margin of victory? Identify these two points on the scatterplot.

**2.168 Height and Weight** The quantitative variables *Height* (in inches) and *Weight* (in pounds) are included in the **StudentSurvey** dataset.

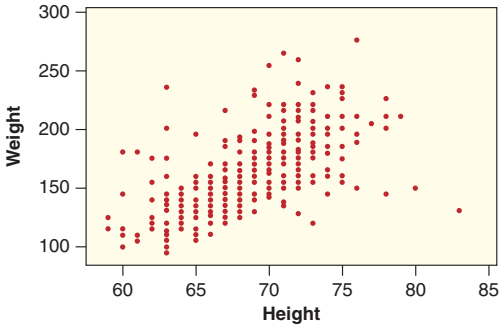



**Figure 2.56** *Scatterplot of height and weight*

(a) What would a positive association mean for these two variables? What would a negative association mean? Which do you expect is more likely?

(b) Figure 2.56 shows a scatterplot of the data. Does there appear to be a positive or negative relationship between height and weight?

How strong does the trend appear to be? Does it appear to be approximately a linear trend?

(c) Describe the person represented by the outlier in the lower right corner.

**2.169 Light Roast or Dark Roast for Your Coffee?** A somewhat surprising fact about coffee is that the longer it is roasted, the less caffeine it has. Thus an "extra bold" dark roast coffee actually has less caffeine than a light roast coffee. What is the explanatory variable and what is the response variable? Do the two variables have a negative association or a positive association?

**2.170 Mother's Love, Hippocampus, and Resiliency** Multiple studies[58] in both animals and humans show the importance of a mother's love (or the unconditional love of any close person to a child) in a child's brain development. A recent study shows that children with nurturing mothers had a substantially larger area of the brain called the hippocampus than children with less nurturing mothers. This is important because other studies have shown that the size of the hippocampus matters: People with large hippocampus area are more resilient and are more likely to be able to weather the stresses and strains of daily life. These observations come from experiments in animals and observational studies in humans.

(a) Is the amount of maternal nurturing one receives as a child positively or negatively associated with hippocampus size?

(b) Is hippocampus size positively or negatively associated with resiliency and the ability to weather the stresses of life?

(c) How might a randomized experiment be designed to test the effect described in part (a) in humans? Would such an experiment be ethical?

(d) Can we conclude that maternal nurturing in humans causes the hippocampus to grow larger? Can we conclude that maternal nurturing in animals (such as mice, who were used in many of the experiments) causes the hippocampus to grow larger? Explain.

**2.171 Commitment Genes and Cheating Genes** In earlier studies, scientists reported finding a "commitment gene" in men, in which men with a certain gene variant were much less likely to commit to a monogamous relationship.[59] That study involved only men (and we return to it later in this text), but a new study, involving birds this time rather than humans, shows that female infidelity may be inherited.[60] Scientists recorded who mated with or rebuffed whom for five generations of captive zebra finches, for a total of 800 males and 754 females. Zebra finches are believed to be a monogamous species, but the study found that mothers who cheat with multiple partners often had daughters who also cheat with multiple partners. To identify whether the effect was genetic or environmental, the scientists switched many of the chicks from their original nests. More cheating by the mother was strongly associated with more cheating by the daughter. Is this a positive or negative association?

**2.172 The Happy Planet Index** The website TED.com offers free short presentations, called TED Talks, on a variety of interesting subjects. One of the talks is called "The Happy Planet Index," by Nic Marks.[61] Marks comments that we regularly measure and report economic data on countries, such as Gross National Product, when we really ought to be measuring the well-being of the people in the countries. He calls this measure *Happiness*, with larger numbers indicating greater happiness, health, and well-being. In addition, he believes we ought to be measuring the ecological footprint, per capita, of the country, with larger numbers indicating greater use of resources (such as gas and electricity) and more damage to the planet. Figure 2.57 shows a scatterplot of these two quantitative variables. The data are given in **HappyPlanetIndex**.

(a) Does there appear to be a mostly positive or mostly negative association between these two variables? What does that mean for these two variables?

(b) Describe the happiness and ecological footprint of a country in the bottom left of the graph.

(c) Costa Rica has the highest *Happiness* index. Find it on the graph and estimate its ecological footprint score.

(d) For ecological footprints between 0 and 6, does a larger ecological footprint tend to be associated with more happiness? What about for ecological footprints between 6 and 10? Discuss this result in context.

[58]Raison, C., "Love key to brain development in children," cnn.com, *The Chart*, March 12, 2012.
[59]Timmer, J., "Men with genetic variant struggle with commitment," arstechnica.com, reporting on a study in the *Proceedings of the National Academy of Science*, 2009.
[60]Millus, S., "Female infidelity may be inherited," *Science News*, July 16, 2011, p. 10.
[61]Marks, N., "The Happy Planet Index," www.TED.com/talks, August 29, 2010.

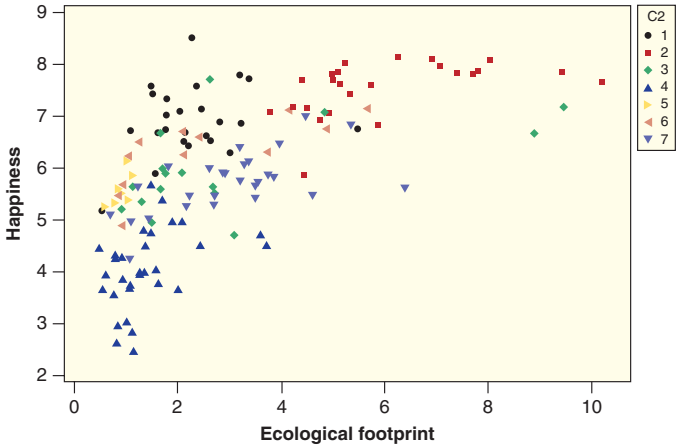



**Figure 2.57** *Happiness and ecological footprint*

(e) Marks believes we should be working to move all countries to the top left of the graph, closer to Costa Rica. What attributes does a country in the upper left of the graph possess?

(f) This graph shows a third variable as well: region of the world. One way to depict a categorical variable on a scatterplot is using different colors or shapes for different categories. The code is given in the top right, and is categorized as follows: 1 = Latin America, 2 = Western nations, 3 = Middle East, 4 = Sub-Saharan Africa, 5 = South Asia, 6 = East Asia, 7 = former Communist countries. Discuss one observation of an association between region and the two quantitative variables.

(g) If the goal is to move all countries to the top left, how should efforts be directed for those in the bottom left? How should efforts be directed for those in the top right?

**2.173 Vegetables and Obesity**

The **USStates** dataset includes information on the 50 US states, including the percent of the population of each state that eats at least five servings of fruits and vegetables a day and the percent of the population of each state that is obese. Figure 2.58 shows a scatterplot of these two variables.

(a) Does the scatterplot show a positive or negative association? Explain why your answer makes sense for these two variables.

(b) Where would a very healthy state be located on the scatterplot: top left, top right, bottom left, bottom right, or middle? What about a very unhealthy state?

(c) Pick a point in a very healthy location in the scatterplot, and use the dataset **USStates** to find the state it represents. Pick a point in a very unhealthy location and find the state it represents.

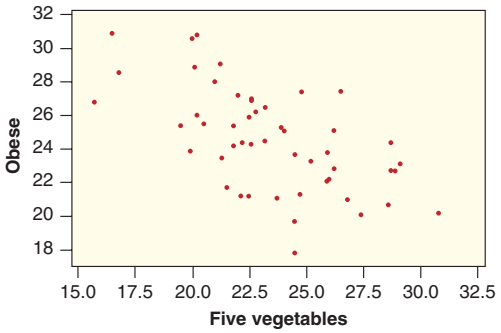



**Figure 2.58** *Percent eating five vegetables a day and percent obese*

(d) Is the data from a sample or a population? What is the correct notation for the correlation?

(e) Which of the following is most likely to be the correlation between these two variables?

$$-1, \quad -0.941, \quad -0.605, \quad -0.083, \quad 0.172, \quad 0.445, \quad 0.955, \quad 1$$

(f) Would a positive correlation imply that eating more vegetables will cause you to gain weight?

(g) Would a negative correlation imply that eating more vegetables will cause you to lose weight?

(h) One state stands out for eating an average number of vegetables but having a particularly low obesity rate. What state is this?

**2.174 Ages of Husbands and Wives** Suppose we record the husband's age and the wife's age for many randomly selected couples.

(a) What would it mean about ages of couples if these two variables had a negative relationship?

(b) What would it mean about ages of couples if these two variables had a positive relationship?

(c) Which do you think is more likely, a negative or a positive relationship?

(d) Do you expect a strong or a weak relationship in the data? Why?

(e) Would a strong correlation imply there is an association between husband age and wife age?

**2.175 Is Your Body Language Closed or Open?** A closed body posture includes sitting hunched over or standing with arms crossed rather than sitting or standing up straight and having the arms more open. According to a recent study, people who were rated as having a more closed body posture "had higher levels of stress hormones and said they felt less powerful than those who had a more open pose."[62]

(a) What are the variables in this study? Is each variable categorical or quantitative? Assume participants had body language rated on a numerical scale from low values representing more closed to larger values representing more open. Assume also that participants were rated on a numerical scale indicating whether each felt less powerful (low values) or more powerful (higher values).

(b) Do the results of the study indicate a positive or negative relationship between the body language scores and levels of stress hormones? Would your answer be different if the scale had been reversed for the body language scores?

(c) Do the results of the study indicate a positive or negative relationship between the body language scores and the scores on the feelings of power? Would your answer be different if both scales were reversed? Would your answer be different if only one of the scales had been reversed?

**2.176 SAT Scores: Math vs Verbal** The **StudentSurvey** dataset includes scores on the Math and Verbal portions of the SAT exam.

(a) What would a positive relationship between these two variables imply about SAT scores? What would a negative relationship imply?

[62] "Don't Slouch!" *Consumer Reports OnHealth*, February 2011; 23(2): p.3.

(b) Figure 2.59 shows a scatterplot of these two variables. For each corner of the scatterplot (top left, top right, bottom left, bottom right), describe a student whose SAT scores place him or her in that corner.

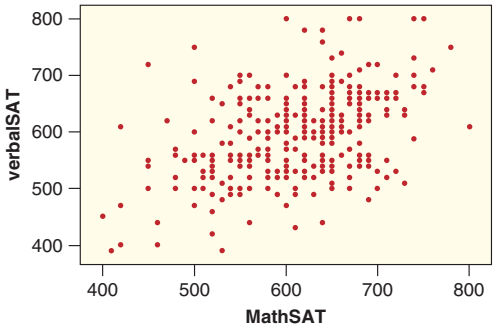



**Figure 2.59** *Math SAT score and Verbal SAT score*

(c) Does there appear to be a strong linear relationship between these two variables? What does that tell you about SAT scores?

(d) Which of the following is most likely to be the correlation between these two variables?

$-0.941, \quad -0.605, -0.235, \quad 0.445, \quad 0.751, \quad 0.955$

**2.177 Exercising or Watching TV?** The **StudentSurvey** dataset includes information on the number of hours a week students say they exercise and the number of hours a week students say they watch television.

(a) What would a positive relationship between these two variables imply about the way students spend their time? What would a negative relationship imply?

(b) For each corner of the scatterplot of these two variables shown in Figure 2.60 (top left, top

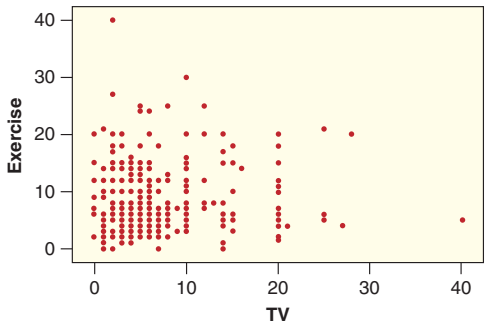



**Figure 2.60** *Number of hours a week of exercise and of television watching*

right, bottom left, bottom right), describe a student whose daily habits place him or her in that corner.

(c) There are two outliers in this scatterplot. Describe the student corresponding to the outlier on the right. Describe the student corresponding to the outlier on the top.

(d) The correlation between these two variables is $r = 0.01$. What does this correlation tell you about the strength of a linear relationship between these two variables?

**2.178 Blood Pressure and Heart Rate** In Example 2.19 on page 79 we computed z-scores for patient #772 in the **ICUAdmissions** dataset, who had a high systolic blood pressure reading of 204 but a low pulse rate of 52 bpm.

(a) Find the point corresponding to patient #772 on the scatterplot of blood pressure vs heart rate shown in Figure 2.61.

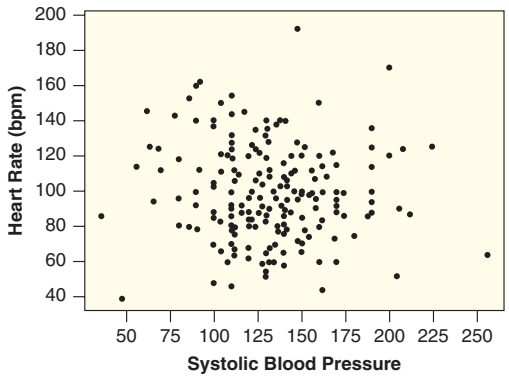



**Figure 2.61** *Blood pressure vs heart rate for ICU patients*

(b) Patient #772 has a high blood pressure reading but a low pulse rate. Does the scatterplot in Figure 2.61 support a conjecture that these two variables have a negative association?

**2.179 An Outlier in Jogging Times** Table 2.32 gives the times for five races in which two joggers participated.

(a) Use technology to construct a scatterplot of the race times.

(b) Use technology to find the correlation.

(c) A sixth race is held on a very windy day, and jogger A takes 50 minutes while jogger B takes a whole hour to complete the race. Recalculate the correlation with this point added.

**Table 2.32** *Jogging times*

| Jogger A | Jogger B |
|---|---|
| 44 | 48 |
| 45 | 49 |
| 43 | 38 |
| 48 | 40 |
| 45 | 50 |

(d) Compare correlations from parts (b) and (c). Did adding the results from the windy day have an effect on the relationship between the two joggers?

**2.180 Comparing Global Internet Connections** In Exercise 2.109 on page 89, we discuss a study in which the Nielsen Company measured connection speeds on home computers in nine different countries in order to determine whether connection speed affects the amount of time consumers spend online.[63] Table 2.33 shows the percent of Internet users with a "fast" connection (defined as 2Mb or faster) and the average amount of time spent online, defined as total hours connected to the web from a home computer during the month of February 2011. The data are also available in the dataset **GlobalInternet**.

**Table 2.33** *Internet connection speed and hours online*

| Country | Percent Fast Connection | Hours Online |
|---|---|---|
| Switzerland | 88 | 20.18 |
| United States | 70 | 26.26 |
| Germany | 72 | 28.04 |
| Australia | 64 | 23.02 |
| United Kingdom | 75 | 28.48 |
| France | 70 | 27.49 |
| Spain | 69 | 26.97 |
| Italy | 64 | 23.59 |
| Brazil | 21 | 31.58 |

(a) What would a positive association mean between these two variables? Explain why a positive relationship might make sense in this context.

(b) What would a negative association mean between these two variables? Explain why a negative relationship might make sense in this context.

[63] "Swiss Lead in Speed: Comparing Global Internet Connections," NielsenWire, April 1, 2011.

(c) Make a scatterplot of the data, using connection speed as the explanatory variable and time online as the response variable. Is there a positive or negative relationship? Are there any outliers? If so, indicate the country associated with each outlier and describe the characteristics that make it an outlier for the scatterplot.

(d) If we eliminate any outliers from the scatterplot, does it appear that the remaining countries have a positive or negative relationship between these two variables?

(e) Use technology to compute the correlation. Is the correlation affected by the outliers?

(f) Can we conclude that a faster connection speed causes people to spend more time online?

**2.181 What's Wrong with the Statement?** A researcher claims to have evidence of a strong positive correlation ($r = 0.88$) between a person's blood alcohol content (BAC) and the type of alcoholic drink consumed (beer, wine, or hard liquor). Explain, statistically, why this claim makes no sense.

**2.182 Iris Petals** Allometry is the area of biology that studies how different parts of a body grow in relation to other parts. Figure 2.62 shows a scatterplot[64] comparing the length and width of petals of irises.

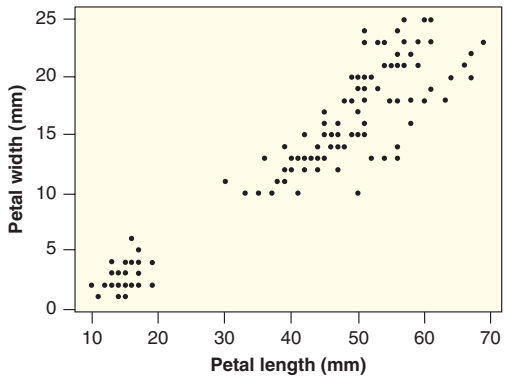



**Figure 2.62** *Iris petals*

(a) Does there appear to be a positive or negative association between petal width and petal length? Explain what this tells us about petals.

(b) Discuss the strength of a linear relationship between these two variables.

(c) Estimate the correlation.

(d) Are there any clear outliers in the data?

[64]R.A. Fishers's iris data downloaded from http://lib.stat.cmu.edu/DASL/Datafiles/Fisher'sIrises.html.

(e) Estimate the width of the petal which has a length of 30 mm.

(f) There are at least two different types of irises included in the study. Explain how the scatterplot helps illustrate this, and name one difference between the types that the scatterplot makes obvious.

**2.183 Create a Scatterplot** Draw any scatterplot satisfying the following conditions:

(a) $n = 10$ and $r = 1$

(b) $n = 8$ and $r = -1$

(c) $n = 5$ and $r = 0$

**2.184 Offensive Rebounds vs Defensive Rebounds** The dataset **NBAPlayers2011** is introduced on page 88, and includes many variables about players in the National Basketball Association in 2010–2011.

(a) Use technology to create a scatterplot for the relationship between the number of offensive rebounds in the season and the number of defensive rebounds. (Put offensive rebounds on the vertical axis.)

(b) Does the relationship appear to be positive or negative? What does that mean for these two variables? How strong is the relationship?

(c) There appear to be three outliers in the top right. Who are they?

(d) Use technology to find the correlation between these two variables.

**2.185 Do Movies with Larger Budgets Get Higher Audience Ratings?** The dataset **HollywoodMovies2011** is introduced on page 93, and includes many variables for movies that were produced in Hollywood in 2011, including *Budget* and *AudienceScore*.

(a) Use technology to create a scatterplot to show the relationship between the budget of a movie, in millions of dollars, and the audience score. We want to see if the budget has an effect on the audience score.

(b) Is there a linear relationship? How strong is it? Give your answer in the context of movies.

(c) There is an outlier with a very large budget. What is the audience rating for this movie and what movie is it? There is another data value with a budget of about 125 million dollars and an audience score over 90. To what movie does that dot correspond?

(d) Use technology to find the correlation between these two variables.

**2.186** **Pick a Relationship to Examine** Choose one of the following datasets: **USStates**, **HollywoodMovies2011**, **AllCountries**, or **NBAPlayers2011**, and then select any two quantitative variables that we have not yet analyzed. Use technology to graph a scatterplot of the two variables and discuss what you see. Is there a linear relationship? If so, is the association positive or negative? How strong is the trend? Are there any outliers? If so, identify them by name. In addition, use technology to find the correlation. Does the correlation match what you see in the scatterplot? Be sure to state the dataset and variables you use.

## 2.6 TWO QUANTITATIVE VARIABLES: LINEAR REGRESSION

In Section 2.5 we investigate the relationship between two quantitative variables. In this section, we discuss how to use one of the variables to predict the other when there is a linear trend.

**DATA 2.12** **Restaurant Tips**

The owner[65] of a bistro called *First Crush* in Potsdam, New York, is interested in studying the tipping patterns of its patrons. He collected restaurant bills over a two-week period that he believes provide a good sample of his customers. The data from 157 bills are stored in **RestaurantTips** and include the amount of the bill, size of the tip, percentage tip, number of customers in the group, whether or not a credit card was used, day of the week, and a coded identity of the server. ■

Image Source/Getty Images, Inc.

**Can we predict the size of a tip?**

For the restaurant tips data, we want to use the bill amount to predict the tip amount, so the explanatory variable is the amount of the bill and the response variable is the amount of the tip. A scatterplot of this relationship is shown in Figure 2.63.

[65]Thanks to Tom DeRosa for providing the tipping data.

Finally, when we plot the data, we also look for outliers that that may exert a strong influence on the regression line, similar to what we see for correlation in Figure 2.53 on page 111.

**Regression Caution #3**

Outliers can have a strong influence on the regression line, just as we saw for correlation. In particular, data points for which the explanatory value is an outlier are often called *influential points* because they exert an overly strong effect on the regression line.

**SECTION LEARNING GOALS**

*You should now have the understanding and skills to:*

- Use technology to find the regression line for a dataset with two quantitative variables
- Calculate predicted values from a regression line
- Interpret the slope (and intercept, when appropriate) of a regression line in context
- Calculate residuals and visualize residuals on a scatterplot
- Beware of extrapolating too far out when making predictions
- Recognize the importance of plotting your data

## Exercises for Section 2.6

**SKILL BUILDER 1**

In Exercises 2.187 to 2.190, two variables are defined, a regression equation is given, and one data point is given.

(a) Find the predicted value for the data point and compute the residual.

(b) Interpret the slope in context.

(c) Interpret the intercept in context, and if the intercept makes no sense in this context, explain why.

**2.187** *Hgt* = height in inches, *Age* = age in years of a child
$\widehat{Hgt} = 24.3 + 2.74(Age)$; data point is a child 12 years old who is 60 inches tall

**2.188** *BAC* = blood alcohol content (% of alcohol in the blood), *Drinks* = number of alcoholic drinks
$\widehat{BAC} = -0.0127 + 0.018(Drinks)$; data point is an individual who consumed 3 drinks and had a BAC of 0.08

**2.189** *Weight* = maximum weight capable of bench pressing (pounds), *Training* = number of hours spent lifting weights a week
$\widehat{Weight} = 95 + 11.7(Training)$; data point is an individual who trains 5 hours a week and can bench 150 pounds

**2.190** *Study* = number of hours spent studying for an exam, *Grade* = grade on the exam
$\widehat{Grade} = 41.0 + 3.8(Study)$; data point is a student who studied 10 hours and received an 81 on the exam

**SKILL BUILDER 2**

Use technology to find the regression line to predict $Y$ from $X$ in Exercises 2.191 to 2.194.

**2.191**

| | | | | | |
|---|---|---|---|---|---|
| $X$ | 3 | 5 | 2 | 7 | 6 |
| $Y$ | 1 | 2 | 1.5 | 3 | 2.5 |

**2.192**

| | | | | | | |
|---|---|---|---|---|---|---|
| $X$ | 2 | 4 | 6 | 8 | 10 | 12 |
| $Y$ | 50 | 58 | 55 | 61 | 69 | 68 |

**2.193**

| | | | | | | |
|---|---|---|---|---|---|---|
| $X$ | 10 | 20 | 30 | 40 | 50 | 60 |
| $Y$ | 112 | 85 | 92 | 71 | 64 | 70 |

**2.194**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $X$ | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 |
| $Y$ | 532 | 466 | 478 | 320 | 303 | 349 | 275 | 221 |

**2.195 Concentration of $CO_2$ in the Atmosphere** Levels of carbon dioxide ($CO_2$) in the atmosphere are rising rapidly, far above any levels ever before recorded. Levels were around 278 parts per million in 1800, before the Industrial Age, and had never, in the hundreds of thousands of years before that, gone above 300 ppm. Levels are now nearing 400 ppm. Table 2.35 shows the rapid rise of $CO_2$ concentrations over the last 50 years, also available in **CarbonDioxide**.[67] We can use this information to predict $CO_2$ levels in different years.

**Table 2.35** *Concentration of carbon dioxide in the atmosphere*

| Year | $CO_2$ |
|---|---|
| 1960 | 316.91 |
| 1965 | 320.04 |
| 1970 | 325.68 |
| 1975 | 331.08 |
| 1980 | 338.68 |
| 1985 | 345.87 |
| 1990 | 354.16 |
| 1995 | 360.62 |
| 2000 | 369.40 |
| 2005 | 379.76 |
| 2010 | 389.78 |

(a) What is the explanatory variable? What is the response variable?

(b) Draw a scatterplot of the data. Does there appear to be a linear relationship in the data?

(c) Use technology to find the correlation between year and $CO_2$ levels. Does the value of the correlation support your answer to part (b)?

(d) Use technology to calculate the regression line to predict $CO_2$ from year.

(e) Interpret the slope of the regression line, in terms of carbon dioxide concentrations.

(f) What is the intercept of the line? Does it make sense in context? Why or why not?

(g) Use the regression line to predict the $CO_2$ level in 2003. In 2020?

(h) Find the residual for 2010.

**2.196 The Honeybee Waggle Dance** When honeybee scouts find a food source or a nice site for a new home, they communicate the location to the rest of the swarm by doing a "waggle dance."[68] They point in the direction of the site and dance longer for sites farther away. The rest of the bees use the duration of the dance to predict distance to the site. Table 2.36 shows the distance, in meters, and the duration of the dance, in seconds, for seven honeybee scouts.[69] This information is also given in **HoneybeeWaggle**.

**Table 2.36** *Duration of a honeybee waggle dance to indicate distance to the source*

| Distance | Duration |
|---|---|
| 200 | 0.40 |
| 250 | 0.45 |
| 500 | 0.95 |
| 950 | 1.30 |
| 1950 | 2.00 |
| 3500 | 3.10 |
| 4300 | 4.10 |

(a) Which is the explanatory variable? Which is the response variable?

(b) Figure 2.69 shows a scatterplot of the data. Does there appear to be a linear trend in the data? If so, is it positive or negative?

(c) Use technology to find the correlation between the two variables.

(d) Use technology to find the regression line to predict distance from duration.

(e) Interpret the slope of the line in context.

(f) Predict the distance to the site if a honeybee does a waggle dance lasting 1 second. Lasting 3 seconds.

[67] Dr. Pieter Tans, NOAA/ESRL, www.esrl.noaa.gov/gmd/ccgg/trends/. Values recorded at the Mauna Loa Observatory in Hawaii.

[68] Check out a honeybee waggle dance on youtube!

[69] Seeley, T., *Honeybee Democracy*, Princeton University Press, Princeton, NJ, 2010, p. 128.

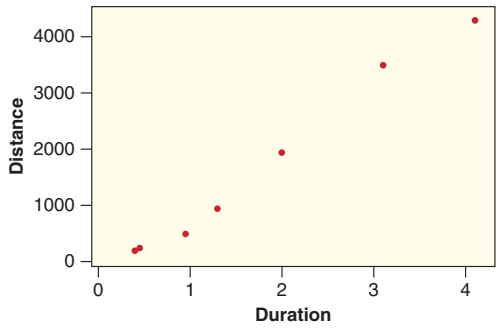



**Figure 2.69** *Using dance duration to predict distance to source*

**2.197 Is It Getting Harder to Win a Hot Dog Eating Contest?** Every Fourth of July, Nathan's Famous in New York City holds a hot dog eating contest, which we discuss in Exercise 2.98. Table 2.21 on page 86 shows the winning number of hot dogs and buns eaten every year from 2002 to 2011, and the data are also available in **HotDogs**. Figure 2.70 shows the scatterplot with the regression line.

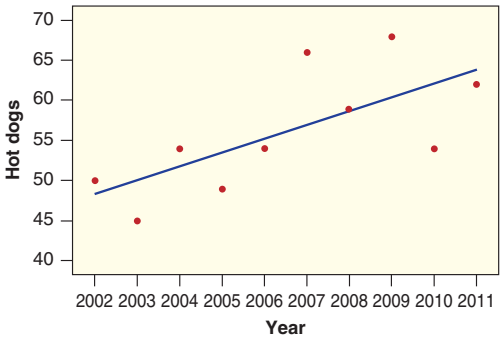



**Figure 2.70** *Winning number of hot dogs and buns*

(a) Is the trend in the data mostly positive or negative?

(b) Using Figure 2.70, is the residual larger in 2007 or 2008? Is the residual positive or negative in 2010?

(c) Use technology to find the correlation.

(d) Use technology to find the regression line to predict the winning number of hot dogs from the year.

(e) Interpret the slope of the regression line.

(f) Predict the winning number of hot dogs in 2012. (*Bonus*: Find the actual winning number in 2012 and compute the residual.)

(g) Why would it not be appropriate to use this regression line to predict the winning number of hot dogs in 2020?

**2.198 Runs and Wins in Baseball** In Exercise 2.136 on page 100, we looked at the relationship between total hits by team in the 2010 season and division (NL or AL) in baseball. Two other variables in the **BaseballHits** dataset are the number of wins and the number of runs scored during the season. The dataset consists of values for each variable from all 30 MLB teams. From these data we calculate the regression line:

$$\widehat{Wins} = 0.362 + 0.114(Runs)$$

(a) Which is the explanatory and which is the response variable in this regression line?

(b) Interpret the intercept and slope in context.

(c) The Oakland A's won 81 games while scoring 663 runs. Predict the number of games won by Oakland using the regression line. Calculate the residual. Were the A's efficient at winning games with 663 runs?

**2.199 Presidential Elections** In Example 2.43 on page 123, we used the approval rating of a president running for re-election to predict the margin of victory or defeat in the election. We saw that the least squares line is $\widehat{Margin} = -36.5 + 0.836(Approval)$. Interpret the slope and the intercept of the line in context.

**2.200 Height and Weight** Using the data in the **StudentSurvey** dataset, we use technology to find that a regression line to predict weight (in pounds) from height (in inches) is

$$\widehat{Weight} = -170 + 4.82(Height)$$

(a) What weight does the line predict for a person who is 5 feet tall (60 inches)? What weight is predicted for someone 6 feet tall (72 inches)?

(b) What is the slope of the line? Interpret it in context.

(c) What is the intercept of the line? If it is reasonable to do so, interpret it in context. If it is not reasonable, explain why not.

(d) What weight does the regression line predict for a baby who is 20 inches long? Why is it not appropriate to use the regression line in this case?

### PREDICTING PERCENT BODY FAT

Exercises 2.201 to 2.203 use the dataset **BodyFat**, which gives the percent of weight made up of body

fat for 100 men as well as other variables such as *Age*, *Weight* (in pounds), *Height* (in inches), and circumference (in cm) measurements for the *Neck*, *Chest*, *Abdomen*, *Ankle*, *Biceps*, and *Wrist*.[70]

**2.201 Using Weight to Predict Body Fat** Figure 2.71 shows the data and regression line for using weight to predict body fat percentage. For the case with the largest positive residual, estimate the values of both variables. In addition, estimate the predicted body fat percent and the residual for that point.

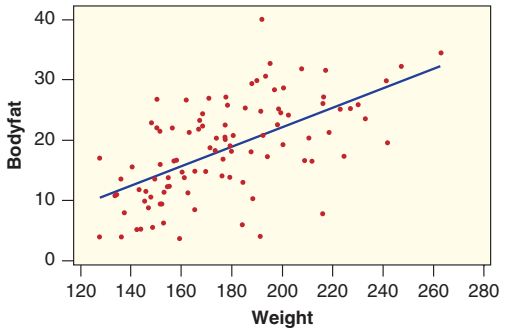



**Figure 2.71** *Using weight to predict percent body fat*

**2.202 Using Abdomen Circumference to Predict Body Fat** Figure 2.72 shows the data and regression line for using abdomen circumference to predict body fat percentage.

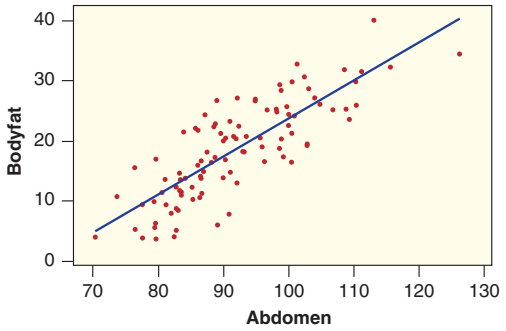



**Figure 2.72** *Using abdomen circumference to predict percent body fat*

(a) Which scatterplot, the one using *Weight* in Figure 2.71 or the one using *Abdomen* in Figure 2.72, appears to contain data with a larger correlation?

(b) In Figure 2.72, one person has a very large abdomen circumference of about 127 cm. Estimate the actual body fat percent for this person as well as the predicted body fat percent.

(c) Use Figure 2.72 to estimate the abdomen circumference for the person with about 40% body fat. In addition, estimate the residual for this person.

**2.203 Using Neck Circumference to Predict Body Fat** The regression line for predicting body fat percent using neck circumference is $\widehat{BodyFat} = -47.9 + 1.75 \cdot Neck$.

(a) What body fat percent does the line predict for a person with a neck circumference of 35 cm? Of 40 cm?

(b) Interpret the slope of the line in context.

(c) One of the men in the study had a neck circumference of 38.7 cm and a body fat percent of 11.3. Find the residual for this man.

**2.204 Cricket Chirps and Temperature** In the **CricketChirp** dataset given in Table 2.31 on page 108, we learn that the chirp rate of crickets is related to the temperature of the air.

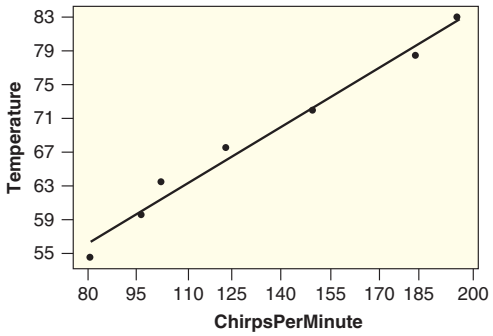



**Figure 2.73** *Draw a length representing a residual*

(a) Figure 2.73 shows the seven points together with the regression line. Does there appear to be a linear relationship between these two variables? How strong is it? Is it positive or negative?

(b) Use technology to find the formula for the regression line for the seven data points.

(c) Calculate the predicted values and the residuals for all seven data points.

[70] A sample taken from data provided by R. Johnson in "Fitting Percentage of Body Fat to Simple Body Measurements," *Journal of Statistics Education*, (1996), http://www.amstat.org/publications/jse/v4n1/datasets.johnson.html.

**2.205 Land Area and Rural Population** Two variables in the dataset **AllCountries** are the size of the country (in 1000 sq km) and the percent of the population living in rural areas. We are interested in using the size of the country (*LandArea*) to predict the percent rural (*Rural*). The values of these variables for a random sample of ten countries is shown, with the 3-letter country codes, in Table 2.37, and is also available in **TenCountries**. Figure 2.74 shows a scatterplot of the data.

**Table 2.37** *Land area (in 1000 sq km) and percent living in rural areas*

| Country | SRB | BHS | SVN | UZB | TUN | ARM | ROU | MKD | LBN | PRK |
|---|---|---|---|---|---|---|---|---|---|---|
| Land Area | 88.3 | 10.0 | 20.1 | 425.4 | 155.4 | 28.5 | 229.9 | 25.2 | 10.2 | 120.4 |
| Rural | 48.0 | 16.3 | 51.4 | 63.2 | 33.5 | 36.1 | 45.8 | 33.1 | 13 | 37.3 |

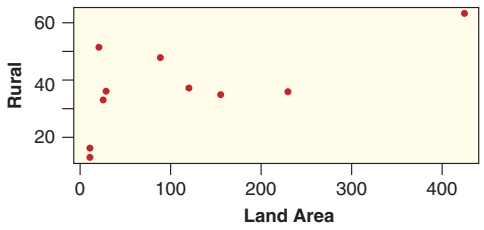



**Figure 2.74** *Scatterplot of land area and percent rural*

(a) What is the explanatory variable? What is the response variable?

(b) Without doing any calculations, which do you think is the most likely correlation between the two variables?

$$0.00, \quad 0.60, \quad -0.60, \quad 60$$

(c) Use technology to find the regression line to predict percent rural from land area, and interpret the slope.

(d) Does the intercept make sense in this situation?

(e) Which country is the most influential on this regression line (use the 3 letter code)?

(f) Use the regression line to predict the percent of the US population living in rural areas given that the area of the US is 9,147.4 thousand sq km in area. Does the prediction seem reasonable? Explain why it is not appropriate to use this regression line to predict the percent rural for the US.

**2.206 Adding One Point to Land Area and Rural Population** In Exercise 2.205, we used a random sample of 10 countries to use the size of a country to predict the percent of the population living in rural areas. We now see how results change if we add the United States (Land Area: 9147.4, Rural: 18.3%) to the sample.

(a) Use technology to find the new regression line using the 11 data points.

(b) The slope of the regression line using the original 10 points in Exercise 2.205 is about 0.08. Compare the slope with US added to the slope without US. Does adding US have a strong effect on the slope? Why or why not? (*Hint*: Plot the data!)

(c) Predict the percent rural for US with the new regression line. Is this prediction better than the prediction given in Example 2.205 (which was 752%)?

**2.207 Predicting World Gross Revenue for a Movie from Its Opening Weekend** Use the data in **HollywoodMovies2011** to use revenue from a movie's opening weekend (*OpeningWeekend*) to predict total world gross revenues by the end of the year (*WorldGross*). Both variables are in millions of dollars.

(a) Use technology to create a scatterplot for this relationship. Describe the scatterplot: Is there a linear trend? How strong is it? Is it positive or negative? Does it look like revenue from a movie's opening weekend is a good predictor of its future total earnings?

(b) The scatterplot contains an outlier in the top right corner. Use the dataset to identify this movie.

(c) Use technology to find the correlation between these variables.

(d) Use technology to find the regression line.

(e) Use the regression line to predict world gross revenues for a movie that makes 50 million dollars in its opening weekend.

**2.208 Using Life Expectancy to Predict Happiness** In Exercise 2.172 on page 114, we introduce the dataset **HappyPlanetIndex**, which includes information for 143 countries to produce a "happiness" rating as a score of the health and well-being of the country's citizens, as well as information on the ecological footprint of the country. One of the variables

used to create the happiness rating is life expectancy in years. We explore here how well this variable, *LifeExpectancy*, predicts the happiness rating, *Happiness*.

(a) Using technology and the data in **HappyPlanetIndex**, create a scatterplot to use *LifeExpectancy* to predict *Happiness*. Is there enough of a linear trend so that it is reasonable to construct a regression line?

(b) Find a formula for the regression line and display the line on the scatterplot.

(c) Interpret the slope of the regression line in context.

**2.209 Pick a Relationship to Examine** Choose one of the following datasets: **USStates**, **StudentSurvey**, **AllCountries**, or **NBAPlayers2011**, and then select any two quantitative variables that we have not yet analyzed. Use technology to create a scatterplot of the two variables with the regression line on it and discuss what you see. If there is a reasonable linear relationship, find a formula for the regression line. If not, find two other quantitative variables that do have a reasonable linear relationship and find the regression line for them. Indicate whether there are any outliers in the dataset that might be influential points or have large residuals. Be sure to state the dataset and variables you use.

**2.210 The Impact of Strong Economic Growth** In 2011, the Congressional Budget Office predicted that the US economy would grow by 2.8% per year on average over the decade from 2011 to 2021. At this rate, in 2021, the ratio of national debt to GDP (gross domestic product) is predicted to be 76% and the federal deficit is predicted to be $861 billion. Both predictions depend heavily on the growth rate. If the growth rate is 3.3% over the same decade, for example, the predicted 2021 debt-to-GDP ratio is 66% and the predicted 2021 deficit is $521 billion. If the growth rate is even stronger, at 3.9%, the predicted 2021 debt-to-GDP ratio is 55% and the predicted 2021 deficit is $113 billion.[71]

(a) There are only three individual cases given (for three different economic scenarios), and for each we are given values of three variables. What are the variables?

(b) Use technology and the three cases given to find the regression line for predicting 2021 debt-to-GDP ratio from the average growth rate over the decade 2011 to 2021.

(c) Interpret the slope and intercept of the line from part (b) in context.

(d) What 2021 debt-to-GDP ratio does the model in part (b) predict if growth is 2%? 4%?

(e) Studies indicate that a country's economic growth slows if the debt-to-GDP ratio hits 90%. Using the model from part (b), at what growth rate would we expect the ratio in the US to hit 90% in 2021?

(f) Use technology and the three cases given to find the regression line for predicting the deficit (in billions of dollars) in 2021 from the average growth rate over the decade 2011 to 2021.

(g) Interpret the slope and intercept of the line from part (f) in context.

(h) What 2021 deficit does the model in part (f) predict if growth is 2%? 4%?

(i) The deficit in 2011 was $1.4 trillion. What growth rate would leave the deficit at this level in 2021?

[71] Gandel, S., "Higher growth could mean our debt worries are all for nothing," *Time Magazine*, March 7, 2011, p. 20.