

## 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.

Retailers 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.

## 3. Parking lots.

The owners of the parking garage can advertise about the availability of parking. They can also communicate with businesses about hours when more spots are available and when they should encourage more business.

## 4. Satellites and global climate change.

This rise and fall of temperature and water levels can help in planning for future problems and guide public policy to protect our safety.

## Section 1.2

## 5. Super Bowl.

When collecting data about the Super Bowl, the games themselves are the *Who*.

## 6. Nobel laureates.

Each year is a case, holding all of the information about that specific year. Therefore, the year is the *Who*.

## 7. Health records.

The sample is about 5000 people, and the population is all residents of the United States of America. The *Who* is the selected subjects and the *What* includes medical, dental, and physiological measurements and laboratory test results.

## 8. Facebook.

The *Who* is the 350 million photos. The *What* might be information about the photos, for example: file format, file size, time and date when uploaded, people and places tagged, and GPS information.

## Section 1.3

## 9. Grade levels

- 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.
- b) If we were studying the relationship between grade-level and height, we would be treating grade level as quantitative.

## 2 *Part I Exploring and Understanding Data*

### 10. 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.

### 11. Voters.

The response is a categorical variable.

### 13. Medicine.

The company is studying a quantitative variable.

### 12. Job hunting.

The answer is a categorical variable.

### 14. Stress.

The researcher is studying a quantitative variable.

## Section 1.4

### 15. Voting and elections.

Pollsters might consider whether a person voted previously or whether he or she could name the candidates. Voting previously and knowing the candidates may indicate a greater interest in the election.

### 16. Weather.

Meteorologists can use the models to predict the average temperature ten days in advance and compare their predictions to the actual temperatures.

### 17. The News.

Answers will vary.

### 18. The Internet.

Answers will vary.

### 19. Gaydar.

*Who* – 40 undergraduate women; *What* – Whether or not the women could identify the sexual orientation of men based on a picture; *Population of interest* – All women.

### 20. Hula-hoops.

*Who* – An unknown number of participants; *What* – Heart rate, oxygen consumption, and rating of perceived exertion; *Population of interest* – All people.

### 21. Bicycle Safety.

*Who* – 2500 cars; *What* – Distance from the bicycle to the passing car (in inches); *Population of interest* – All cars passing bicyclists.

### 22. Investments.

*Who* – 30 similar companies; *What* – 401(k) employee participation rates (in percent); *Population of interest* – All similar companies.

### 23. Honesty.

*Who* – Coffee drinkers at a Newcastle University coffee station; *What* – amount of money contributed to the collection tray; *Population of interest* – All people in honor system payment situations.

### 24. Blindness.

*Who* – 24 patients suffering from blindness; *What* – Whether the patient had Stargardt's disease or dry age-related macular degeneration, and whether or not the stem cell therapy was effective in treating the condition; *Population of interest* – All people with these eye conditions.

### 25. Not-so-diet soda.

*Who* – 474 participants; *What* – whether or not the participant drank two or more diet sodas per day, waist size at the beginning of the study, and waist size at the end of the study; *Population of interest* – All people.

**26. Molten iron.**

*Who* – 10 crankshafts at Cleveland Casting; *What* – The pouring temperature (in degrees Fahrenheit) of molten iron; *Population of interest* – All crankshafts at Cleveland Casting.

**27. Weighing bears.**

*Who* – 54 bears; *Cases* – Each bear is a case; *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 4 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.

**28. Schools.**

*Who* – Students; *Cases* – Each student is an individual case; *What* – Age (probably in years, though perhaps in years and months), race or ethnicity, number of absences, 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 state 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. Number of absences (days), age (years?), 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?

**29. Arby's menu.**

*Who* – Arby's sandwiches; *Cases* – Each sandwich is a case; *What* – type of meat, number of calories (in calories), and serving size (in ounces); *When* – Not specified; *Where* – Arby's restaurants; *Why* – These data might be used to assess the nutritional value of the different sandwiches; *How* – Information was gathered from each of the sandwiches on the menu at Arby's, resulting in a census. *Variables* – There are three variables. Number of calories and serving size (ounces) are quantitative variables, and type of meat is a categorical variable.

**30. Age and party.**

*Who* – 1180 Americans; *Cases* – Each of the 1180 Americans surveyed is a case in this poll; *What* – Region, age (in years), political affiliation, and whether or not the person voted in the 2006 midterm Congressional election; *When* – First quarter of 2007; *Where* – United States; *Why* – The information was gathered for presentation in a Gallup public opinion poll. *How* – Phone Survey; *Variables* – There are four variables. Region, political affiliation, and whether or not the person voted in 1998 are categorical variables, and age is a quantitative variable.

**31. Babies.**

*Who* – 882 births; *Cases* – Each of the 882 births is a case. *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 pounds and ounces), 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 (years), length of pregnancy (years), 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.

## 4 *Part I Exploring and Understanding Data*

### 32. Flowers.

*Who* – 385 species of flowers; *Cases* – Each of the 385 species at each of the 47 years is a case. *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.

### 33. Herbal medicine.

*Who* – experiment subjects; *Cases* – Each subject is a case. *What* – herbal cold remedy or sugar solution, and cold severity (0 to 5 scale); *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 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.

### 34. Vineyards.

*Who* – American Vineyards; *Cases* – Each vineyard is a case. *What* – Size of vineyard (in acres), number of years in existence, state, varieties of grapes grown, average case price (in dollars), gross sales (probably in dollars), and percent profit; *When* – Not specified. *Where* – United States; *Why* – Business analysts hoped to provide information that would be helpful to producers of American wines. *How* – Not specified. *Variables* – There are five quantitative variables and two categorical variables. Size of vineyard, number of years in existence, average case price, gross sales, and percent profit are quantitative variables. State and variety of grapes grown are categorical variables.

### 35. Streams.

*Who* – Streams; *Cases* – Each stream is a case; *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* – Upstate New York; *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, temperature, and BCI are quantitative variables.

### 36. Fuel economy.

*Who* – Every model of automobile in the United States. *Cases* – Each vehicle model is a case. *What* – Vehicle manufacturer, vehicle type, weight (probably in pounds), horsepower (in horsepower), and gas mileage (in miles per gallon) for city and highway driving; *When* – This information is collected currently. *Where* – United States; *Why* – The Environmental Protection Agency uses the information to track fuel economy of vehicles. *How* – The data is collected from the manufacturer of each model. *Variables* – There are six variables. City mileage, highway mileage, weight, and horsepower are quantitative variables. Manufacturer and type of car are categorical variables.

### 37. Refrigerators.

*Who* – 353 refrigerator models; *Cases* – Each of the 353 refrigerator models is a case. *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* – 2013; *Where* – United States; *Why* – The information was compiled to provide information to the readers of *Consumer Reports*. *How* – Not specified. *Variables* – There are 7 variables. Brand, type, and overall rating are categorical variables. Cost, size, estimated energy cost, and repair history are quantitative variables.

### 38. Walking in circles.

*Who* – 32 volunteers; *Cases* – Each volunteer is a case. *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 5 variables. Sex, handedness, and side of the field are categorical variables. Height and number of yards walked are quantitative variables.

**39. Kentucky Derby 2018.**

*Who* – Kentucky Derby races; *What* – Year, winner, jockey, trainer, owner, and time (in minutes, seconds, and hundredths of a second); *When* – 1875 to 2018; *Where* – Churchill Downs, Louisville, Kentucky; *Why* – Not specified. To examine the trends in the Kentucky Derby? *How* – Official statistics are kept for the race each year. *Variables* – There are 8 variables. Winner, jockey, trainer and owner are categorical variables. Date and duration are quantitative variables.

**40. Indy 2018.**

*Who* – Indy 500 races; *What* – Year, winner, time (in minutes, seconds, and hundredths of a second), and average speed (in miles per hour); *When* – 1911 – 2018; *Where* – Indianapolis, Indiana; *Why* – Not specified. To examine the trends in Indy 500 races? *How* – Official statistics are kept for the race every year; *Variables* – There are 4 variables. Driver is a categorical variable. Year, time, and speed are quantitative variables.

**41. Kentucky Derby 2018 on the computer.**

- a) Fonso was the winning horse in 1880.
- b) The length of the race changed in 1895, from 1.5 miles to 1.25 miles.
- c) The winning time in 1974 was 124 seconds.
- d) Secretariat ran the Derby in under 2 minutes in 1973.

**42. Indy 500 2018 on the computer.**

- a) The average speed of the winner in 1920 was 88.618 miles per hour.
- b) Bill Vukovich won the Indy 500 twice in the 1950s.
- c) There were only 6 Indy 500 races in the 1940s.

