Chapter 2 – Data

SECTION EXERCISES

SECTION 2.1

1.

a) Each row represents a different house that was recently sold. It is best described as a case.
b) Including the house identifier, there are six variables in each row.

2.

a) Each row represents a different transaction (not customer or book). It is best described as a case.
b) Including the transaction identifier, there are eight variables in each row.

SECTION 2.2

3.

a) House_ID is an identifier (special type of categorical); Neighbourhood is categorical (nominal); YR_BUILT is quantitative (units—year), but could also be treated as categorical (ordinal); FULL_MARKET _VALUE is quantitative (units—dollars); SFLA is quantitative (units—sq. ft.).
b) These data are cross-sectional. Each row corresponds to a house that recently sold—at approximately the same fixed point in time.

4.

a) Transaction ID is an identifier (special type of categorical); Customer ID is an identifier (special type of categorical); Date is categorical or may be treated as numerical if redefined as how many days ago the transaction took place; ISBN is an identifier (special type of categorical); Price is quantitative (units—dollars); Coupon is categorical (simply nominal); Gift is categorical (simply nominal); Quantity is quantitative (unit—counts).
b) These data are cross-sectional. Each row corresponds to a transaction at a fixed point in time. However the date of the transaction has been recorded. Consequently, since a time variable is included, the data could be reconfigured as a time series.

SECTION 2.3

5. The real estate data in Exercise 1 are not from a designed survey or experiment. Rather, the real estate major's data set was derived from transactional data (on local home sales). The major concern with drawing conclusions from this data set is that we cannot be sure that the sample is representative of the population of interest (e.g., all recent local home sales or even all recent national home sales).

6. The student is using a secondary data source (from the Internet). The main concerns about using these data for drawing conclusions is that the data were collected for a different purpose (not necessarily for developing a stock investment strategy) and information about how, when, and why they were collected may not be available.

CHAPTER EXERCISES

7. Canadian labour force.

a) Someone on vacation from a full-time job is employed (they are on the payroll of their employer).
b) Someone who is not working, has a job offer, but is trying to find a better offer is unemployed because they are not working but are available and searching for work.
c) Someone who looked for work until six months ago but then gave up looking is not in the labour force because they are not searching for work.

8. Non-employment in Canada

a) The labour force

b) Unemployed and non-employed are both "available" for work. Unemployed are the subgroup that are actively "searching." So the answer is yes.

9. Domestic credit in Canada. Imagine collecting the data and putting it in a table with one row for each year and two columns corresponding to the two variables: domestic credit and GDP. The rows identify the *Who* of the data (i.e., the years) and the column headings identify the *What* (domestic credit and GDP). *When* is recent years, *Where* is Canada, and *Why* is to investigate possible future trends. The two variables of domestic credit and GDP are both quantitative and measured in $ billion. *Concerns*—none.

10. Oil spills. The description of the study has to be broken down into its components in order to understand it. *Who*—50 tankers having recent major oil spills**.** *What*—what is being measured–:date, spillage amount (no specified unit), and cause of puncture. *When*—recent years. *Where*—United States**.** *Why*—not specified but probably to determine whether spillage amount per oil spill has decreased. *How*—how was the study conducted Not specified, although it is mentioned that the data are online**.** *Variables*—what is the variable being measured There are three variables—the date, the spillage amount (which is quantitative), and the cause of the puncture (which is categorical). *Concerns*—more detail needed on the specifics of the study.

11. Sales. The description of the study has to be broken down into its components in order to understand it. *Who*—who or what was actually sampled: months at a major Canadian company. *What*—what is being measured: money spent on advertising ($ thousands) and sales ($ million). *When*—monthly for the past three years; *Where* Canada (assumed). *Why*—to compare money spent on advertising to sales. *How*—how was the study conducted: not specified. *Variables*—what is the variable being measured There are three variables—the date, the amount of money spent on advertising (which is quantitative), and sales (which are quantitative). *Concerns*—none.

12. Food store. *Who*—who or what was actually sampled: existing stores**.** *What*—what is being measured: weekly sales ($), town population (thousands), median age of town (years), median income of town ($), and whether the stores sell beer/wine. *When*—not specified. *Where*—Canada. *Why*—the food retailer is interested in understanding if there is an association among these variables in order to determine where to open the next store. *How*—how was the study conducted: data collected from their stores. *Variables*—what is the variable being measured: sales ($), town population (thousands), median age of town (years), and median income of town ($), which are all quantitative. Whether or not the stores sell beer/wine is categorical. *Concerns*—none.

13. Sales II. *Who*—who or what was actually sampled: quarterly data from a major Canadian company. *What*—what is being measured: quarterly sales ($ millions), unemployment rate (%), inflation rate (%). *When*—quarterly for the past three years. *Where*—Canada. *Why*—to determine how sales are affected by the unemployment rate and inflation rate. *How*—how was the study conducted: not specified. *Variables*—what is the variable being measured: quarterly sales ($ millions), unemployment rate (%), and inflation rate (%), which are quantitative. *Concerns*—none.

14. Subway's menu. *Who*—Subway sandwiches. *What*—type of meat, number of calories, and serving size (in ounces). *When*—not specified. *Where*—Subway restaurants. *Why*—to assess the nutritional value of the different sandwiches. *How*—information gathered on each of the sandwiches offered on the menu. *Variables*—the number of calories and serving size (ounces) are quantitative, and the type of meat is categorical. *Concerns*—none.

15. MBA admissions. *Who*—MBA applicants. *What*—sex, age, whether or not accepted, whether or not attended, and reasons for not attending (if they did not attend). *When*—not specified. *Where*—the school. *Why*—the researchers wanted to investigate any patterns in female student acceptance and attendance in the MBA program. *How*—data obtained from the admissions office. *Variables*—sex, whether or not students accepted, whether or not they attended, the reasons for not attending (all categorical), and age (years), which is quantitative. *Concerns*—none.

16. **Climate.** *Who*—385 species of flowers. *What*—date of first flowering (in days). *When*—data gathered over the course of 47 years. *Where*—southern England. *Why*—the researchers wanted to investigate if the first flowering is indicating a warming of the overall climate. *How*—not specified. *Variables*—date of first flowering is a quantitative variable. *Concerns*–date of first flowering should be measured in days from January 1 to address leap year issues.

17. **MBA admissions II.** *Who*—MBA students. *What*—each student's standardized test scores and GPA in the MBA program. *When*—the past five years. *Where*—London. *Why*—to investigate the association between standardized test scores and performance in the MBA program over the past five years. *How*—not specified. *Variables*—standardized test scores and GPA, both quantitative variables. *Concerns*–none.

18. **Canadian schools.** *Who*—students. *What*—age (years or years and months), number of days absent, grade level, reading score, math score, and any disabilities/special needs. *When*—ongoing and current. *Where*–a Canadian province. *Why*—keeping this information is a provincial requirement. *How*—data are collected and stored as part of school records. *Variables*—there are seven variables. Grade level and disabilities/special needs are categorical variables. Number of absences, age (years or years and months), reading scores, and math scores are quantitative variables. *Concerns*—what tests are used to measure reading and math ability and what are the units of measurement?

19. **Pharmaceutical firm.** *Who*—experimental participants. *What*—herbal cold remedy or sugar solution, and cold severity. *When*—not specified. *Where*—major pharmaceutical firm. *Why*—scientists were testing the effectiveness of a herbal compound on the severity of the common cold. *How*—scientists conducted 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 might be difficult to quantify (beneficial to add actual observations and measurements, such as body temperature). Also, scientists at a pharmaceutical firm could have a predisposed opinion about the herbal solution or may feel pressure to report negative findings about the herbal product.

20. **Startup company.** *Who*—customers of a startup company. *What*—customer name, ID number, region of the country, date of last purchased, amount of purchase ($), and item purchased. *When*—present day. *Where*—Canada (assumed). *Why*—the company is building a database of customers and sales information. *How*—assumed that the company records the needed information from each new customer. *Variables*—there are six variables: name, ID number, region of the country, and item purchased, which are categorical, and date and amount of purchase ($), which are quantitative. *Concerns*—although region is coded as a number, it is still a categorical variable.

21. **Cars.** *Who*—cars parked in executive and staff lots at a large company. *What*—make, country of origin, type of vehicle (car, van, SUV, etc.), and age of vehicle (probably in years). *When*—not specified. *Where*—a large company. *Why*—not specified. *How*—data recorded in executive and staff lots of a large company. *Variables*—make, country of origin, and type of vehicle are three categorical variables. Age is the single quantitative variable. Whether or not the vehicle is in an executive or staff lot is also a categorical variable. *Concerns*—none.

22. **Canadian vineyards.** *Who*—vineyards. *What*—size, number of years in existence, province, varieties of grapes grown, average case price, gross sales, and percent profit. *When*—not specified. *Where*—Canada. *Why*—business analysts hope to provide information that would be helpful to grape growers in Canada. *How*—not specified. *Variables*—size of vineyard (hectares), number of years in existence, average case price ($), gross sales ($), and percent profit are five quantitative variables. Province and variety of grapes grown are categorical variables. *Concerns*—none.

23. **Environment.** *Who*—streams. *What*—name of stream, substrate of the stream (limestone, shale, or mixed), acidity of the water (measured in PH), temperature (degrees Celsius), and BCI (a measure of biological diversity—unknown units). *When*—not specified. *Where*—Alberta. *Why*—research conducted for an ecology class. *How*—not specified. *Variables*—there are five variables. Name of stream and substrate of the stream (limestone, shale, or mixed) are categorical variables. Acidity of the water (PH), temperature (degrees Celsius), and BCI are quantitative variables. *Concerns*—none.

24. Canadian voters. *Who*—1180 Canadian voters. *What*—region, age, party affiliation, whether or not the person owned any shares of stock, and their attitude toward unions. *When*—not specified. *Where*—Canada. *Why*—the information was gathered as part of a Gallup public opinion poll. *How*—telephone survey. *Variables*—there are five variables. Region (East, West, Prairie, etc.), party affiliation, and whether or not the person owned any shares of stock are categorical variables. Age (in years) and attitude (scale of 1 to 5) toward unions are quantitative variables. *Concerns*—none.

25. CTA. *Who*—all airline flights in Canada. *What*—type of aircraft, number of passengers, whether departures and arrivals were on schedule, and mechanical problems. *When*—the information is currently recorded. *Where*—Canada. *Why*—the information is required by the CTA. *How*—data is collected from airline flight information. *Variables*—there are four variables. Type of aircraft, whether departures and arrivals were on schedule, and mechanical problems are categorical variables. Number of passengers is a quantitative variable. *Concerns*—none.

26. Mobile phones. *Who*—mobile phone manufacturers. *What*—sales. *When*—pastthree years. *Where*—worldwide. *Why*—to project the future of the mobile phone business. *Variable*—sales, which is a quantitative variable measured in $. *Concerns*—none

27. Canadian families. *Who*—all Canadians (since it is a census). *What*—family type. *When*—every five years. *Where*—Canada. *Why*—to investigate social trends. *Variable*—family type, which is a categorical variable. *Concerns*—none

28. Canadian oil and gas production. *Who*—crude oil, natural gas, natural gas by-products. *What*—value and volume. *When*—every year. *Where*—Canada. *Why*–not specified. *Variables*—value and volume, both quantitative (measured in $ and m3, respectively). *Concerns*—none

29. Overnight visitors to Canada. *Who*—overnight visitors to Canada. *What*—number of nights spent in Canada and money spent in Canada. *When*—each year. *Where*—Canada. *Why*—to provide information for the tourism industry. *Variables*—number of nights spent in Canada and money spent in Canada, both of which are quantitative variables (with no units and with units of $, respectively). *Concerns*—none

30. Stock market. *Who*—students in an MBA statistics class. *What*—total personal investment in stock market ($), number of different stocks held, total invested in mutual funds ($), and the name of each mutual fund. *When*—not specified. *Where*—a business school in Toronto. *Why*—the information was collected for use in classroom illustrations. *How*—an online survey was conducted, and participation was probably required for all members of the class. *Variables*—there are four variables. Total personal investment in stock market ($), number of different stocks held, and total invested in mutual funds ($) are quantitative variables. The name of each mutual fund is a categorical variable. *Concerns*—none.

31. Theme park sites. *Who*—potential theme park locations. *What*—country of site, estimated cost (in Euros), potential population size within one hour drive of site (counts), size of site (hectares), whether or not mass transportation is within five minutes of site. *When*—2017. *Where*—Europe. *Why*—present to potential developers on the feasibility of various sites. *How*—not specified. *Variables*—there are five variables. Country of site and whether or not mass transportation is within five minutes of site are both categorical variables. Estimated cost (€), potential population size (counts), and size of site (hectares) are quantitative. *Concerns*—none.

32. Indy. *Who*—Indianapolis 500 races. *What*—year, winner, car model, time (hrs), speed (mph), and car number. *When*—1911–2012. *Where*—Indianapolis, Indiana. *Why*—examine trends in Indianapolis 500 race winners. *How*—official statistics kept for each race every year. *Variables*—there are six variables. Winner, car model, and car number are categorical variables. Year, time (hrs), and speed (mph) are quantitative variables. *Concerns*—none.

33. Kentucky Derby. *Who*—Kentucky Derby races. *What*—date, winner, winning margin (in lengths), jockey, winner's payoff ($), duration of the race (minutes and seconds), and track conditions. *When*—1875–2012. *Where*—Churchill Downs, Louisville, Kentucky. *Why*—examine trends in Kentucky Derby winners. *How*—

official statistics kept for each race every year. *Variables*—there are seven variables. Winner, winning jockey, and track conditions are categorical variables. Date, winning margin (in lengths), winner's payoff ($), and duration of the race (minutes and seconds) are quantitative variables. *Concerns*—none.

34. Mortgages. Each row represents each individual mortgage loan. Headings of the columns would be: borrower name and mortgage amount ($).

35. Employee performance. Each row represents each individual employee. Headings of the columns would be Employee ID Number (to identify the row instead of the name), contract average ($), supervisor's rating (1–10), and years with the company.

36. Company performance. Each row represents a week. Headings of the columns would be week number of the year (to identify each row), sales prediction ($), sales ($), and difference between predicted sales and realized sales ($).

37. Command performance. Each row represents a Broadway show. Headings of the columns would be the show name (identifies the row), profit or loss ($), number of investors, and investment total ($).

38. Car sales. Cross-sectional are data taken from situations that vary over time but are measured at a single time. This problem focuses on data for September only, which is a single time period. Therefore, the data are cross-sectional.

39. Motorcycle sales. Time series data are measured over time. Usually the time intervals are equally spaced (e.g., every week, every quarter, or every year). This problem focuses on the number of motorcycles sold by the dealership in each month of last year; therefore, the data are measured over a period of time and are time series data.

40. Cross sections. Time series data are measured over time. Usually the time intervals are equally spaced (e.g., every week, every quarter, or every year). This problem focuses on the average diameter of trees brought to a sawmill in each week of a year; therefore, the data are measured over a period of time and are time series data.

41. Series. Cross-sectional data are taken from situations that vary over time but are measured at a single time. This problem focuses on data for attendance at the third World Series game. Therefore, the data are cross-sectional.

42. Canadian immigrants.

a) *Who*: years; *What*: percentage unemployment rates; *When*: 2009–2013; *Where*: Canada; *Why*: to compare unemployment rates among demographic groups; *How*: complied from the Statistics Canada Labour Force Survey.

b)

i. All data are quantitative.
ii. For a given year, the data are cross-sectional. Overall the data are time series.
iii. All data are secondary, since they are derived from the primary interview data in the Canadian Labour Force Survey.

43. Interpreting Published Data. Answers will vary.

Mini Case: Ottawa Senators

PLAN	Setup Clarify the objective.	Identify the types of data in the data file.
DO	Mechanics Describe the W's for the data.	**WHAT:** The answer comes from the column headings. The shooter's first and last names, the team, the shooter's positions, total shots, total goals, shooting percentage, and the number of game-deciding goals. **WHO:** The shooter, since this is what each row of the table is about. **WHERE:** The data are from the NHL for games played across North America. **WHEN:** Data entries were recorded after each game in the 2007–08 season. **WHY:** The data has been collected to record the results of NHL shootouts.
	Identify the types of variables that the data consists of.	**Variable Type** The shooter's name, team, and position are primary data, since they have not been processed or summarized. The rest of the data are secondary, since they are a summary of the results of shots taken during multiple games and have therefore been processed. The whole table is cross-sectional, since it all applies to the 2007–08 season. The data on the shooter's name, team, and position are categorical and the rest of the data are quantitative.
REPORT	Conclusion Summarize the results.	We have identified the W's and the variable types of our data. The data is partly primary and partly secondary. The data is cross-sectional. The data is partly categorical and partly quantitative.

Mini Case: Credit Card Company

PLAN	Setup: State the objective	To gain a clear understanding of the data available.
DO	Mechanics: *List the W's for these data:*	Large format tables and graphs (if any) are placed below this PLAN/DO/REPORT table *Who* – company cardholders *What* – offer status (type of offer made to cardholder), credit card charges made by cardholder in August 2008, September 2008, and October 2008, marketing segment, industry segment, amount of spend lift after promotion, average spending on card pre- and post-promotion, whether or not cardholder is a retail customer or enrolled in the program and whether or not the spend lift was positive.

	Classify each variable as categorical or quantitative; if quantitative identify the units:	*Why* – to determine what types of offers are most effective in increasing credit card spending *When* – most likely in 2008 *Where* – not specified *How* – demographic data most likely collected when credit card account was opened and spending data collected during transactions *Variables:* *Offer Status* – categorical *Charges August 2008* – quantitative ($) *Charges September 2008* – quantitative ($) *Charges October 2008* – quantitative ($) *Marketing Segment* – categorical *Industry Segment* – categorical *Spend Lift After Promotion* – quantitative ($) *Pre Promotion Avg Spend* – quantitative ($) *Post Promotion Avg Spend* – quantitative ($) *Retail Customer* – categorical *Enrolled in Program* – categorical *Spend Lift Positive* – categorical
REPORT	Conclusion: State the conclusion in the context of the original objective	We have clarified what the data consist of and the details are given above.

Mini Case: Canadian Immigrants

PLAN	State the objective in its context.	Clarify the implications of the differences in unemployment rates of immigrants and people born in Canada
DO	Mechanics (a) Which data was Anjali referring to? (b) What other explanations are there of the data other than "Canadian employers are against immigrants."	 (a) The data on Canadian-born males indicate a marked difference in unemployment from 6.2–8.5% for high school graduates to 2.9–3.7%, for university graduates. However for immigrant females, the difference is only from 8.8–11.8% to 7.7–9.2%. (b) Employers could have clear reasons for not employing immigrants, other than being against them— e.g., poor knowledge of English/French, graduation from an overseas university with lower standards than the average Canadian university, graduation from a Canadian university. Also some immigrants, such as doctors, are required to re-qualify in order to work in Canada, and many become unemployed in the interim.

	(c) What additional data do you suggest Statistics Canada should collect in order to clarify this issue?	(c) Statistics Canada could collect data on the universities attended by people born in Canada and by immigrants, and the factors employers consider when hiring employees (in addition to their academic qualifications)
REPORT	Conclusion. State the conclusion in the context of the original objective.	The data clearly show higher unemployment rates for immigrants than for people born in Canada even when they have the same level of academic qualifications. In order to draw conclusions from these data, more detail is needed on whether the academic qualifications are comparable and what other factors employers consider.

Chapter 2
Data

What's It About?

In this chapter we introduce students to data. We talk about the importance of context (the W's), about variables, and make the distinction between categorical and quantitative data. We begin to introduce the vocabulary of Statistics.

Comments

It is valuable to get students involved with data from the start. We don't take a "big picture" approach at this time. There will be plenty of time to build models and draw inferences later. For now, let's just get our hands dirty playing with the data. When students have a good sense of what kinds of things data can say to us, they learn to expect to listen to the data. Throughout the course, we insist that no analysis of data is complete without telling what it means. This is where that understanding starts.

Rather than head directly for the "real purpose" of the course in the inference chapters, we prefer to emphasize the connection between our work with data and what they tell us about the world. No analysis is complete without a connection back to the real-world circumstances. Setting that stage is the underlying motivation for this chapter. We'll spend the next 5 chapters or so looking at and exploring data without making formal inferences.

Looking Ahead

You might have the students thumb through the book and read the opening of some chapters. Each one starts with a story about a company or business sector and data, and most have additional stories and more data inside. Statistics is about the real world. Among other topics, we'll be discussing Future Shop, MBNA, PotashCorp, Whole Foods Market, and even Canada's Wonderland. We need to get students thinking about the context of data and able to make the distinction between categorical and quantitative data. These are fundamental skills for everything that follows, and they'll be used throughout the course.

Class Do's

Get the class thinking about what the term "data" means. Students need to understand that data are not just numbers and that they must have a context (the W's). When data are quantitative, they should also have units. There are two ways we treat data: ***categorical*** and ***quantitative***. Don't get distracted by worrying about ratio, interval, and other distinctions. These are problematic and don't matter for the concepts and methods discussed in this book. Emphasize that the distinction between treating data as categorical or quantitative may be more about how *we* display and analyze data than it is about the variable itself. The variable "sex" is data, but just because we might label the males as 1 and the females as 0 doesn't mean that it's quantitative. On the other hand, taking the average of those 0's and 1's does give us the percentage of males. How about *age*? It is often quantitative, but could be categorical if broken down only into *child*, *adult*, and *senior*. Students should recognize that every discipline has its own vocabulary, and Statistics is no exception. They'll need to understand and use that vocabulary properly. Unfortunately, many Statistics words have a common everyday usage that's not quite the same. We'll be pointing those out as we go along.

Emphasize vocabulary words as they come up. One of the first should be *variable*. Make the point that it does not mean exactly the same thing as it did in Algebra. There, we call "x" a variable, but

often that means that we just don't currently know its value. In Statistics a variable is an attribute or characteristic of an individual or object whose value varies from case to case.

Point out that summaries of data can be verbal, visual, and numerical. All are important. In fact, any complete analysis of data almost always includes all three of these.

Hope that someone objects to finding an overall average shoe size or to comparing men's and women's sizes—shoe sizes are inconsistent in terms of units. This adds emphasis to the importance of units and the W's.

The Importance of What You Don't Say

We are laying a foundation here. Stretching up to the attic at this point just makes everyone feel unsafe. Many fundamental Statistics terms are left unmentioned in this chapter. You'll find it best to leave it that way. We'll get to them when the students have a safe place to file them along with their other knowledge. So we have an unusually long list of terms we recommend leaving for later in the course. In particular, avoid saying the following:

Hypothesis, Inference. These are certainly important in this course, but we have no background for discussing them honestly now, so they would just be confusing and intimidating.

Nominal, Ordinal, Interval, Ratio. "Nominal" is used by some software packages as a synonym for "categorical" as "continuous" is used for "quantitative." These distinctions arise from studies of measurement scales. But it isn't correct to claim that each variable falls into one of these categories. It is the use to which the data are put that determines what properties the variable must have. Ordinal categorical data may come up, but there are no special techniques for dealing with ordered categories in this course. And any differences between interval- and ratio-scaled data are commonly ignored in statistical analyses. If any of these terms were mentioned now, they'd never come up again anyway.

Random, Probability, Correlation. Everyone has some intuitive sense of these terms, and we'll deal with them formally—but not for a while. Students may want to use these terms, but at this early stage in the course, we don't need them. Without background and careful definition, they are likely to be misused and can simply be frightening.

Class Examples

1. Ask students to tell some things they learned about the class from inspecting the data collected in the opening day's survey. You can use that discussion to develop several of the important points of the chapter.
2. Consider 17, 21, 44, and 76. Are those data? Context is critical—they could be test scores, ages in a golf foursome, or uniform numbers of the starting backfield on the football team. In each case, our reaction changes.
3. Run through some other examples of data, asking about the W's, the variables (what are they, what type is each used as, and what are the units), and so on.

   •A Consumer Reports article on energy bars gave the brand name, flavor, price, number of calories, and grams of protein and fat.

   •A report on a charity run in Toronto listed each runner's gender, country, age, and time.

Solution:

Consumer Reports
Who: energy bars
What: brand name, flavor, price, calories, protein, fat
When: not specified
Where: not specified
How: not specified. Are data collected from the label? Are independent tests performed?
Why: information for potential consumers
Categorical variables: brand name, flavor
Quantitative variables: price ($), number of calories (calories), protein (grams), fat(grams)

Charity Run
Who: charity runners
What: gender, country, age, time
When: not specified
Where: Toronto
How: not specified. Presumably, the data were collected from registration information.
Why: race result reporting
Categorical variables: gender, country
Quantitative variables: age (years), time (hours, minutes, seconds)

Resources

ActivStats[*]

• Start with Lesson 1 to let students familiarize themselves with the features of the software. Lesson 2 examines types of data and context.

Web Links

• The Data and Story Library (DASL; http://lib.stat.cmu.edu/DASL/) is a source of data for student projects and classroom examples.

• Statistics Canada has a wealth of information that is free to educational institutions if accessed through the institutional library, or you may want to visit the Statistics Canada website for educators and students at: http://www.statcan.gc.ca/edu/power-pouvoir/ch1/5214857-eng.htm

This website features not only a vast amount of data by different subjects, but also exercises that you can use with your students as well.

Other

• Read polls, studies, or other reports in newspaper and magazine articles. It's always interesting to see how well (or poorly) they provide information about the W's.

If you have a computer and projection capabilities in class, you can find daily surveys at Gallup and other polling organizations. Current data are often particularly interesting to students. But don't use results of voluntary-response online surveys. We'll be making the point that these are fatally flawed—but we can't say that clearly without concepts and terms that we haven't developed yet.

[*] ActivStats is available at www.MyStatLab.com

Basic Exercises

1. The following data show responses to the question "What is your primary source for news?" from a sample of students at the University of Manitoba in June 2015.

Internet	Newspaper	Internet	TV	Internet
Newspaper	TV	Internet	Internet	TV
Newspaper	TV	TV	Newspaper	TV
Internet	Internet	Internet	Internet	Internet
TV	Internet	Internet	TV	TV

   a. Identify the W's for this data
   b. Is the data categorical or quantitative?
   c. Is the data cross sectional or time series?.

2. In September 2015, a cable company surveyed its customers in Burnaby, BC, and asked how likely they were to watch more than 2 hours of TV per day. The following data show the responses.

Very Likely	Unlikely	Unlikely	Very Likely
Likely	Unlikely	Likely	Likely
Unlikely	Unlikely	Likely	Likely
Very Likely	Unlikely	Unlikely	Very Likely
Unlikely	Unlikely	Unlikely	Likely

   a. Identify the W's for this data
   b. Is the data categorical or quantitative?
   c. Is the data cross sectional or time series?.

ANSWERS

1.

a. Who: students
   What: primary source of news
   When: June 2015
   Where: University of Manitoba
   How: not specified.
   Why: not specified

b. Categorical

c. Cross-sectional

2.

a. Who: Cable company customers

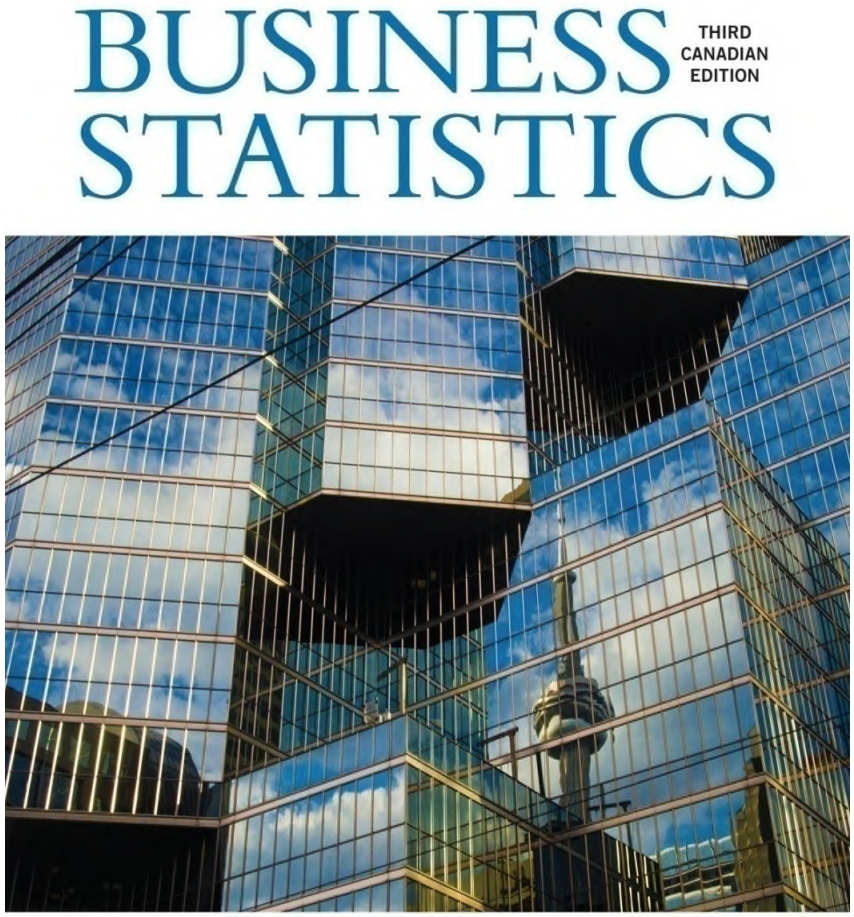
What: likelihood of watching more than 2 hours of TV per day
When: September 2015
Where: Burnaby, BC
How: survey
Why: not specified

b. Categorical

c. Cross-sectional

Business Statistics

Third Canadian Edition



Pearson

Chapter 2

Data

Ch.2: Data

Learning Objectives

1) Identify the context of your data
2) Distinguish different types of data

2.1 What *Are* Data? (1 of 11)

- Data values or *observations* are systematically recorded information, whether numbers or labels, together with its context.

- Data can be numbers, names, or other labels
- Data are useless without an understanding of their *context* (The answers to *Who* and *What* are essentials.)

- Data are often organized into a *data table* like that below

2.1 What *Are* Data? (2 of 11)

Table 2.1 An example of data with no context. It's impossible to say anything about what these values might mean without knowing their context.

10675489	B0000010AA	10.99	Chris G.	905	Quebec	15.98
Samuel P.	Nova Scotia	10783489	12837593	N	B000068ZVQ	15783947
Ontario	Katherine H.	16.99	Alberta	N	11.99	N
B000002BK9	902	Monique D.	Y	819	B0000015Y6	403

2.1 What *Are* Data? (3 of 11)

A row of a data table corresponds to an individual *case* about *Whom* (or about *Which* – if they are not people) *we* record some characteristics.

These characteristics may be collected on or about …

- *respondent* – individual who answers a survey
- *subject* or *participant* – person on whom we experiment
- *experimental units* – a company, website, or other inanimate subject

2.1 What *Are* Data? (4 of 11)

Table 2.2 Example of a data table. The variable names are in the top row. Typically, the Who of the table are found in the leftmost column.

Cases

Purchase Order Number	Name	Ship to Province	Price	Area Code	Gift?	ASIN
10675489	Katherine H.	Alberta	10.99	403	N	B0000015Y6
10783489	Samuel P.	Nova Scotia	16.99	902	Y	B000002BK9
12837593	Chris G.	Quebec	15.98	819	N	B000068ZVQ
15783947	Monique D.	Ontario	11.99	905	N	B000001OAA

2.1 What *Are* Data? (5 of 11)

The *characteristics* recorded about each individual or case are called *variables*.

These are usually shown as the columns of a data table and identify *What* has been measured.

2.1 What *Are* Data? (6 of 11)

Table 2.2 Example of a data table. The variable names are in the top row. Typically, the Who of the table are found in the leftmost column.

Variables

Purchase Order Number	Name	Ship to Province	Price	Area Code	Gift?	ASIN
10675489	Katherine H.	Alberta	10.99	403	N	B0000015Y6
10783489	Samuel P.	Nova Scotia	16.99	902	Y	B000002BK9
12837593	Chris G.	Quebec	15.98	819	N	B000068ZVQ
15783947	Monique D.	Ontario	11.99	905	N	B000001OAA

2.1 What *Are* Data? (7 of 11)

Data tables are cumbersome for complex data sets, so often two or more separate data tables are linked together in a *relational database.*

Each data table included in the database is a *relation* because it is about a specific set of cases with information about each of these cases for all (or at least most) of the variables.

2.1 What *Are* Data? (8 of 11)

Example: A typical relational database is provided consisting of three relations: customer data, item data, and transaction data

For example, we can look up a **customer** to see what items they purchased, or we may look up an **item (product number)** to see who purchased it

2.1 What *Are* Data? (9 of 11)

Table 2.3 A relational database shows all the relevant information for the three separate relations linked by customer and product numbers.

Customers

Customer Number	Name	City	Province	Postal Code	Customer Since	Gold Member
473859	Rahini, R.	Magog	QC	J1X SV8	2007	No
127389	Li, V.	Guelph	ON	N1K 2H9	2000	Yes
335682	Marstas, J.	Calgary	AB	T2E O89	2003	No

2.1 What *Are* Data? (10 of 11)

Table 2.3 A relational database shows all the relevant information for the three separate relations linked by customer and product numbers.

Items

Product ID	Name	Price	Currently in Stock
SC5662	Silver Cane	43.50	Yes
TH2839	Top Hat	29.99	No
RS3883	Red Sequinned Shoes	35.00	Yes
…			

2.1 What *Are* Data? (11 of 11)

Table 2.3 A relational database shows all the relevant information for the three separate relations linked by customer and product numbers.

Transactions

Transaction Number	Date	Customer Number	Product ID	Quantity	Shipping Method	Free Ship?
T23478923	9/15/17	473859	SC5662	1	UPS 2nd Day	N
T23478924	9/15/17	473859	TH2839	1	UPS 2nd Day	N
T63928934	10/22/17	335473	TH2839	3	UPS Ground	N
T72348299	12/22/17	127389	RS3883	1	FedEx Ovnt	Y

2.2 Variable Types (1 of 13)

When a variable names categories and answers questions about how cases fall into those categories, it is called a *categorical variable* (also called *qualitative variable).*

When a variable has measured numerical values with *units* and the variable tells us about the quantity of what is measured, it is called a *quantitative variable.*

2.2 Variable Types (2 of 13)

Categorical variables …

- arise from descriptive responses to questions like “What kind of advertising do you use?” or “Do you invest in stock market?”
- may only have two possible values (like “Yes” or “No”)
- may be a number like a telephone area code

2.2 Variable Types (3 of 13)

Table 2.4 Some examples of categorical variables.

Question	Categories or Responses
Do you invest in the stock market?	__Yes__No
What kind of advertising do you use?	__Magazines__Internet__Direct Mailings
I would recommend this course to another student.	__Strongly Disagree__Slightly Disagree__Slightly Agree__Strongly Agree
How satisfied are you with this product?	__Very Unsatisfied__Unsatisfied__Satisfied__Very Satisfied

2.2 Variable Types (4 of 13)

Some quantitative variables have *units*. The units indicate …

- how each value has been measured
- the corresponding *scale* of measurement
- how much of something we have
- how far apart two values are

Other quantitative variables have no *units,* such as …

- Number of visits to a web site
- Number of shares of a company traded in Toronto Stock Exchange

2.2 Variable Types (5 of 13)

Some variables can be both categorical and quantitative

How data are classified depends on *Why* we are collecting the data

For example, variable *Age* is obviously the quantitative value, measured in years, that may be used for finding the average age of customers

Age categories such as Child, Teen, Adult, or Senior can be the categorical value used to decide in which music to offer in a special deal – folk, jazz, hip hop or reggae.

2.2 Variable Types (6 of 13)

Counts

In gathering data, we often count things.

Counts can be used in two different ways:

1) To summarize categorical variables.
2) To measure the amount of things.

2.2 Variable Types (7 of 13)

Counts

A company may count the number of purchases shipped by different means to summarize the categorical variable *Shipping Method*.

Table 2.5 A summary of the categorical variable *Shipping Method* that shows the counts, or number of cases, for each category.

Shipping Method	Number of Purchases
Ground	20,345
Second-day air	7,890
Overnight	5,432

2.2 Variable Types (8 of 13)

Identifiers

An *identifier variable* is a unique identifier assigned to each individual or item in a group.

For example, social insurance numbers, student ID numbers, tracking numbers, transactions numbers, are all identifier variables for people or items.

2.2 Variable Types (9 of 13)

Identifiers

Identifier variables …

- are special kind of categorical variable
- do not have units
- are useful in combining data from different sources to avoid duplication
- are not variables to be analyzed

2.2 Variable Types (10 of 13)

Other Data Types

Categorical variables used *only* to name categories are sometimes called *nominal variables*

When data values can be ordered, we say that the variable has *ordinal* values. For example, employees can be ranked according to the number of days worked in the company.

2.2 Variable Types (11 of 13)

Other Data Types

We may differentiate quantitative variables according to whether their measured values have a defined value for zero.

For example, 80^{o}F is not twice as hot as 40^{o}F because 0^{o} is an arbitrary value.

Data without a defined value for zero are said to be on an *interval scale*, otherwise, they are said to be on a *ratio scale.*

2.2 Variable Types (12 of 13)

Cross-Sectional and Time Series Data

Variables that are measured at regular intervals over time are called *time series*. For example, the share price of the Royal Bank of Canada at the end of each day for the past year.

When several variables are all measured at the same point in time, the data is called *cross-sectional data*. For example, collecting data on sales revenue, number of customers, and expenses totalled over the past month at each Starbucks location.

2.2 Variable Types (13 of 13)

Primary and Secondary Data

Primary data are data we collect ourselves and that the researchers have a very clear idea of the meaning of the data they collect from surveys, since they themselves design the wording of every question in those surveys and conduct the interviews.

Secondary data are data collected by another party, like Statistics Canada. It is very important to read all the guidelines and footnotes provided in order to get a precise idea of what the secondary data mean.

2.3 Where, How, and When (1 of 3)

When and *Where* data are collected can be important.

- Values recorded in 1803 may mean something different from similar values recorded last year.
- Values measured in Tanzania may differ in meaning from similar measurements made in Mexico.

2.3 Where, How, and When (2 of 3)

How data are collected can make the difference between insight and nonsense

For example, data that come from a voluntary survey on the Internet are almost always worthless.

However, data provided by agencies and businesses via the Internet can be extremely useful.

2.3 Where, How, and When (3 of 3)

Data can be found …

- by performing an experiment and actively manipulating variables
- by purchasing from the public or private agencies
- by searching on the Internet

What Can Go Wrong?

- Don’t label a variable as categorical or quantitative without thinking about the data and what they represent

- Don’t automatically assume variables that are numbers are quantitative

- Always be skeptical, understand the context of the data and look for bias in how the data were collected.

What Have We Learned? (1 of 2)

- We've learned that data are information gathered in a specific context
- The five W's (*Who*, *What*, *Why*, *Where*, and *When*) help nail down the context of the data
- We must know at least the *Who*, *What*, and *Why* to be able to say anything useful based on the data
  - The *Who* are the cases
  - The *What* are the variables
  - The *Why* help us decide which way to treat the data

What Have We Learned? (2 of 2)

- We've learned that data can be divided into quantitative/categorical, cross-sectional/time-series, and primary/secondary
- Categorical variables identify a category for each case
- Quantitative variables record measurements or amounts of something and include units
- Variables can be categorical or quantitative depending on what we want to learn from them
- Identifier variables are categorical variables that name each case