

Chapter 1: Stats Starts Here

Section 1.1

1. **Grocery shopping.** Discount cards at grocery stores allow the stores to collect information about the products that the customer purchases, what other products are purchased at the same time, whether or not the customer uses coupons, and the date and time that the products are purchased. This information can be linked to demographic information about the customer that was volunteered when applying for the card, such as the customer's name, address, sex, age, income level, and other variables. The grocery store chain will use that information to better market their products. This includes everything from printing out coupons at the checkout that are targeted to specific customers to deciding what television, print, or Internet advertisements to use.
2. **Online shopping.** Amazon hopes to gain all sorts of information about customer behavior, such as how long they spend looking at a page, whether or not they read reviews by other customers, what items they ultimately buy, and what items are bought together. They can then use this information to determine which other products to suggest to customers who buy similar items, to determine which advertisements to run in the margins, and to determine which items are the most popular so these items come up first in a search.

Section 1.2

3. **Super Bowl.** When collecting data about the Super Bowl, the games themselves are the *who*.
4. **Nobel laureates.** Each prize category is a case, holding all of the information (such as who won in this category for each year) about that specific category. Therefore, the prize category is the *who*. (Or, depending on your purpose, the data could also be organized by year, with years as cases, showing the winners for all categories for each single year.)

Section 1.3

5. **Grade level.**
 - a) If we are, for example, comparing the percentage of first-graders who can tie their own shoes to the percentage of second-graders who can tie their own shoes, grade-level is treated as categorical. It is just a way to group the students. We would use the same methods if we were comparing boys to girls or brown-eyed kids to blue-eyed kids.

2 Part I: Exploring and Understanding Data

- b) If we were studying the relationship between grade-level and height, we would be treating grade level as quantitative.
- 6. **ZIP codes.**
 - a) ZIP codes are categorical in the sense that they correspond to a location. The ZIP code 14850 is a standardized way of referring to Ithaca, NY.
 - b) ZIP codes generally increase as the location gets further from the east coast of the United States. For example, one of the ZIP codes for the city of Boston, MA is 02101. Kansas City, MO has a ZIP code of 64101, and Seattle, WA has a ZIP code of 98101.
- 7. **Affairs.** The response is a (nominal) categorical variable.
- 8. **Economic outlook.** The response is categorical. If we ignored the response 'Don't know', we could also say it is ordinal.
- 9. **Medicine.** The company is studying a quantitative variable.
- 10. **Stress.** The researcher is studying a quantitative variable.

Chapter Exercises

- 11. **The news** Answers will vary.
- 12. **The Internet** Answers will vary.
- 13. **Ovulation** *Who:* 40 undergraduate women. *What:* Whether or not the women could identify the sexual orientation of men based on a picture. *Why:* To see if ovulation affects a woman's ability to identify sexual orientation of a male. *How:* Showing very similar photos to the women, with half gay. *Variables:* Categorical variable: 'He's gay' or 'He's not gay'.
- 14. **Blindness** *Who:* 24 patients. *What:* Whether or not stem cell therapy is effective in treating Stargardt's disease or dry age-related macular degeneration. *When:* 2011. *Why:* To see if stem cell can treat these conditions. *Variables:* Some measure of vision or improvement in vision, perhaps on a precise quantitative scale (or at least ordinal), or if completely blind patients, possibly categorical (Blind, Not Blind).
- 15. **Investments** *Who:* 48 China/India/Chindia funds listed at globeinvestor.com. *What:* 1 month, 1 year, and 5 year returns for each fund. *When:* The most recent periods of time. *Where:* globeinvestor.com website. *Why:* To compare investment returns for future investment decisions. *How:* globeinvestor.com uses reports from the fund companies. *Variables:* There are three variables, all of which are quantitative. 1 month return; 1 year return; 5 year return, annualized; all variables are measured as percentages.

16. **Biomass of trees** *Who:* 25 trees from each of three different species in an unspecified B.C. forest. *What:* Species, diameter, and height of each tree. *When:* Not specified. *Where:* An unspecified B.C. forest. *Why:* To learn a cheap, simple method for estimating biomass of trees. *How:* Went out in the wild and used tape measures, a triangularization tool for height, and a picture sheet identifying the different species of tree. *Variables:* There are three variables. Species is a categorical variable, and diameter (unit of measurement not specified, but probably in cm) and estimated height (unit of measurement not specified, but probably in cm or m) are quantitative variables.
17. **Air travel** *Who:* All airline flights in Canada. *What:* Type of aircraft, number of passengers, whether departures and arrivals were on schedule, and mechanical problems. *When:* This information is currently reported. *Where:* Canada. *Why:* This information is required by Transport Canada and the Canadian Transportation Agency. *How:* Data is collected from airline flight information. *Variables:* There are four variables. Type of aircraft, departure and arrival timeliness, and mechanical problems are categorical variables, and number of passengers is a quantitative variable.
18. **Tracking sales** *Who:* Customers of a start-up company. *What:* Customer name, ID number, region of the country, date of last purchase, amount of purchase (probably in dollars), and item purchased. *When:* Present time. *Where:* Canada. *Why:* The company is building a database of customers and sales information. *How:* Presumably, the company records the information from each new customer and their purchase. *Variables:* There are six variables. Name, ID number, region of the country, and item purchased are categorical variables. Date and amount of purchase are quantitative variables. *Concerns:* Region is a categorical variable, and it is potentially confusing to record it as a number.
19. **Cars** *Who:* Automobiles. *What:* Make, country of origin, type of vehicle, and age of vehicle (probably in years). *When:* Not specified. *Where:* A large university. *Why:* Not specified. *How:* A survey was taken in campus parking lots. *Variables:* There are four variables. Make, country of origin, and type of vehicle are categorical variables, and age of vehicle is a quantitative variable.
20. **Stats students** *Who:* Students in a statistics class. *What:* Height (units not specified, but presumably in cm or metres), shoe size, sex, degree program, and birth order. *When:* Not specified. *Where:* Not specified. *Why:* The information was collected for use in classroom illustrations. *How:* An online survey was conducted. Presumably, participation was required for all members of the class. *Variables:* There are five variables. Sex and degree program are categorical variables. Height and shoe size are quantitative variables. Birth order can be either categorical or quantitative, depending on how it is used. For example, we could find the proportion of the

students who were second born. In this case, we are treating the variable as categorical. If we found the average birth order, we would be using the variable as a quantitative variable. *Concerns:* Shoe sizes for men and women are not measured on the same scale.

21. **Honesty** *Who:* Workers who buy coffee in an office. *What:* Amount of money contributed to the collection tray. *Where:* Newcastle. *Why:* To see if people behave more honestly when feeling watched. *How:* Counting money in the tray each week. *Variables:* Amount contributed (pounds) is a quantitative variable.
22. **Molten iron.** *Who:* 10 crankshafts at Cleveland Casting. *What:* The pouring temperature (in degrees Fahrenheit) of molten iron. *When:* Sometime before the 1995 journal article. *Where:* Cleveland. *Why:* To ensure the pouring temperature of molten iron is close to 2550 degrees. *How:* Not specified. *Variables:* Temperature (in degrees Fahrenheit) is a quantitative variable.
23. **Weighing bears** *Who:* 54 bears. *What:* Weight, neck size, length (no specified units), and sex. *When:* Not specified. *Where:* Not specified. *Why:* Since bears are difficult to weigh, the researchers hope to use the relationships between weight, neck size, length, and sex of bears to estimate the weight of bears, given the other more observable features of the bear. *How:* Researchers collected data on 54 bears they were able to catch. *Variables:* There are four variables; weight, neck size, and length are quantitative variables, and sex is a categorical variable. No units are specified for the quantitative variables. *Concerns:* The researchers are (obviously!) only able to collect data from bears they were able to catch. This method is a good one, as long as the researchers believe the bears caught are representative of all bears, in regard to the relationships between weight, neck size, length, and sex.
24. **Schools** *Who:* Students. *What:* Age (probably in years, though perhaps in years and months), race or ethnicity, days absent, current grade level, reading score, math score, and disabilities/special needs. *When:* This information must be kept current. *Where:* Not specified. *Why:* Keeping this information is a provincial/territorial requirement. *How:* The information is collected and stored as part of school records. *Variables:* There are seven variables. Race or ethnicity, grade level, and disabilities/special needs are categorical variables. Days absent, age, reading test score, and math test score are quantitative variables. *Concerns:* What tests are used to measure reading and math ability, and what are the units of measure for the tests?

25. **Tim Hortons doughnuts** *Who:* Doughnut types for sale at Tim Hortons. *What:* Various nutritional characteristics (see variables below). *When:* Not stated, but presumably the measurements were taken recently. *Where:* Tim Hortons website. *Why:* To help customers make good nutritional choices. *How:* Further research would be needed to learn how they made these measurements, but presumably at some specialized food analysis lab. *Variables:* There are eight variables, all quantitative: Number of calories (kcal/s), amounts of trans fat (g), total fat (g), sodium (mg), sugar (g), protein (g), % daily value of iron (percentage), and % daily value of calcium (percentage). Units (found at the website) are all per serving.
26. **Trudeaumania?** *Who:* Sample of 1006 adult Canadians. *What:* Gender, province, party preference, age group, view of Justin Trudeau. *When:* April 2013. *Where:* Respondents were spread across Canada. *Why:* For the public interest, and perhaps the interest of political parties. *How:* A telephone poll using Harris/Decima database and methodology. *Variables:* There are five variables, all categorical: gender, province, party preference, age group (18-34, 35-54, 55+), view of Justin Trudeau.
27. **Babies** *Who:* 882 births. *What:* Mother's age (in years), length of pregnancy (in weeks), type of birth (Caesarean, induced, or natural), level of prenatal care (none, minimal, or adequate), birth weight of baby (unit of measurement not specified, but probably grams), gender of baby (male or female), and baby's health problems (none, minor, major). *When:* 1998-2000. *Where:* Large city hospital. *Why:* Researchers were investigating the impact of prenatal care on newborn health. *How:* It appears that they kept track of all births in the form of hospital records, although it is not specifically stated. *Variables:* There are three quantitative variables: mother's age, length of pregnancy, and birth weight of baby. There are four categorical variables: type of birth, level of prenatal care, gender of baby, and baby's health problems.
28. **Flowers** *Who:* 385 species of flowers. *What:* Date of first flowering (in days). *When:* Not specified. *Where:* Southern England. *Why:* The researchers believe that this indicates a warming of the overall climate. *How:* Not specified. *Variables:* Date of first flowering is a quantitative variable. *Concerns:* Hopefully, date of first flowering was measured in days from January 1, or some other convention, to avoid problems with leap years.
29. **Herbal medicine.** *Who:* Experiment volunteers. *What:* Herbal cold remedy or sugar solution, and cold severity. *When:* Not specified. *Where:* Major pharmaceutical firm. *Why:* Scientists were testing the efficacy of an herbal compound on the severity of the common cold. *How:* The scientists set up a controlled experiment. *Variables:* There are two variables. Type of treatment (herbal or sugar solution) is categorical, and severity rating (on a scale from 0 to 5) is quantitative. *Concerns:* The severity of a cold seems subjective and difficult to quantify. Also, the scientists may feel pressure to report negative findings about the herbal product.

30. **Vineyards.** *Who:* Vineyards. *What:* Size of vineyard (in acres), number of years in existence, province, varieties of grapes grown, average case price (in dollars), gross sales (probably in dollars), and percent profit. *When:* Not specified. *Where:* Canada. *Why:* Business analysts hope to provide information that would be helpful to producers of Canadian wines. *How:* Not specified. *Variables:* There are five quantitative variables and two categorical variables. Size of vineyard (in acres), number of years in existence, average case price, gross sales, and percent profit are quantitative variables. Province and variety of grapes grown are categorical variables.
31. **Streams** *Who:* Streams. *What:* Name of stream, substrate of the stream (limestone, shale, or mixed), acidity of the water (measured in pH), temperature (in degrees Celsius), and BCI (unknown units). *When:* Not specified. *Where:* Northern Ontario. *Why:* Research is conducted for an ecology class. *How:* Not specified. *Variables:* There are five variables. Name and substrate of the stream are categorical variables, and acidity (pH), temperature (in degrees Celsius), and BCI are quantitative variables.
32. **Fuel economy** *Who:* Every model of automobile in Canada. *What:* Vehicle manufacturer, vehicle type, weight (probably in kg), horsepower, and gas mileage (in L/100km) for city and highway driving. *When:* This information is collected currently. *Where:* Canada. *Why:* Transport Canada and Natural Resources Canada use the information to track fuel economy of vehicles. *How:* The data are collected from the manufacturer of each model. *Variables:* There are six variables. City mileage, highway mileage (L/100km), weight, and horsepower are quantitative variables. Manufacturer and type of car are categorical variables.
33. **Refrigerators** *Who:* 41 models of refrigerators. *What:* Brand, cost (probably in dollars), size (in cu. ft.), type, estimated annual energy cost (probably in dollars), overall rating, and repair history (in percent requiring repair over the past five years). *When:* 2002-2006. *Where:* United States. *Why:* The information was compiled to provide information to the readers of *Consumer Reports*. *How:* Not specified. *Variables:* There are seven variables. Brand, type, and overall rating are categorical variables. Cost, size (cu. ft.), estimated energy cost, and repair history (percentage) are quantitative variables.
34. **Walking in Circles** *Who:* 32 volunteers. *What:* Sex, height, handedness, the number of yards walked before going out of bounds, and the side of the field on which the person walked out of bounds. *When:* Not specified. *Where:* Not specified. *Why:* The researcher was interested in whether people walk in circles when lost. *How:* Data were collected by observing the people on the field, as well as by measuring and asking the participants. *Variables:* There are five variables. Sex, handedness, and side of the field are categorical variables. Height and number of yards walked are quantitative variables.

- 35. Kentucky Derby 2015** *Who:* Kentucky Derby races. *What:* Year, winner, jockey, trainer, owner, and time (in minutes, seconds, and hundredths of a second). *When:* 1875 to 2015. *Where:* Churchill Downs, Louisville, Kentucky. *Why:* It is interesting to examine the trends in the Kentucky Derby. *How:* Official statistics are kept for the race each year. *Variables:* There are six variables. Date, Winner, jockey, trainer, and owner are categorical variables. Duration is quantitative variables.
- 36. The Stanley Cup 2015** *Who:* Annual Stanley Cup championship hockey series. *What:* Year, winner, loser, series outcome, and time and scorer of the winning goal in the final game. *When:* Annually from 1926–2015. *Where:* Various cities in North America. *Why:* General public/sports fan interest. *How:* Official records found at www.tmlfever.com. *Variables:* There are six variables. Time of winning goal is quantitative variables. Year, winner, loser, and scorer of winning goal are categorical variables. Series results can be viewed in different ways: listing all possible results (categorical), or wins by winner (quantitative) and wins by loser (quantitative), or length of series (quantitative). The variable 'Series (W-L)' though is not really measuring the same thing until after 1939, when the best-of-seven format stabilizes. The series format can be viewed as a categorical variable, too.