CHAPTER 5

Organization of Data for Analysis

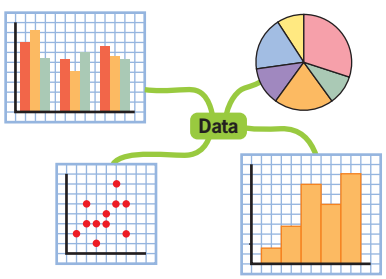
In the 21st century, farming has evolved to use data in order to maximize crop yield. For example, machines can sort green tomatoes from red ones at a rate of more than 3000 tomatoes per minute. Fertilizer spreaders use GPS and soil sample data to customize how much fertilizer is applied to specific parts of a field.

Key Terms

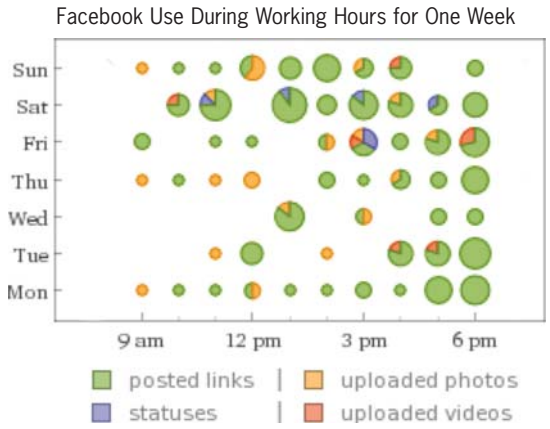
numerical (quantitative) data
categorical (qualitative) data
ordinal data
nominal data
population
sample
variability (in samples)
treatment group
control group
bias
primary source data
microdata
secondary source data
aggregate data
response bias
sampling bias
measurement bias
non-response bias

Literacy Strategy

You can use a mind map to organize the different ways that data can be displayed and situations where one method is better than another. As you complete the chapter, you will learn about different types of data. Create a mind map to show how to display different types of data graphically. Include information about when to use each type of graph.



Career Link



Data Miner

A data miner takes a large set of data and "mines" it for patterns, trends, and relationships. Often data miners do not collect the data themselves but rather "dig in" to large sets of data that others have created. For example, social media websites such as Facebook use data mining to customize advertising for each user. Although computers do the actual "mining" of a user's Facebook page, data miners and others develop the algorithms that make it work. How do you think the ads on your Facebook page might differ from the ones on your friends' or parents' pages?

Chapter Problem

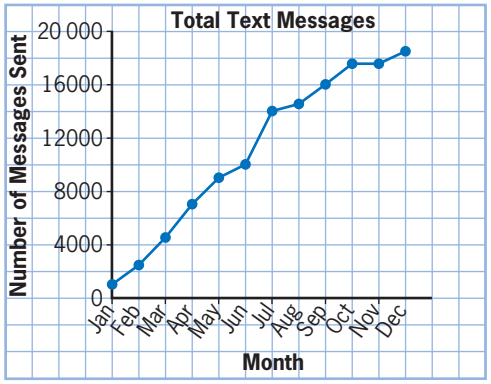
Food Production

It is estimated that by 2050 there will be over 9 billion people living on the planet. Some estimates suggest that by then we will need 30–70% more food and 40% more water. Imagine you have been hired by Agri-Research Consultants, a firm that helps farmers maximize the yield of various crops. List some ways data could help farmers increase their yields.

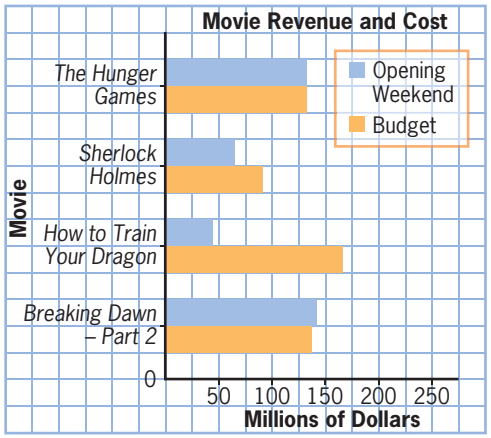
Identifying and Analysing Graphs

1. Name each type of graph.

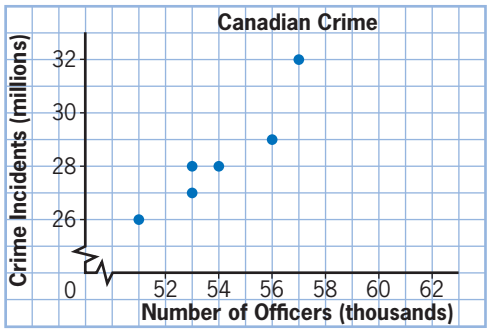
a)



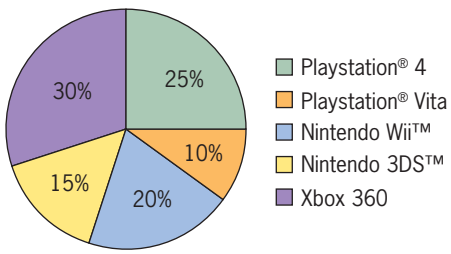
b)



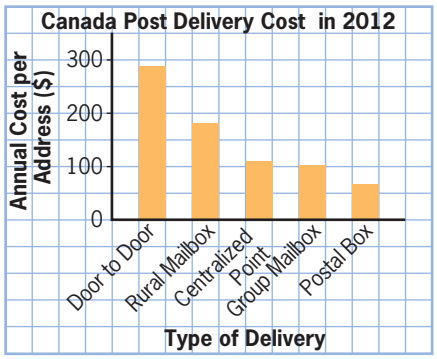
c)



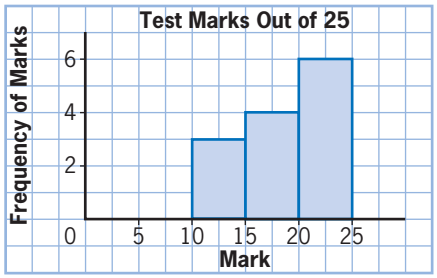
d) Favourite Game Console



e)



f)



2. Use the graphs in #1 to answer the questions.

a) What intervals does the histogram show?

b) Approximately how much does it cost to deliver mail to a centralized point?

c) What trend seems to exist between number of officers and crime incidents?

d) What is the most popular game console?

e) During which month do you think students had their cell phones taken away?

f) Which movie made the most money on its opening weekend?

3. What is the difference between a bar graph and a histogram?

Use a Spreadsheet to Create Graphs

4. Create a bar graph to represent the data in the table. Include a title and label the axes.

TV Show	Viewers (millions)
America's Got Talent	9.8
Big Bang Theory	18.6
The X Factor	4.9
Elementary	9.1

5. Create a circle graph to represent the data in the table.

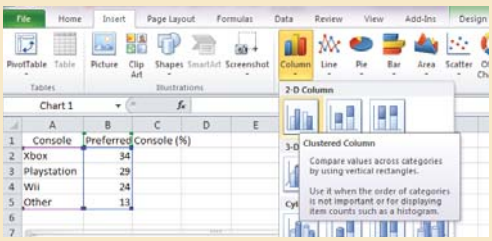
Favourite Food	Amount of People
Spaghetti	20%
Pizza	40%
Hamburgers	15%
Subs	25%

Example: Create a bar graph and a circle graph to represent the data in the table.

Console	Preferred Console (%)
Xbox	34
PlayStation	29
Wii	24
Other	13

Enter the data in a table in a spreadsheet.

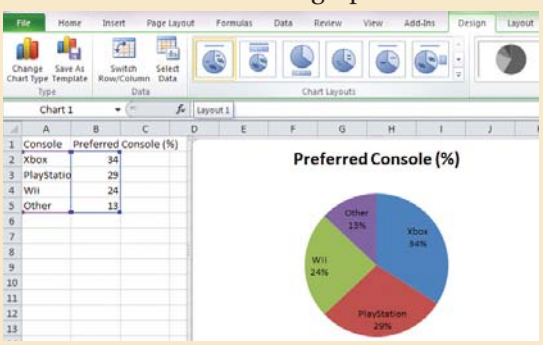
To create a bar graph, highlight the data. Then select **2-D Clustered Column** from the **Insert** menu.



Under **Chart Tools**, click **Layout**. Label the axes by clicking on **Axis Titles**.

To create a circle graph, highlight the data. Then select **2-D Pie Chart** from the **Insert** menu.

Under **Chart Tools**, choose the **Design** tab to show the labels on the graph.



6. Use the same steps as above to create a scatter plot. Instead of selecting **Column** or **Pie** on the **Insert** menu, choose **Scatter**. Then, select **Scatter with only Markers**.

Year	2007 Toyota Camry Value ($)	2007 Ford Fusion Value ($)
2010	15 575	13 997
2011	14 230	12 924
2012	13 107	11 308
2013	12 224	9 831

7. The data describe the change in population in New York and Toronto since 2001. Create a cluster bar graph to display the data.

City	Population in 2001 (millions)	Population in 2012 (millions)
Toronto	2.25	2.73
New York	7.86	8.34

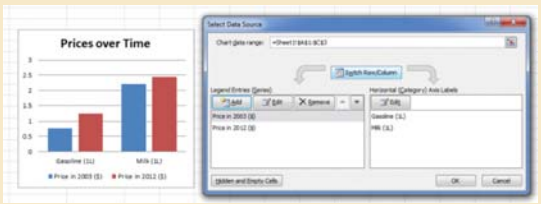
Example: Create a cluster bar graph to compare the changes in price.

	2003 Price ($)	2012 Price ($)
Gasoline (1 L)	0.76	1.25
Milk (1 L)	2.20	2.45

Enter the data in a table in a spreadsheet.

Highlight the data. Then select **Clustered Column** from the **Insert** menu.

Reformat the chart to compare the price change in gas and milk independently. Right click on the chart and click on **Select Data**. Click on **Switch Row/Column**. The chart will now directly compare the price of each item.



5.1

Data Concepts and Graphical Summaries

Learning Goals

I am learning to

- show how data are used and misused in statistical studies
- identify different types of data
- understand that there is variability in data
- see that you can analyse single sources of data or related sources

Minds On...

Data are everywhere. What story does the data in the image tell?

Action!

Investigate Interpreting Data

Materials

- calculator
- grid paper and ruler or graphing software
- computer with Internet access

On its maiden voyage in 1912, the RMS *Titanic* struck an iceberg and sank in less than three hours. The table shows survival statistics.

Category	Survivors		Deceased		Total
	Women and Children	Men	Women and Children	Men	
First Class	145	54	11	119	329
Second Class	104	15	24	142	285
Third Class	105	69	119	417	710
Crew	214		685		899
Total	706		1517		2223

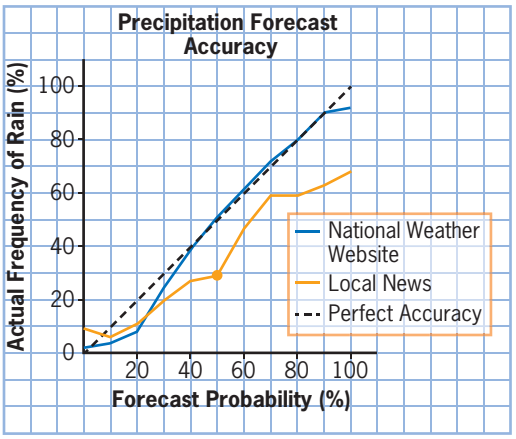
Source: Statistics of Passengers Rescued and Lost, White Star Momentos

1. Create a graph that lawyers could use to help families of the affected passengers. Explain what type of graph you chose and why.
2. The cruise line released a statement with the headline, "Hundreds of Lives Saved in *Titanic* Mishap." Create a graph or highlight some data that the cruise line could use to justify this headline.
3. **Reflect** How do the graphs from steps 1 and 2 show that the way in which you represent information can change people's perceptions?
4. **Extend Your Understanding** Search the Internet for some examples of data displays that are accurate but misrepresent the information.

Example 1

Variability in Data

Sometimes people can use accurate information to tell different stories. Probability of precipitation (PoP) is a common measurement used in weather forecasts. The graph shows the accuracy of a national weather website and local news channels compared to perfect accuracy.



Source: Data from *The Signal and the Noise: Why So Many Predictions Fail—But Some Don't*

a) What does the yellow dot indicate?

b) Which outlet is more accurate, the national weather website or the local news?

c) If both outlets base their forecasts on information collected by Environment Canada, what reasons can you suggest for their differences?

Solution

a) This data point shows that when the local news predicted a 50% chance of precipitation, it actually rained only about 30% of the time.

b) For most of the predictions, the national weather website was closer to perfect than the local news. The local news consistently overestimated the chance of rain, in some cases by as much as 30 percentage points. The national weather website's predictions were more accurate than the local news except where their predictions were between 0% and just over 20% PoP.

c) Environment Canada collects weather data at many locations in small geographic areas. The national weather website may be forecasting for a larger area than the local news. If precipitation is recorded in a localized region, it may be outside the area forecast by the local news, but included in the forecast for the national weather website. Also, although the raw data come from the same source, the calculations used to make the predictions may be different. This results in different analyses of the same data.

Your Turn

Researchers often have conflicting opinions even though they use the same data. Research one of the following topics to find conflicting opinions:

- climate change
- vaccinations
- fluoridated water

numerical (quantitative) data

- data in the form of any number

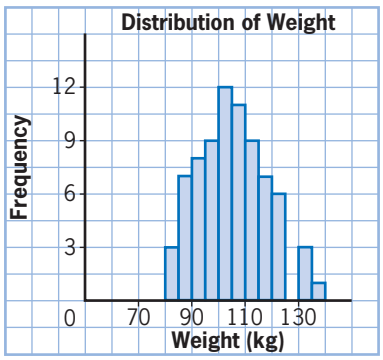
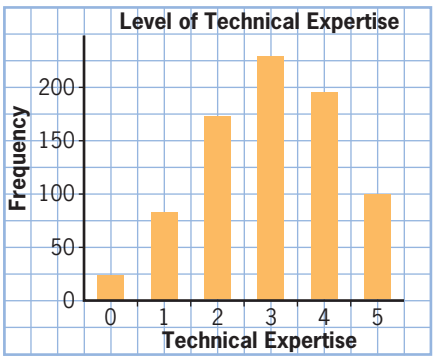
categorical (qualitative) data

- data that can be sorted into distinct groups or categories

There are two main types of data: **numerical (quantitative)** and **categorical (qualitative)**. Numerical data are either continuous or discrete.

Continuous data can have any value in a range (including decimal numbers). For example, the weight of a person or the amount of time an experiment takes could have any value in a range. We often use a histogram to display continuous data. When the bars in a histogram are touching, it means that the data can be any value in a range.

Discrete data are data that only have specific values (usually whole numbers). We often represent discrete data with a bar graph. The bars do not touch, indicating that there are no possible values in between.

Continuous Numerical Data	Discrete Numerical Data
	
Since the weight of a person can be any value, weight is continuous data.	When someone fills out a survey and uses a rating scale, the scale is measured in whole numbers, so it is discrete data.

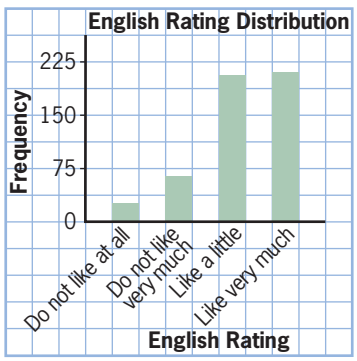
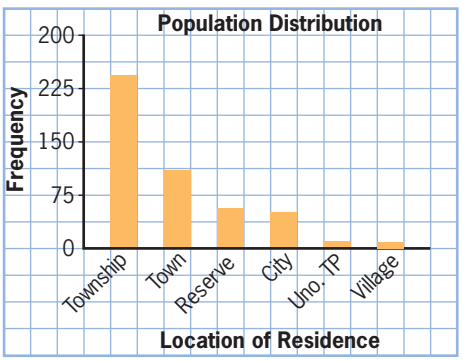
ordinal data

- qualitative data that can be ranked
- examples: poor, fair, good, very good

nominal data

- qualitative data that cannot be ranked
- examples: blue eyes, green eyes, brown eyes

There are two main types of categorical data: **ordinal data** and **nominal data**.

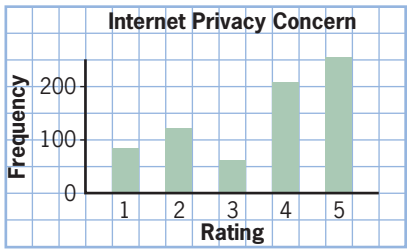
Categorical Ordinal Data	Categorical Nominal Data
	
Since the person is answering using a rating scale, these are ordinal data.	The data are categorized by the type of place. There is no logical order, so these nominal data are placed from the highest bar to the lowest.

Example 2

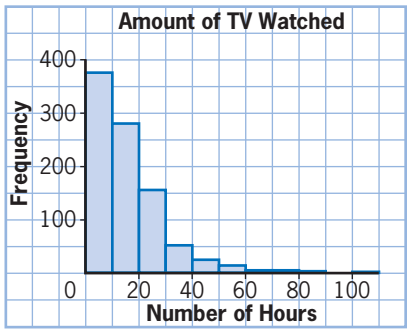
Comparing Types of Data

For each graph, identify the type of data, give reason(s) for your choice, and write one statement about what the data show.

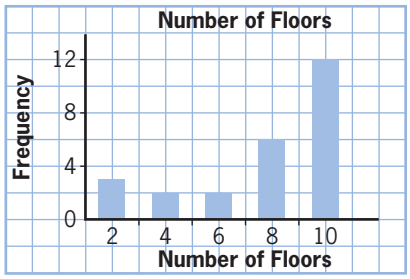
a) A survey asks people to rate how concerned they are with Internet privacy on a scale where 1 is not concerned and 5 is very concerned.



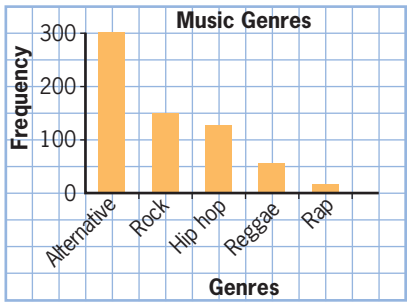
b) A survey asks 1000 people how many hours a week they watch TV.



c) A town planner records how many floors each apartment building has.



d) The songs in a digital music library are sorted by genre.



Solution

a)
- categorical, ordinal data since they are non-numerical and ranked
- most people are concerned with Internet privacy

b)
- numerical, continuous data since the numbers can be any value
- more than one-third of the respondents watch 0–10 hours of TV

c)
- numerical, discrete data since each value can be only a distinct whole number
- most apartment buildings in the town have 10 floors

d)
- categorical, nominal data since they are non-numerical and not ranked
- the owner of the library likes alternative music more than any other

Project Prep

Being able to identify types of data will be useful in organizing the information for your project.

Your Turn

Your teacher will provide you with a file called **YouthSurvey.ftm**. Create one graph for every type of data. Are there any types of data that could not be represented as a graph? Explain.

Example 3

Data With More Than One Variable

Literacy Link

Per capita means per person.

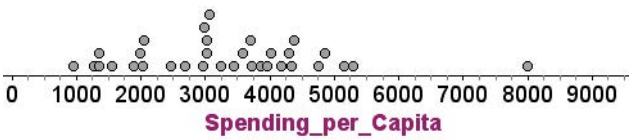
Researchers often measure more than one variable of a particular item. Then, they can analyse the data to see if the measurements are connected to each other. The table shows the life expectancy in years and the health care spending per capita in dollars from various countries. Your teacher will provide you with a Fathom™ file called **HealthCare.ftm** to use if you wish.

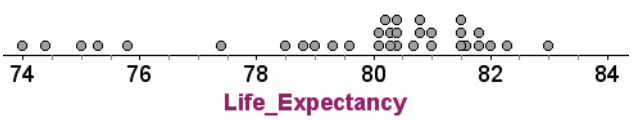
Country	Spending per Capita ($)	Life Expectancy (years)	Country	Spending per Capita ($)	Life Expectancy (years)
Australia	3734	81.6	Japan	3025	83
Austria	4345	80.4	Korea	1895	80.4
Belgium	3874	80.1	Luxembourg	4755	80.7
Canada	4309	81	Mexico	957	74
Chile	1283	78.8	Netherlands	4870	80.8
Czech Republic	2039	77.4	New Zealand	2984	80.8
Denmark	4390	79	Norway	5300	81
Estonia	1371	75	Poland	1356	75.8
Finland	3259	80.1	Portugal	2692	79.6
France	3962	81.5	Slovak Republic	2063	75.3
Germany	4187	80.3	Slovenia	2470	79.3
Greece	2977	80.3	Spain	3080	81.8
Hungary	1567	74.4	Sweden	3703	81.5
Iceland	3597	81.8	Switzerland	5157	82.3
Ireland	4037	80.2	United Kingdom	3456	80.4
Israel	1991	81.5	United States	8006	78.5
Italy	3030	82			

Source: Table 2: Total expenditure on health per capita, OECDiLibrary, October 11, 2013 and Table 11: Life expectancy at birth, total population, OECDiLibrary, December 6, 2013

a) Which country spends the most per person on health care? the least? Where is Canada on the list?

b) The dot plots show the spending per capita and the life expectancy of each country. In each plot, each dot represents one country.





For each graph, identify the lowest and highest ranked country, as well as Canada and the United States. What story does the data seem to tell?

c) Create a scatter plot of life expectancy versus health care spending. Which countries lie outside the general trend? How do you know?

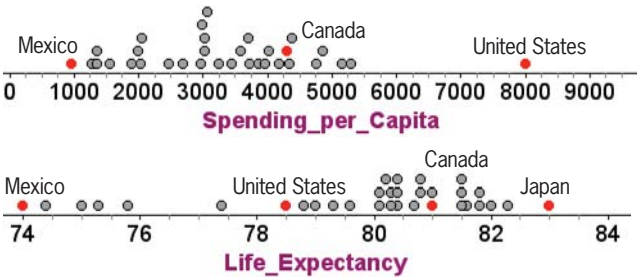
d) In what area of the graph would a country want to be located? Why?

e) In Canada, everyone has access to health care services. This is not the case in the United States. How can these data be used to argue against more health care spending? How can they be used to argue in favour of more health care spending?

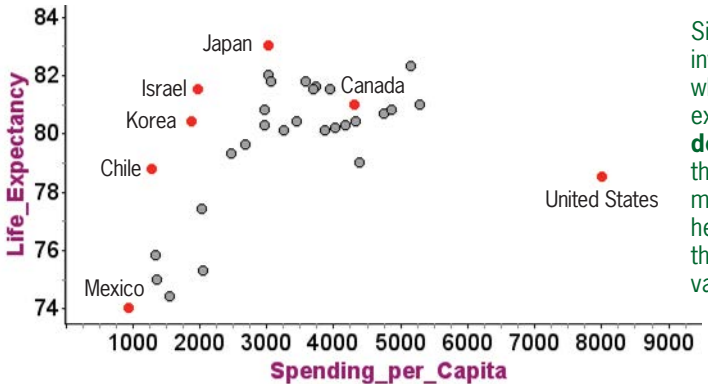
Solution

a) The United States spends the most per person on health care ($8006). Mexico spends the least ($957). Canada is in the middle, at $4309.

b) The United States has the highest health care spending, but is in the middle in terms of life expectancy. Mexico has the lowest life expectancy and the lowest spending per person. Japan has relatively low health care spending yet it has the highest life expectancy.



c) You can use technology to create the scatter plot. Your teacher will provide you with a Fathom™ file called **HealthCare.ftm** or a CSV file called **HealthCare.csv**. Refer to the Prerequisite Skills section on page 195 for help creating a scatter plot using technology.



Since you are investigating whether life expectancy **depends on** the amount of money spent on health care, it is the dependent variable.

The country clearly outside of the trend is the United States, since it spends almost twice as much as every other country. Chile, Korea, Israel, and Japan could also be considered outside the trend since they all seem to have a higher life expectancy than many of those in the trend that spend similar amounts.

d) The top left area of this graph indicates that countries have a high life expectancy with low spending per capita.

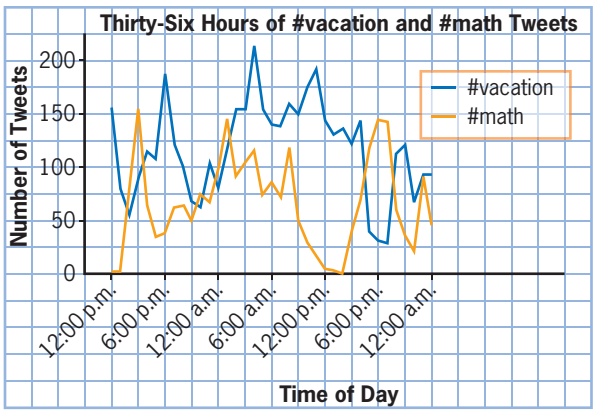
Literacy Link

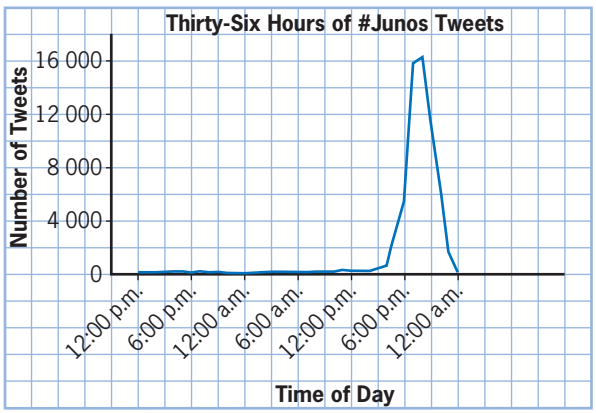
CSV stands for *comma-separated values*. Data in a list are separated by commas, which then instructs the spreadsheet that the data entry belongs in the next cell of the column.

e) Someone who is in favour of spending less on health care could argue that Americans are already paying almost twice as much per person for health care as many countries, yet their life expectancy is in the middle of the range. It would be hard for someone in favour of more health care spending to use the data. However, these data show that other factors must affect life expectancy since spending in the United States is so high and life expectancy is not affected. Perhaps the issue is less about the amount of money being spent and instead is about how that money is being spent. Or, perhaps it is about lifestyle choices, diet, exercise, and other environmental factors.

Your Turn

The graphs show the number of tweets at various times of day that contain the hashtags #vacation, #math, and #Junos.





a) What do these graphs tell you about the frequency of tweets with each type of hashtag?

b) What do the data seem to indicate about #math and #vacation as Twitter topics?

c) The Juno Awards is the Canadian music industry's yearly award show. When do you think the show was broadcast?

d) Could these graphs be compared to each other as they are? Explain why or why not.

Consolidate and Debrief

Key Concepts

- Depending on who is analysing the data and their intention, the information taken from the data can be very different.
- Variability in data exists due to errors in measurement or varying conditions in experiments.
- Different people can interpret data in different ways.
- There are two main types of data: numerical and categorical. Numerical data may be classified as continuous or discrete. Categorical data may be classified as ordinal or nominal.
- When researchers collect data on more than one variable, they can compare the data to see if there is a relationship.

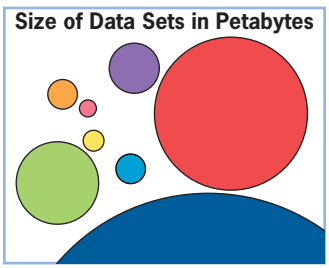
Reflect

R1. Studies show that political experts' predictions are correct or mostly correct 46% of the time and incorrect or mostly incorrect 47% of the time. The remaining 7% are a mix of correct and incorrect predictions. What does this suggest about how reliable experts might be when making predictions?

R2. In the last 20 years, the amount of data being moved and stored online has become staggeringly large. In 2012 alone, almost 3 zetabytes of information moved online. Discuss the positives and negatives of collecting this large amount of data each year.

Literacy Link

A *terabyte* is 10^{12} bytes. A *petabyte* is 10^{15} bytes. A *zetabyte* is 10^{21} bytes.



Practise

1. Use examples to describe similarities and/or differences between the two types of data.

a) continuous versus discrete

b) ordinal versus nominal

c) numerical versus categorical

2. List the different types of data.

a)

	A	B	C	D	E	F
1	Province	Height	Armspan	Handedness	Eye Colour	Fav Sport
2	Ontario	156.5	122	Right	Green	Dancing
3	Ontario	156	139	Right	Blue	Swimming
4	Ontario	163.5	163	Right	Brown	Hockey
5	.	.	.	.	.	.
6	.	.	.	.	.	.

b)

	A	B	C	D	E	F
1	Province	Number of Languages Spoken	Reaction Time	Mode of Travel	Travel Time	Number of Siblings at Home
2	Ontario	1	0.33	Bus	15	2
3	Ontario	2	0.485	Walk	10	3
4	Ontario	1	0.28	Car	25	1
5	.	.	.	.	.	.

3. In 2012, the Council of Ontario Universities published a report showing the average annual salary for different university majors two years after graduation. The table shows the 10 lowest paid majors. Create a graph that unfairly shows how poorly someone with a major in Fine and Applied Arts will be paid.

Rank	Major	Salary
1	Fine & Applied Arts	$35 539
2	Humanities	$38 696
3	Theology & Religious Vocations	$39 333
4	Kinesiology/Recreation/ Phys-Ed	$39 779
5	Architecture & Landscape Architecture	$40 733
6	Social Sciences	$42 585
7	Journalism	$42 901
8	Agricultural & Biological Services	$43 466
9	Forestry	$43 889
10	Food Science & Nutrition	$43 952

Source: 2012 Grad Survey, Council of Ontario Universities

Apply

4. A teacher creates a new blog about mathematics and statistics. She collects data showing the number of pageviews in the first few months.

Month	Pageviews
Jan	120
Feb	315
Mar	434
Apr	502
May	596
Jun	537
Jul	472
Aug	645
Sept	848

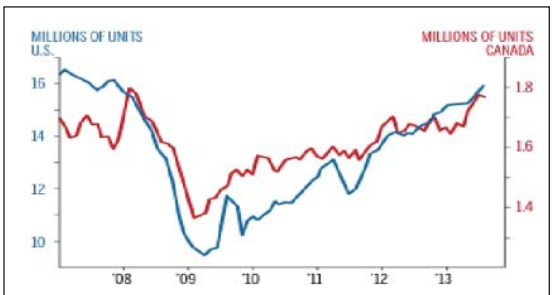
a) Create a graph of the data.

b) What is the general trend for the traffic on this blog?

c) The graph shows a dip in pageviews. Suggest reasons why this might happen.

d) Predict the number of pageviews expected in December.

5. Consider the headline and graph from *Maclean's*:

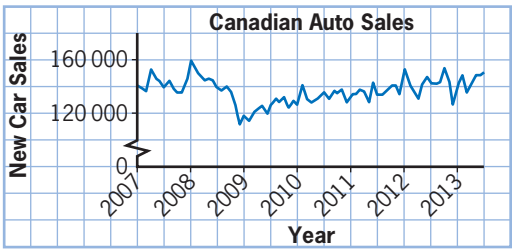


Back on track
Five years after the financial crisis started and nearly destroyed the North American auto industry, car sales have finally recovered.

Source: "Chart of the Week," *Maclean's*, September 30 2013, p. 36

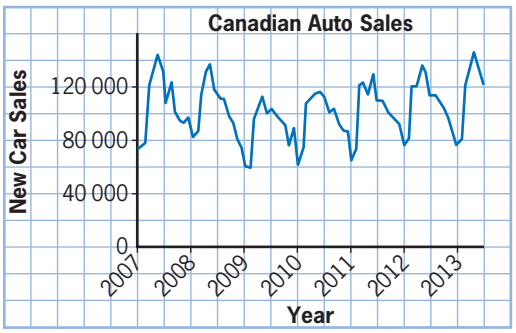
a) The graph uses real data. How have the data been presented to increase impact of the statement?

b) This graph shows the same data. What is it about the original statement that this graph does not show as well?



Source: CANSIM Table 079-0003, Statistics Canada, April 11, 2014

6. Thinking The graph in #5 shows the seasonal adjusted data. This graph shows the actual number of sales.



Source: CANSIM Table 079-0003, Statistics Canada, April 11, 2014

a) Why do you think *Maclean's* chose to show seasonally adjusted data?

b) Why do you think auto sales are compared to the same month in the previous year rather than to the previous month?

Literacy Link

When data are *seasonally adjusted*, they are changed so that fluctuations due to seasonal factors are removed. For example, some farm workers will be unemployed during the winter months, so unemployment rates will be higher in the winter.

Achievement Check

7. The table shows some statistics for superhero searches on YouTube.

Superhero	Views	Hours Watched
Batman	3 000 000 000	71 000
Thor	2 100 000 000	66 000
Superman	1 700 000 000	14 000
Iron Man	1 400 000 000	20 000
The Avengers	1 000 000 000	31 000
Wolverine	540 000 000	7 800
Spider-Man	340 000 000	7 400
Captain America	280 000 000	4 900
Justice League	220 000 000	3 200
Deadpool	200 000 000	8 900

a) Create a graph that tells a story about the data. Your teacher will provide you with a file called **Superheros.csv**.

b) How can the large numbers be expressed so that they have more meaning?

c) Your teacher will direct you to a website that creates graphs based on search criteria. Enter three superheroes as separate search items. Describe the graphs.

Extend

8. Your teacher will direct you to a website that shows the popularity of various names.

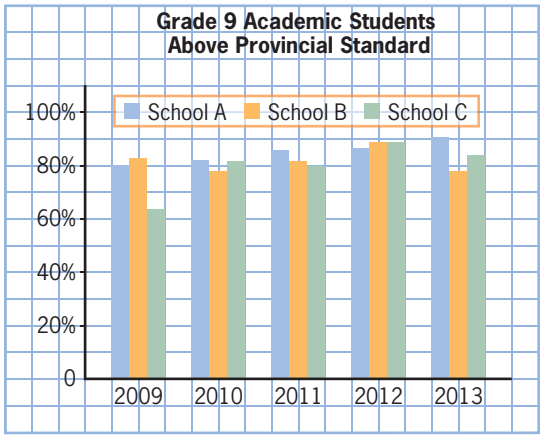
a) Search for your name and describe how its popularity has changed over the years.

b) Enter a common name at your school. Is it just as popular on this site? Explain.

c) Determine three names that are currently not popular but were at one time.

d) Determine three names that were not popular until the last decade.

9. Communication The graph shows the results of EQAO tests at three schools.

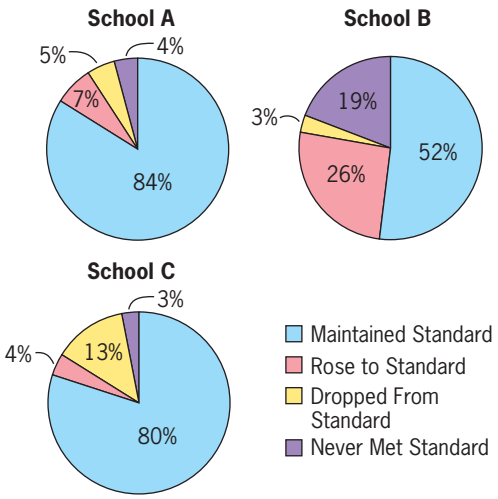


a) Describe the schools' results over time. Which school is doing the best?

b) Some people prefer to compare the results of the same cohort of students. The circle graphs show the cohort data for 2013 from the three schools.

Literacy Link

A *cohort* is a group of people.



Describe some positive and negative results for each school. Based on this data, which school appears to have the best results?

5.2

Principles of Data Collection

Learning Goals

I am learning to

- distinguish between a population and a sample
- understand why sampling a population can give information about that population
- understand that when sampling data the results can vary
- sample data in various ways

Minds On...

What if your job were to test the quality of french fries? At a french fry factory, quality control experts measure and assess batches of fries. Why do you think this is done?

Action!

Investigate Political Party Support

population

- all the individuals in a group that is being studied

sample

- a group of items or people selected from the population

variability (in samples)

- shows how samples are different from each other
- the more similar the samples are to each other, the lower the variability and the more accurately the samples represent the population

In the months leading up to an election, media outlets often predict which party will get the most votes from the population. They use the results published by polling firms, which contact a sample of the registered voters to try to predict who will win.

1. Work in groups or as a class to design a survey to determine which political party people would vote for.
2. Have everyone in the class answer the survey anonymously.
3. In small groups, choose several other groups of people to sample. What are some factors you might want to consider to get a sample that represents the voting population?
4. Survey the rest of your groups.
5. Compare the results of your surveys between the sample groups. Are the responses similar or is there variability in the samples?
6. **Reflect**
   a) Is this is an accurate method to measure the support for political parties?
   b) When polls are conducted, the number of people polled is often a very small proportion of the voting public. How will this affect the accuracy of the estimates?

c) What if only one sample were done? How would this affect the accuracy?

d) What if the sample were much smaller? How would this affect the accuracy?

7. Extend Your Understanding Research the results of a recent political poll for Ontario. Your teacher will provide you with a list of websites. Are the results of the poll similar to the results from your investigation? Why or why not?

For a sample to be representative of a population, each member of the population must have an equally likely chance of being selected in the sample. So, selection for the sample must be random. If the sample is not random, it is biased and not as reliable. There are different types of sampling methods. Some are better than others.

Type of Sample	Example
Simple Random • randomly choose a specific number of people • examples: stratified samples and systematic samples	Put all the names in a population into a hat and draw one or several names. Each person has an equal chance of being chosen.
Systematic • put the population in an ordered list and choose people at regular intervals	Order all the patients of a doctor in some way (e.g., alphabetically) and choose one randomly. Select the rest of the data at regular intervals from the original starting point (e.g., every tenth name after the original).
Stratified • divide the sample into groups with the same proportions as those groups in the population • time- and cost-efficient to conduct	Survey factory employees about new safety initiatives. There are 1000 employees in the factory, of which 633 are women and 367 are men. Randomly select 63 women and 37 men to take the survey.
Cluster • divide the population into groups, randomly choose a number of the groups, and sample each member of the chosen groups	Survey Little League Canada baseball players. Randomly select five districts in each province and give the survey to every player in those districts.
Multistage • divide the population into a hierarchy and choose a random sample at each level	Conduct an employee wellness survey by randomly selecting 10 stores. Randomly select three departments in each store, and randomly select 10 employees in each of those departments.
Convenience • choose individuals from the population who are easy to access • can yield unreliable results since it inadvertently omits large portions of the population • often very inexpensive to conduct	To get the public's input on a new pet by-law, a local politician goes to a local park and asks people their opinion.
Voluntary • allow participants to choose whether or not to participate • often the only people who respond are either heavily in favour or heavily against what the survey is about	Conduct an online poll asking people whether banning junk food in schools will fight obesity.

Project Prep

If you choose to collect data for your project, you will have to discuss the validity of your sampling method.

Example

Types of Samples

For each situation, identify the type of sample and discuss whether each member of the population is equally likely to be chosen.

a) A teacher wishes to get feedback from the class about a recent presentation. He plans to draw five students' names out of a hat. All 30 students' names will be in the hat.

b) A telephone company wants to determine whether a fitness centre would be used by its 3000 employees. The company plans to survey 300 employees by interviewing every tenth person on the payroll list.

c) A chain store is trying to decide whether to open a store in Huntsville, Ontario. The company decides to survey 25% of the population of Huntsville and three nearby towns. The table shows the population of each location.

Location	Population
Huntsville	19 056
Kearney	841
Emsdale	2 317
McMurrich	779

d) A market research company mails surveys to all of the adult residents in a town. The survey asks about brands of consumer products. The residents are asked to mail back their responses in a prepaid envelope.

e) A reporter stops people on a downtown street to ask what they think of the city's waterfront.

f) Researchers want to investigate the use of pesticides by apple farmers in Ontario. They divide the province into 10 sections and choose five sections at random. They sample all farms within the five sections.

g) The province wants to randomly choose 250 students. It randomly selects five school boards from the 72 in Ontario. Then it randomly selects five schools in each of those boards. Finally, it randomly chooses 50 students in each of those schools.

Solution

a) This is a simple random sample. The draw includes all possible students, so each person has an equally likely chance of being selected.

b) Since the employees are ordered, and every tenth person is chosen, this is a systematic sample. As long as the first person is chosen randomly, everyone has an equal chance of being selected.

c) If 25% of each town is sampled, this is a stratified sample since the number from each town in the sample would be in the same proportion as in the population (the company would survey 4764 people from Huntsville, 210 from Kearney, 579 from Emsdale, and 195 from McMurrich). Everyone has an equal opportunity to be selected since the 25% of people being sampled are randomly selected from each town.

d) Since the residents choose whether or not they respond to the survey, it is a voluntary response survey. Even though all residents received the survey, the data will reflect responses only from those who are interested. Usually only people who feel strongly one way or another will respond. Thus, the entire population will not be represented in the sample.

e) Since the reporter surveys anyone who is walking on the street, the data are collected in a way that is easily accessible. So, this is a convenience sample. The entire population is not represented since only people in that area at that time are surveyed.

f) This is a cluster sample because the population is divided into groups, a number of sections are randomly chosen, and all farms in those sections are surveyed. Since there is random selection in the first set of groups and sections, all groups and sections are equally likely to be chosen. It is more efficient for the researchers to visit the farms within five areas than it would be to travel to different farms throughout the whole province.

g) This population has a hierarchy of organization and random sampling occurs at each level of the organization, so this is multistage sampling. Each student is equally likely to be chosen since there is random sampling at every level.

Your Turn

In each case, identify the type of sample.

a) You want to find out if your town is in favour of starting a composting pickup service. You ask everyone on your street.

b) A university is polling its students. It selects 200 students at random in the same proportions as the enrollment in each department.

c) There are 139 swim clubs in Ontario. Swim Ontario conducts a survey to vote on its new logo. The organization randomly selects 10 swim clubs and surveys every member in each of those clubs.

d) A coach puts the names of all the basketball players into a hat and draws one name for a free basketball.

e) A questionnaire is sent to every ninth person on an alphabetical list of a store's credit card customers. The first person chosen from the list is picked randomly.

f) The student council invites all students to provide ideas for activities.

g) A marketing firm wants to collect information on certain products in a city of 800 000 people. The researchers randomly select 10 neighbourhoods. In each neighbourhood they randomly select five streets, and on each street they randomly select 10 households.

Consolidate and Debrief

Key Concepts

- A population is the entire group of a set of people or things. A sample is a smaller portion of that population.
- You can learn a lot about a population by examining samples of that population, as long as all members of the population are equally likely to be part of the sample.
- When multiple samples are taken from the same population, they are different from each other. This is called variability of samples. The smaller the differences in the samples, the more likely the sample closely represents the population.
- There are many types of sampling techniques. Some types of samples work better in certain situations. A good sample is random, and each person in the population has an equally likely chance to be chosen.

Reflect

R1. What are the differences and similarities between a cluster sample and a multistage sample?

R2. If a population is either heavily in favour or heavily against a certain topic, the sample size for a survey does not need to be as large as it would if the opinions were mixed. Why might this be?

Practise

1. Describe the sampling method used.

a) A seat belt factory randomly selects a time each hour and then tests the next 10 seat belts on the factory line.

b) A city randomly selects 500 residential addresses from its database.

c) A charity mails a survey to its 450 members.

d) The manager of a golf course knows that about 40% of the members are female. He randomly selects 75 females and 112 males to survey.

2. What is the difference between a population and a sample? Use examples to explain.

3. Write several tweets (140 characters or less) to describe each type of sampling method.

Apply

4. Use the data to describe three sampling methods you could use to conduct a survey.

Grade 9		Grade 10		Grade 11		Grade 12	
Girls	Boys	Girls	Boys	Girls	Boys	Girls	Boys
78	91	102	95	91	95	68	62
Classes		Classes		Classes		Classes	
7		8		7		7	

5. **Open Question** A candy factory wants to do some quality control on its production line to see if it has the right proportion of each flavour in its coloured candy mix. Outline a possible plan to sample the product.

6. Use the Internet to research Audrey Tobias and her connection to the Canadian census.

Literacy Link

A *census* is a count or survey of a population. For example, surveying everyone at your school.

7. Describe how each pair of sampling methods are similar and different. Provide examples to support your answers.

a) multistage versus stratified

b) convenience versus voluntary

Processes

Representing
How could you show your answer? Think about visual ways to compare.

8. A car dealership conducts a phone survey to determine customer satisfaction. The dealership will like to use a stratified sample based on the type of vehicle purchased.

Type of Vehicle	Number of Customers
SUV/truck	858
Minivan	1213
Midsize car	478
Economy car	987
Sports car	221

a) What is the population?

b) If the dealership wishes to conduct 250 surveys, how many calls should it make for each type of vehicle?

c) Why would the dealership choose to do a phone survey rather than mailing a survey to each customer?

d) What else can be done to ensure the survey results represent the population?

Achievement Check

9. Tomato farmers use an automated tomato picker to pick the tomatoes and eliminate non-ripe tomatoes. The machine works very quickly but the elimination of non-ripe tomatoes is not perfect.

a) The farmer sells a 41.9915-tonne load to a buyer who takes a single scoop of tomatoes from a random spot in the load. If the sample has approximately 300 tomatoes and 92% are acceptable, how much is the load worth? Tomatoes sell for $94.40 per tonne.

b) At least 67% the tomatoes in a load must be acceptable for the buyer to purchase the load. If the farmer expects to get 150 loads from his fields each year, by how much could the farmer's income vary? How do you think the variability in the yields affects how the farm's finances are managed?

c) What type of sampling method is used? Do you think it is accurate enough to be reliable? Explain.

Extend

10. Communication Prior to 2011, the Canadian census required approximately 80% of households to complete the short form questionnaire and about 20% to complete the long form (more detailed) questionnaire. In the 2011 census, the long form was replaced by the National Household Survey, which was no longer mandatory. Describe some pros and cons of this change.

11. Thinking It is estimated that there are over 4000 moose in Algonquin Provincial Park. Research how the moose count is done. Does it use a sampling method similar to those described in this section? Explain.

5.3

Collecting Data

Learning Goals

I am learning to

- collect primary data by designing surveys and experiments
- describe the characteristics of an effective survey

Minds On...

People are asked survey questions all the time. Sometimes the way the question is asked or the subject matter of the question can influence the results of the survey.

- If you were asked whether you would help at the local food bank, how would you answer?
- How might the question itself influence your answer?

Action!

Investigate Conducting an Experiment

Materials

- wastepaper basket, bucket, or other container
- at least 3 different types of balls

Conduct an experiment to determine which combination of ball and throwing style will yield the highest probability of a basket. Follow the steps to complete the table.

	Ball 1		Ball 2		Ball 3	
Distance	Underhand	Overhand	Underhand	Overhand	Underhand	Overhand
1 m						
3 m						
5 m						

1. Place the empty basket against a wall.
2. Stand 1 m from the basket. Use an underhand throw to toss one type of ball into the basket. Repeat 10 times and record how many times you got the ball in.
3. Repeat step 2 from a distance of 3 m and 5 m. Record your results.
4. Repeat steps 2 and 3 using an overhand toss.

5. Repeat steps 2 to 4 using a different type of ball. Then repeat steps 2 to 4 using the third type of ball.

6. Based on your data, which type of ball and throwing style combination gives you the best chance of getting the ball in the basket?

7. How is organizing your information in a table helpful for answering the question?

8. When researchers conduct experiments, they often try to keep all variables constant while changing only one thing at a time. How did this experiment follow that model?

9. **Reflect** Why is changing only one variable at a time a good thing to do, in general, when conducting experiments?

10. **Extend Your Understanding** You are planning an experiment involving measuring the time it takes for a soup can to roll down ramps with different angles of elevation. Which variables do you need to keep constant?

There are two different types of experiments: observational studies and experimental studies.

Literacy Link

An *inference* is a conclusion based on reasoning.

In observational studies, researchers look at situations that are already occurring and try to make inferences. For example, a researcher might compare two groups of people, one with members who exercise and another with members who do not exercise, to see if one group is healthier than the other.

In experimental studies, researchers control what is going on and make inferences based on those controls. For example, a researcher might randomly choose two similar groups and have members of one group perform rigorous exercise once a day for 30 days while members of the other group continue with their normal lifestyle. The researchers would then measure the fitness of both groups at the end of the month.

In this example, the group that exercises is the **treatment group** while the one that does not exercise is the **control group**. In an experimental study there is a greater chance of determining the cause of a particular behaviour. Three things need to occur to determine the cause:

- Control: as many aspects of the experiment need to be controlled as possible so that if there is an effect, the researchers know what caused it.
- Randomization: when groups are chosen, subjects need to be randomized so that no biases occur in any of the groups.
- Replication: even though the groups are random, when researchers repeat an experiment they should be similar in make-up so that changes from one group to another are easier to detect.

treatment group

- the participants in an experiment who receive the specific treatment being measured

control group

- the participants in an experiment who do not receive the specific treatment being measured
- compared to the treatment group

Example 1

What Type of Study?

For each case, answer the following questions:

a) Is it an experimental study or an observational study? Give reasons why.

b) If it is an experimental study, what is the control group and the treatment group? What effect is being studied? If it is an observational study, what are some things that could cause the effect to happen?

Case 1: People with headaches are randomly divided into two groups. One group gets pain medication, and the other does not. One hour later, the participants are asked about their pain.

Case 2: You go to all of the houses in your neighbourhood and ask whether they use fertilizer on their lawns. You then check if their grass is green.

Solution

Case 1:

a) Since there is randomization and some aspect of control (some are given the medicine while the others are not), this is an experiment. Even though you cannot tell if the two groups have similar make-up, the other two criteria are met, so this is a well-designed experimental study.

b) Control group: those who did not get medication
Treatment group: those who did get medication
Effect being studied: whether medication decreases headache pain

Case 2:

a) This is an observational study. There are no controls for who used fertilizer and who did not, and subjects are not randomized.

b) The effect being studied is whether fertilizer makes lawns green. Factors other than fertilizer, such as watering, can affect the colour of a lawn. Without good control, it is hard to know what causes any effect.

Your Turn

1. A researcher interviews people as they leave the gym and finds that they get fewer colds compared to people who do not go to the gym.

a) Why is this an observational study?

b) What could be done to turn this into an experimental study?

2. A botanist is studying the effects of acidity on rate of growth. She grows one group of plants using water with neutral pH. She grows each other group using water with increasingly acidic pH levels.

a) Which are the control and which are the experimental groups?

b) Why do you think groups of plants were used rather than one single plant for each pH level?

Example 2

Survey Questions

Conducting surveys is another way to collect primary data. Surveys are less controlled than experiments, but a well-written survey can provide useful information. The following survey is being distributed to 100 students:

School Feelings Survey

Name: _____ Age: ☐ 10–12 ☐ 12–16 ☐ 16–17 ☐ Over 18

Gender: ☐ Male ☐ Female

1. Would you like to have the freedom to use your cell phone in class or would you rather be alone with no communication? ☐ Cell phones ☐ No cell phones
2. Which of the following is your favourite subject? ☐ Math ☐ English ☐ Drama
3. Do you think it's important for students to attend church? ☐ Yes ☐ No
4. How do you like the new cafeteria menu?
   ☐ Pretty good ☐ Good ☐ Great ☐ Fantastic ☐ Awesome
5. How do you like the new school logo and mascot at the school?
   ☐ Like them ☐ Do not like them

Project Prep

If you are collecting your own data for your project, creating valid survey questions will help make sure the data are accurate and represent the population.

Check to make sure the survey asks appropriate questions.

a) Is the survey anonymous? Why might anonymity be important?

b) Are the choices for ages appropriate and clear? If not, how can you fix this section? Why is this important information to collect?

c) Is #1 a leading question? If so, explain why this is a problem and rewrite the question.

d) Do any questions provide a limited number of options? If so, rewrite the question.

e) Does the survey ask any personal questions that people might prefer not to answer?

f) Is the rating scale in #4 a well-written scale? If not, rewrite it.

g) Question #5 asks for an opinion about two things in one question. Explain why this is a problem and rewrite the question.

Literacy Link

A *leading question* is phrased so that it could influence the way a person answers.

Solution

a) No, the survey asks for a name. People may be less likely to answer truthfully, especially if the questions are personal or embarrassing.

b) No. There are several issues with the age range boxes:

- The ranges overlap: 12 and 16 are both covered in two boxes.
- The ranges vary: the 12–16 age group is larger than the others.
- There is no box for an 18-year-old to check.

Use individual ages:

■ 12 and under ■ 13 ■ 14 ■ 15 ■ 16 ■ 17 ■ 18 ■ 19 and over

Demographic information can reveal trends within the population.

Literacy Link

Demographic information includes things like age, gender, level of education, income, residency, ethnicity, and so on.

bias

- occurs when there is a prejudice for or against an idea or response
- biased samples can result from problems with either the sampling technique or the data collection method
- example: a survey question that asks whether you agree that the government should continue to waste money is biased because it leads people to change their opinion toward government spending

c) The cell phone question suggests that having a cell phone is better than not having one. This is a **biased** question. Phrase questions in a way that is neutral and does not reveal your personal opinion. When asking for an opinion, include a "No Opinion" option.

Should cell phones be allowed in class? ■ Yes ■ No ■ No opinion

d) The question about a favourite subject includes only three options. Instead, allow respondents to fill in the blanks.

What is your favourite subject? _____

Open questions can sometimes result in unexpected answers. For example, if a respondent says Shakespeare is his favourite subject, you might choose to categorize that as English.

e) The question about going to a church is unethical. By asking only about church, the question is excluding a potentially large number of people who attend a different place of worship. It is also unrelated to the topic of School Feelings. Consider whether you really need the question in the survey. If you do, rephrase the question and include a third option:

Do you think it is important for students to attend a place of worship?
■ Yes ■ No ■ Prefer not to answer

f) The scale gives only positive choices. Also, the choices are vague. Use clear wording and put the worst option at one end of the scale and the best option at the other end. Place the other choices evenly between.

How do you like the new cafeteria menu?
■ Dislike ■ Do not like or dislike ■ Like

g) A person who likes either the mascot or the logo, but not both, has nowhere to answer. Either ask two separate questions, or give options that cover all the situations.

How do you like the new school logo and mascot at the school?
■ I like both ■ I like the logo but not the mascot
■ I like the mascot but not the logo ■ I dislike both

Your Turn

In each case, identify the problem with the question and rewrite it so that it is more appropriate.

a) What is your favourite game system?
■ Xbox ■ PlayStation® ■ Wii™

b) Running a business is hard. Leadership training will help your business run smoother.
■ Agree fully ■ Somewhat agree ■ Agree a little

c) How important do you think speed and quality of service are?
■ Very important ■ Important ■ Modestly important
■ Of little importance ■ Unimportant

Example 3

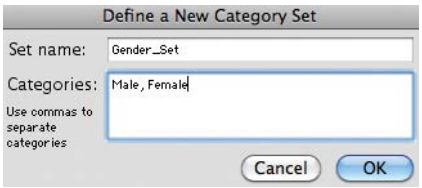
Create a Survey Using Technology

Create a survey to compare how much males and females like baseball.

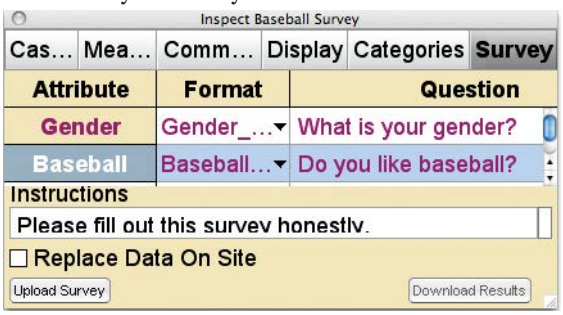
Solution

Method 1: Use Fathom™

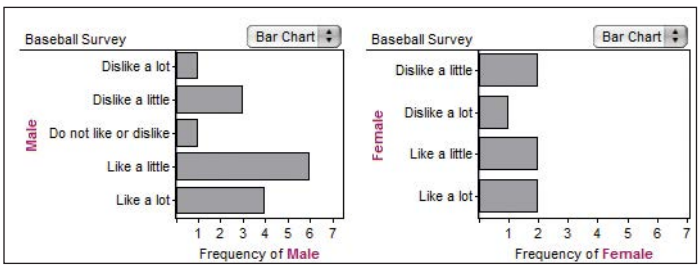
- Open Fathom™.
- Drag a new **Collection** to the workspace. Double click on the icon and name it "Baseball Survey." With your collection selected, from the **Collection** menu choose **Create Survey**.
- Fill in the **attribute** name, "Gender." Under **Format**, choose **Define a New Category Set**. Type in "Male, Female" (a comma separates each choice). Click OK. Under **Question**, type: "What is your gender?"



- Fill in the next **attribute** name, "Baseball." Under **Format**, choose **Define a New Category Set**. Type: "Like a lot, Like a little, Do not like or dislike, Dislike a little, Dislike a lot." Click OK. Under **Question,** type: "Do you like baseball?" In the **Instructions** box, type the instructions for the survey: "Please fill out this survey honestly."



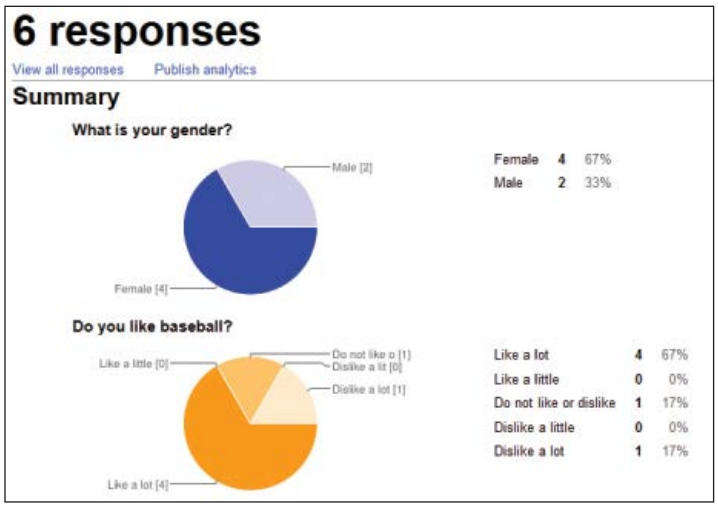
- Click **Upload Survey** to create the online portion of your survey. Your teacher will provide you with a login and password.
- Once uploaded, press **View Survey** to see the survey and get the link. Anyone taking the survey will need the same login and password.
- Once your survey is filled out, click **Download Results** to gather the information in Fathom™.



Method 2: Use Google Docs

- Sign in to or create a Google account and click on **Google Drive**.
- Click on **Create** and then select **Form** from the drop down list.
- Call the form "Baseball Survey."
- Title question 1: "What is your gender?" Select **Multiple Choice** and enter the two options as "Female" and "Male." Click **Done**.
- Click **Add Item**. Then title question 2: "Do you like baseball?" Select **Choose from a list** and enter the following five options: "Like a lot," "Like a little," "Do not like or dislike," "Dislike a little," "Dislike a lot." Click **Done**.

- Click on **Send form**. You can send the survey by entering email addresses, posting a link to the survey on Facebook or Twitter, or embedding it in a website.
- After getting several responses, log in to your Google Drive account and click on **Baseball Survey form (Responses)** to open the spreadsheet that contains all the survey responses. You can cut and paste the data into a spreadsheet or into a Fathom™ collection.
- To get a visual representation of the data, click on the **Form** menu and select **Show Summary of Responses**.



Your Turn

Use a technology of your choice to create a survey with two questions. Share your survey with your classmates.

Consolidate and Debrief

Key Concepts

- In an observational study, the researcher records behaviour and tries to draw conclusions based on the observations.
- Experimental studies try to determine the cause and effect relationship between two variables by controlling for one variable to see what effect it has on the other variable.
- Effective experiments have good control, randomize the members of the treatment and control groups, and try to have a similar demographic make-up in each group.
- Surveys are a powerful way to gain information about a group of people.
- Surveys should be anonymous but can ask for precise demographic information.
- Items on surveys should be clear, concise, and ask only one question that is free of bias.
- Rating scales on a survey should be evenly distributed between good and bad outcomes.
- Data from surveys can be efficiently collected using technology.

Reflect

R1. Why do you think asking clear and concise survey questions is important for gaining information about people?

R2. Clinical tests of medication use a third experimental group called the placebo group. This group takes a medication that has no active ingredient. Members of the placebo group do not know that they are not getting the real medication. Why do you think this group is also studied?

R3. Researchers want to determine whether smoking causes decreased lung capacity. They gather two randomly chosen groups of 100 people. One group smokes one pack of cigarettes a day while the other group does not smoke at all. After two years, the researchers assess the lung capacity in both groups. This is called a longitudinal study. Discuss any possible ethical issues with this study.

Literacy Link

A *longitudinal study* is a type of observational study where researchers measure the same variables over a long period of time, often years or decades.

Practise

Choose the best answer for #1 and #2.

1. Which of the following is an unacceptable rating scale?

A Agree Undecided Disagree

B Frequently Occasionally Rarely

C Usually true Occasionally true Sometimes true

D Important Moderately important Unimportant

2. Which of the following statements is true?

A Observational studies are used to determine the cause of various events.

B Experimental studies must have variables controlled throughout the study.

C Observational studies should have all of their participants randomized.

D Experimental studies keep track of the behaviour of groups that are created by the participants.

3. For each of the survey questions, explain how the question could be improved. Then, rewrite it.

a) Do you exercise daily and get enough sleep at night? ■Yes ■No

b) The *Star Wars* saga is one of the best science fiction stories of all time. How do you feel about Disney taking over the franchise?
■Strongly agree ■Agree ■Don't know
■Disagree ■Strongly disagree

c) Which is your favourite type of music?
■Rap ■Electronic ■Pop

d) How do you feel about the following statement: We should not reduce the number of recycling days in the school.
■Strongly agree ■Agree ■Don't know
■Disagree ■Strongly disagree

4. Use the scale to write a survey question.
■Very frequently ■Frequently
■Occasionally ■Rarely ■Never

5. Identify each case as an observational or experimental study. Indicate what relationship is being studied.

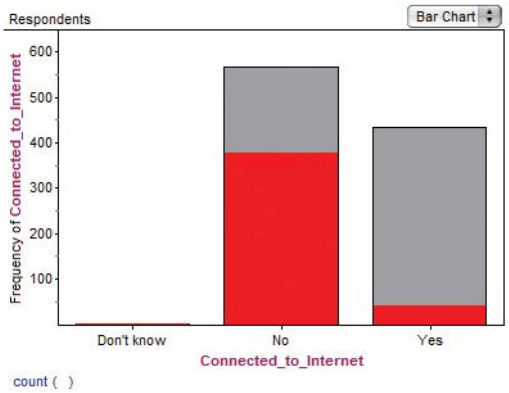
a) People are asked if they have had magnetic therapy and whether their pain is reduced.

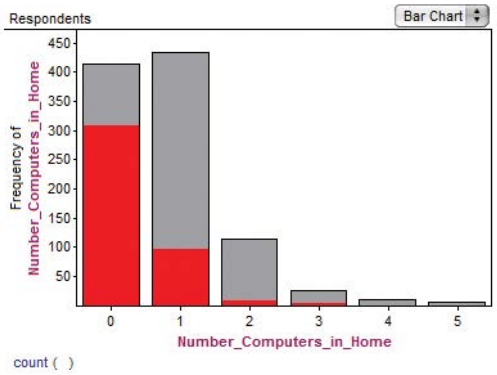
b) As part of a study to gain information regarding travel to Mars, NASA is paying subjects $18 000 to lie in bed for 70 days.

c) Two randomly selected groups of 100 students are studied. One group has all left-handed students while the other has right-handed students. Their math grades are collected and analysed.

Apply

6. In 2000, a survey was given to 1000 Canadians about their use of technology. The results for two of the questions are shown. The red bars show people who answered no when asked whether they had used the Internet.





a) Approximately how many people had no computers in their home?

b) Why do you think a large portion of the people who had no computers are highlighted in red?

c) This survey was a cross-sectional study. How do you think these graphs would change if the same questions were asked now?

Literacy Link

A *cross-sectional study* is a type of observational study where researchers measure the variables from a population at one specific point in time.

d) Your teacher will provide you with a file called **TechUse.ftm**. Create three graphs using the data in it. Describe the story those graphs tell.

7. **Open Question** Write three ineffective survey questions. Exchange with a partner and fix each other's questions.

8. Your teacher will provide you with a file called **2001CensusOntario.ftm**.

a) Double click on the collection and click on the Comments tab. What kind of information is provided?

b) Create two different graphs using location data. Describe what each graph shows.

c) Create three graphs using ethnicity data. What does each graph show?

d) Why is it important to have these data?

9. **Open Question** Create a survey with five questions. Make each question collect a different type of data (categorical, nominal, discrete, numerical, etc.). You could use a situation such as entertainment, sports, the environment, school, or technology.

10. **Open Question** Create a survey with five questions that ask for opinions. Use a different rating scale for each one.

11. For #9 and #10, use technology to create a survey. Have your classmates fill it out. Create a graph to show the results for each question.

12. Organic farmers grow their crops using only natural fertilizers, while traditional farmers use chemical fertilizers. Researchers decide to contact 24 farmers and collect data on the yields of corn grown on 11 organic farms and 13 traditional farms. They find that the organic farms tend to have between 5% and 34% lower crop yields than the traditional farms.

a) What kind of study did they likely conduct?

b) Describe a study that could more formally test whether organic crops yield less than traditional farms.

Achievement Check

13. You want to find out if there is a link between eating habits and grades at your school.

a) **Application** Write 5 to 10 questions to collect data on this topic.

b) Create this survey using technology and have your classmates complete it.

c) Create three or more graphs that compare eating habits and grades.

d) **Thinking** Design an experimental study to see if there is a link between eating habits and grades.

14. **Thinking** A researcher wants to determine whether there is a connection between academic success and birth order. Explain why this could not be done as an experimental study.

Extend

15. **Thinking** The effects of coffee on the human body and mind are widely studied.

a) Search the Internet for three to five different studies on the effects of drinking coffee. Do any of the studies seem questionable? In what way?

b) For each study, identify the type of study and what effect was being measured.

c) Search online for a 2013 study called "Association of Coffee Consumption with All-Cause and Cardiovascular Disease Mortality." What type of study was this? What were the conclusions? What kind of evidence, if any, exists that suggests the conclusions could be reliable?

16. **Communication** Clinical trials often use placebos to help determine if the medication is working. Search for studies that have used placebos. Find at least one where the placebo helped to show the medication worked and one where it did not.

5.4

Interpreting and Analysing Data

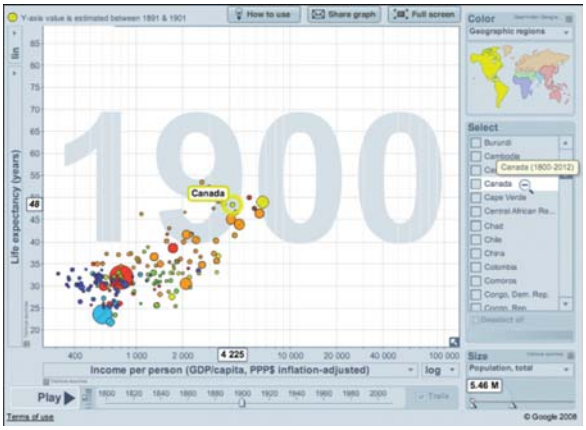
Learning Goals

I am learning to

- distinguish between primary and secondary sources of data
- distinguish between microdata and aggregate data
- collect and analyse data from primary and secondary sources
- collect and analyse data obtained through experimentation

Minds On...

The graphs show data about life expectancy and income per person for 200 countries. In 1900, Canadians had a life expectancy of 48 years, the average income per person was $4225 per year, and there were 5.46 million people in the country. In 2012, Canadians had a life expectancy of 81 years, the average income per person was $35 993 per year, and there were 35 million people in the country. What kind of stories do these data tell about Canada compared to other countries?



Source: Free material from Gapminder World

Source: Free material from Gapminder World

Action!

Investigate Statistics Canada Data

The branch of the Canadian government that is in charge of the census and all the other data collection is called Statistics Canada. The public can access much of the information through CANSIM (Canadian Socio-Economic Information Management System) or actual census data. In this investigation, you will explore how to get the price of goods and services from CANSIM.

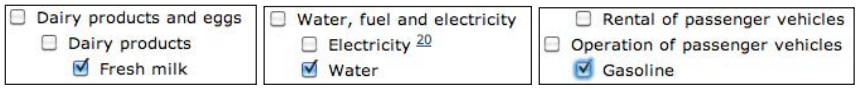
1. Go to the Statistics Canada website and choose **Browse by subject** from the menu. Choose **Prices and price indexes**, then **Consumer price indexes**, and then **Detailed tables from CANSIM**.

2. Scroll down the list until you see **Table 326-0020** on the Consumer Price Index. This table contains information on the price of over 280 goods and services. Click on **Description** and determine the range of years these data cover and what year equals 100. These values do not represent the price of these items but instead the price index.

3. Click on **Table 326-0020**. Describe what you see. Click on the **Add/Remove data** tab.

Source: CANSIM Table 326-0020, Statistics Canada

4. Scroll down and click on **+Expand** from **Step 2-Select: Products and product groups**. Deselect anything currently checked. Put a check beside Fresh milk, Water, and Gasoline.



5. From **Step 3-Select the time frame**, choose the earliest possible year and month and the latest possible year and month.

6. From **Step 4-Select the Screen output format**, choose **HTML table, time as rows**. Then choose **Apply**.

7. What is the earliest date for which data exist for these items? Choose **Add/Remove data** and change the beginning date to the actual time when the data start and choose **Apply**. Be sure to select **HTML table, time as rows** after changing the dates.

8. Select all the data and paste them into a spreadsheet program. Create a single graph that has all three sets of data on it. Refer to the Prerequisite Skills on page 195 for help on plotting multiple series using spreadsheet software.

9. Which set of data did not have data for the entire time frame?

10. **Reflect** Compare how each set of data changed its value over time. Which one seems to vary the most? What is the general trend for all three?

11. **Extend Your Understanding** Determine the current price of each item. Use the Consumer Price Index and the fact that the prices will be proportional to the Consumer Price Index to estimate the cost of each item over this time span. How does this affect the graph of your data?

Literacy Link

The *Consumer Price Index (CPI)* is a value that represents a "basket" of goods that are typically purchased by consumers. The prices are averaged with each item being weighted by importance. When the CPI is high it means that prices in general are high, and vice versa.

Project Prep

Statistics Canada is a huge resource that includes many topics and years. When looking for information for your project, you may wish to search there for the most up to date Canadian data.

primary source data

- data that have been collected directly by the researcher and have not been manipulated or summarized

microdata

- an individual set of data about a single respondent

secondary source data

- data used by someone other than those who actually collected them

aggregate data

- data that are combined or summarized in such a way that the individual microdata can no longer be determined

When you collect and analyse data, it is considered **primary source data**. The individual responses in a survey are called **microdata**. Each line in the table is microdata about each person. The entire table is the researcher's primary data about a group of people.

	A	B	C	D	E
1	Person ID	Gender	Age	Marital Status	Employment
2	1224	Male	43	Married	Employed
3	1225	Female	25	Single	Employed
4	1226	Male	17	Single	Student
5	1227	Male	35	Divorced	Unemployed

Any data that you have not collected on your own represent **secondary source data**. These are often in aggregate form, meaning they have been manipulated in some way. For example, if you collect data about the height of each person in your class, the individual heights are microdata, while the summary of the heights in the form of an average is **aggregate data**.

Example 1

Interpreting Data from Statistics Canada

The table shows the average domestic airfares for 10 Canadian cities.

City	2010	2011	2010 to 2011
	Dollars		% Change
Canada	182.5	190.7	4.5
Calgary	165.5	176.2	6.5
Edmonton	160.8	170.0	5.7
Halifax	172.0	179.3	4.2
Montréal	191.1	194.1	1.6
Ottawa	196.0	194.8	–0.6
Regina	168.1	177.8	5.8
Saskatoon	170.2	178.8	5.1
Toronto	205.2	214.9	4.7
Vancouver	199.2	206.7	3.8
Winnipeg	181.0	189.4	4.6

Source: Table 1 Average domestic air fares for 10 major Canadian cities, *The Daily*, Wednesday, January 9, 2013, Statistics Canada

a) Does this table show microdata or aggregate data? How do you know?

b) Is this table a primary or secondary source of data? Justify your answer.

c) Identify the independent and dependent variables.

d) What kind of story do the data in this table tell?

e) Locate this report from *The Daily* archive on the Statistics Canada website. What other types of things are mentioned in the report?

f) What type of sampling was used? Where did you find that information?

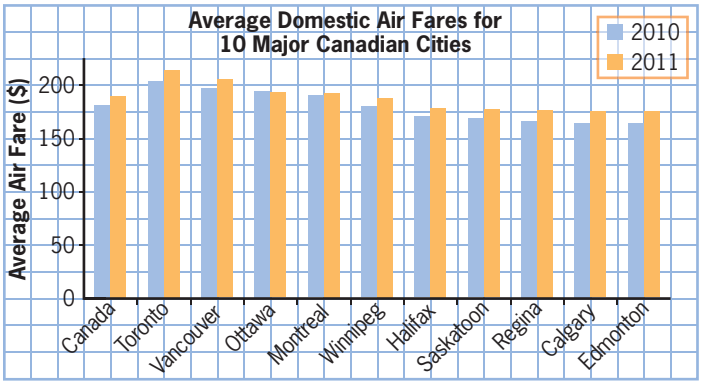
Solution

a) The data represent averages, so this table shows aggregate data.

b) The table is primary data for Statistics Canada, because Statistics Canada collected the data. It is secondary data for anyone else.

c) The independent variable is the city and the dependent variable is the average airfare.

d) Create a graph to help see a pattern in the data.



When organizing categorical data, often it is appropriate to arrange in order of size. Note that Canada goes first because it is not a city.

Except for Montreal and Ottawa, all the cities have a similar increase in ticket cost. Montreal's increase was half as large as the next highest, and Ottawa had a drop in price.

e) Besides a brief summary of the data, there is also a link to the summary table as well as links to the four sets of data that were used to create the table. There is also a link to the actual survey information.

f) A stratified random sample was used. The survey method was found under the section titled "Definitions, data sources and methods."

Your Turn

One of the tables that was used to collect the above data was CANSIM Table 401-0004. A condensed version of the table is shown.

a) Can you see any patterns in the data?

b) What kinds of comparisons could be made between 2008 and 2011?

Average Domestic Fares for Canada and 10 Major Cities

Geography	2008	2009
Canada	196.30	173.00
Halifax	197.20	170.80
Montréal	194.30	177.80
Ottawa	205.80	189.30
Toronto	219.80	194.40
Winnipeg	191.50	169.90
Saskatoon	184.00	160.50
Regina	x	160.00
Calgary	185.20	156.40
Edmonton	180.50	154.20
Vancouver	209.10	182.60

Symbol legend: x Suppressed to meet the confidentiality requirements of the *Statistics Act*

Source: CANSIM Table 401-0004, Average domestic fares for Canada and ten major cities, Statistics Canada, February 18, 2014

Example 2

Analysing a Database

Literacy Link

Many digital files and collections are really *databases*. For example, a music library is a database because every digital recording contains information.

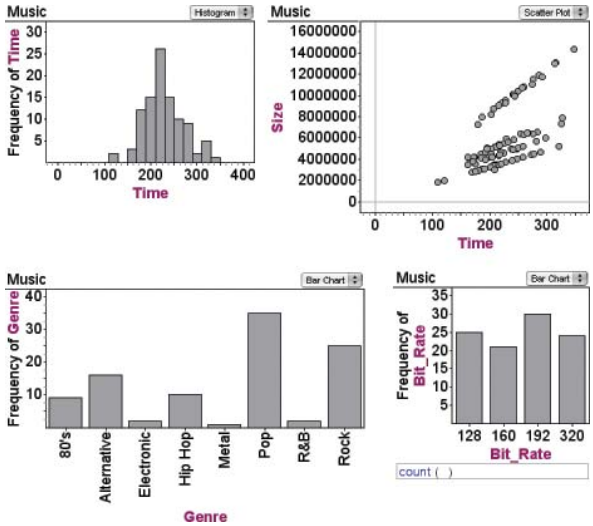
A music library has the attributes shown in the table. Use the table and the graphs to answer the following questions:

Inspect Music

Cases	Measu...	Com...	Dis...	Cat...

Attribute	Value	Fc
Name	Sk8er Boi	
Artist	Avril Lavigne	
Composer	Avril Lavigne/Matrix	
Album	Let Go	
Genre	Rock	
Size	8157076	
Time	203	
Year	2002	
Bit_Rate	320	
Plays	2	

23/100 Show Details



a) What type of data are these?

b) How many songs are in this library?

c) What is a more appropriate arrangement of the Genre graph?

d) What kind of story or stories do the data in the graphs show?

e) How does bit rate relate to the scatter plot of Size vs. Time?

f) Your teacher will provide you with a file called **Music.ftm**. Use it to determine which song(s) is the most played.

Literacy Link

Digital versions of songs are created when the analogue versions are converted piece by piece. This is done by sampling the analogue version and converting these samples to digital pieces. The *bit rate* represents how many samples are taken per second.

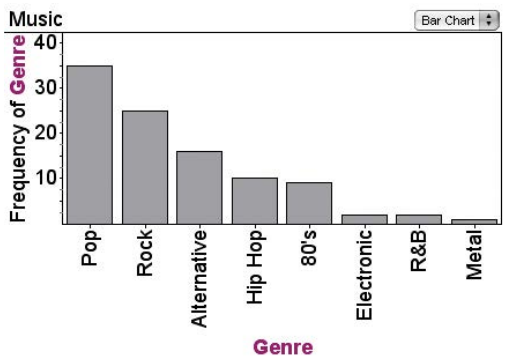
Solution

a) These are microdata. Since they come from a personal library of music, they are also primary data.

b) There are 100 songs in the library.

At the bottom of the attribute list, 23/100 means you are looking at the 23rd piece of data out of 100. How could you use the graphs to see how many songs there are?

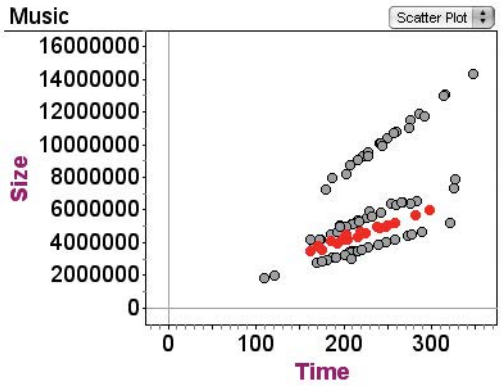
c) Since they are non-ordinal, categorical data, it makes the most sense to arrange the bars from largest to smallest.

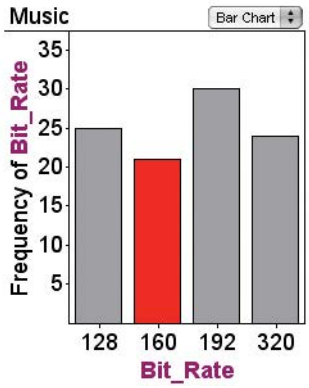


To move the bars around on a categorical plot, drag the names with your mouse.

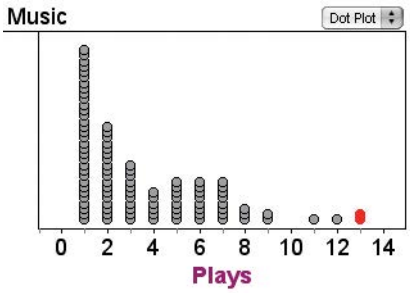
d) Pop and Rock music make up more than half of the songs, and there are more than twice as many Pop songs as there are Alternative songs. In general, the longer the song, the larger the size of the song.

e) Since there are four distinct "lines" of data on the scatter plot, each must correspond to a specific sample rate. For example, if the data with 160 bits per second are selected, you can see their distinct line on the graph. The higher the bit rate, the steeper the line (and the larger the files for the same length of song).





f) Use the attribute called Plays to create a tally of how many times each song was played. Drag a **Graph** from the tool bar and drag the Plays attribute to the horizontal axis. Click on each dot plotted at 13. The name of the song will appear in the list.



Two songs have been played 13 times each: "I Kissed a Girl" and "Remember the Name."

Your Turn

Using the data set from the example, determine the following:

a) What information is given about each song?

b) What kind of story does the Time histogram tell you?

c) Are each of the artist genres equally distributed in terms of the bit rate?

d) If you knew how long a song was, could you determine its size? Explain.

Example 3

Interpreting a Trend on a Time Graph

One of the gases directly linked to the greenhouse effect and climate change is carbon dioxide (CO_2). Keeping track of the amount of carbon dioxide in the atmosphere is an indirect method of measuring the potential for climate change. At the Mauna Loa Observatory in Hawaii, scientists have been collecting atmospheric data since the 1950s. Due to its remote location and minimal influence from vegetation or human activity, this site has become important in the collection of many types of atmospheric data.

Year	CO_2 Reading (ppm)
1993	355.97
1994	356.90
1995	359.57
1996	361.02
1997	362.02
1998	364.16
1999	367.52
2000	368.36
2001	369.61
2002	371.43
2003	373.65
2004	376.03
2005	377.61
2006	380.10
2007	381.80
2008	384.06
2009	385.66
2010	387.51
2011	390.06
2012	392.17
2013	394.66

Source: National Oceanic and Atmospheric Administration, August 28, 2013

a) This data set shows the average monthly carbon dioxide levels (in parts per million) in January of each year for two decades. What type of data are these? Explain.

b) Without graphing, suggest what appears to be happening over time.

c) Create a graph showing these data. Your teacher will provide you with a file called **MaunaLoaSmall.csv** if you wish to use technology. What story do the data tell?

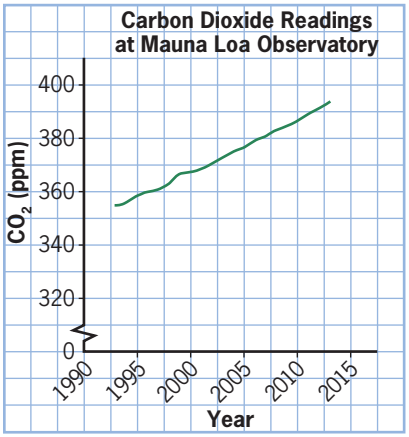
d) Your teacher will provide you with a file called **MaunaLoa.csv**. This is a more complete set of data. Create a graph of the data. Compare the story this graph tells to that of your graph from part c).

Solution

a) Since these are monthly averages, the data have been manipulated. So, they are aggregate data. Although the researchers collected the readings and thus used primary data to generate the summary, these are secondary data for anyone else who uses them.

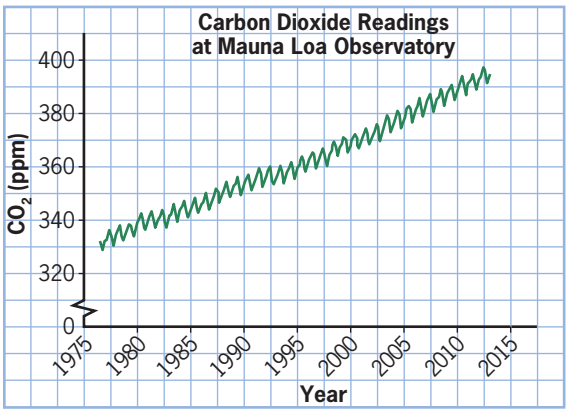
b) The carbon dioxide levels appear to be rising. At the beginning of the data, the CO_2 level was at 355.97 ppm in 1993, and it was at 394.66 ppm in 2013. That is a rise of 38.69 ppm, or 11%.

c) The graph confirms there has been a fairly steady rise in CO_2 readings in the past two decades that does not appear to be slowing down.



Why is it appropriate to use a line graph rather than a scatter plot?

d) This file has seasonal data rather than the annual data, so the graph shows more detail. Although the CO_2 levels are going up as in part c), they are fluctuating as they go up. A closer look seems to indicate that the levels may fluctuate with the seasons.



Your Turn

A student drops a ball from various heights and measures the height of the bounce. The table shows the results.

Drop Height (cm)	Bounce Height (cm)
0	0
10	7
20	14
30	21
40	30
50	38
60	46
70	51
80	60
90	62
100	70

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

b) What is being controlled and what is being changed in this activity?

c) What is the relationship between the drop height and the bounce height?

d) Create a graph of the data. Describe the relationship in the graph.

e) If you dropped the ball from a height of 130 cm, could you use this information to predict its bounce height? If so, explain how.

Consolidate and Debrief

Key Concepts

- Primary sources of data are collected directly from the source and are not manipulated or summarized in any way.
- Microdata are the individual pieces of data that make up all of the primary data.
- Secondary sources of data are used by someone who did not collect them. Often these data have been manipulated and summarized. Data found in the media often are secondary data.
- Data that are summarized in some way are called aggregate data.
- Large sources of data are available for analysis on the Internet.
- Sources of data also are hidden in digital items like songs and photos.

Reflect

R1. How are websites for the NHL, MLB, NBA, and NFL like databases?

R2. Which is more visible to the public, primary or secondary data? Why do you think this is?

Practise

Choose the best answer for #1 and #2.

1. Which of the following is microdata?

A batting averages for a baseball team

B the height of each person in your class

C monthly sales for each store in a chain

D number of births in each country in 2013

2. Go to the Statistics Canada website and search the CANSIM database. Which of the following tables contain information about vacations?

A 128-0014 **B** 203-0026

C 427-0004 **D** 361-0013

3. For each of the CANSIM tables in #2, describe the topic of the data.

4. In each of the following cases, determine whether the data are primary or secondary.

a) Your town sends a survey to each household and collects the data.

b) After collecting the survey from part a), your town publishes the average number of people per house on each street.

c) The newspaper reports the salary of every city council member.

d) A magazine lists the total number of home sales each month for the last year.

e)

Person ID #	Age	Job	Hours Worked
345	24	Shipper	38
231	38	Custodian	42
124	29	Mail clerk	40

f)

Grade	Percent of Students Who Take the Bus to School	
	Male	Female
9	78	82
10	85	80
11	74	81
12	65	64

Apply

5. Go to the Statistics Canada website and choose **Census of Canada**, **Data products**, and **Census Profile**. Search for the city or town closest to where you live.

a) What percent of people are in your age group?

b) What is the largest age group in your city or town?

c) Answer parts a) and b) again, for Toronto. If you live in Toronto, use Sarnia.

d) Compare the population, population density, and percent of married people between the two cities.

6. **Thinking** The data below represent all the earthquakes over magnitude 4 on the Richter scale that occurred in North or South America in 2012.

Month	Day	Name	Depth (km)	Magnitude
1	30	Peru: ICA	43	6.4
3	20	Mexico: Guerror, Oaxaca	20	7.4
3	25	Chile: Parral, Santiago	41	7.2
4	17	Chile: Valparaiso	29	6.7
5	14	Chile: Arica; Peru: Tacna	10	6.3
5	19	Brazil: Montes Claros	10	4.1
8	27	El Salvador: Off the coast	28	7.3
9	5	Costa Rica: Nicoya	35	7.6
10	28	Canada: Queen Charlotte Islands	14	7.7
11	7	Guatemala: San Marcos, San Cristobal Cochu	24	7.3

Source: National Geophysical Data Center

a) A total of 48 earthquakes over magnitude 4 occurred all over the world. What percent were in North or South America?

b) Where in Canada did the only earthquake above magnitude 4 occur? Was this a big earthquake in comparison to the others?

c) Represent this information in a graph. Why did you choose that graph type?

d) Your teacher will provide you with a file called **Earthquake2012.ftm**. It contains more detailed, world-wide data about all of the earthquakes over magnitude 4 in 2012. Use Fathom™ to create a new graph. Drag the **Longitude** to the horizontal axis and the **Latitude** to the vertical axis. What does this graph represent? Drag the **Magnitude** onto the centre of this graph to create a "heat map." Why do you think this is called a heat map?

e) Where is there a concentration of large magnitude earthquakes?

f) Search the Internet for the Significant Earthquake Database. Get all of the earthquake data for the last complete year available. Create a graph and compare to 2012.

7. **Thinking** Every time you take a picture, your camera records details about the photo. These data are called the Exif information.

a) Your teacher will provide you with a file called **Images.ftm**. It contains Exif information for over 500 images. What kind of information is stored about each image?

b) Create three graphs that use numerical data and three graphs that use categorical data. Why do you think this sort of information would be important to a photographer?

c) Create a scatter plot comparing file size to the number of pixels in the image. Some cameras predict how many more photos you can take before the memory card runs out of space. How can this comparison be used by your camera to make this prediction?

8. Some of the largest databases are weather-related. One of the wettest places in Canada is Henderson Lake, BC. Weather stations there measure an average of 9082 mm of rain every year. Consider some of the wettest communities listed below.

Canada's Wettest Communities	
Location	**Rain (mm)**
Hartley Bay, BC	4549
Holberg, BC	3912
Port Renfrew, BC	3671
Tofino, BC	3306
Prince Rupert, BC	3111
Gold River, BC	2846
Wreck Cove Brook, NS	1945
Pools Cove–Fortune Bay, Nfld	1829
Louisbourg, NS	1599
Sydney, NS	1505
Alliston, PEI	1182
Charlottetown, PEI	1173
Sept-Iles, Que	1156
Stratford, Ont	1064
Cameron Falls, Alta	1103

Source: "Rainiest Places in Canada" Current Results

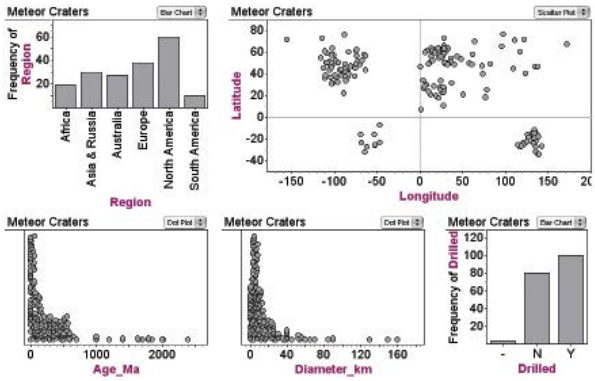
a) Are these data primary or secondary? Give reasons.

b) People in Stratford probably think they get a lot of rain. Compare Stratford, Hartley Bay, and Henderson Lake to get a sense of how much rain these places are getting.

c) Create a graph of the data. How does the graph indicate that BC is likely the wettest province? Your teacher will provide you with a file called **WettestCommunities.csv** if you wish.

Achievement Check

9. There are over 180 known meteor craters on the planet. Most of these impact sites are millions of years old. The graphs show some data that scientists have collected about them.



a) Is this is a primary or secondary source of data? Justify your answer.

b) Are these data the result of a study or an experiment? Justify your answer.

c) North America has had a substantially larger number of impact craters than other regions. Why might this be of interest to scientists? Is there any reason why this could be perfectly normal?

d) Your teacher will provide you with a file called **MeteorCraters.ftm**. Which region has had the lowest proportion of craters drilled to have their rocks tested? Why might this make sense?

e) Does it appear that the older the impact, the larger the diameter? Justify your answer. Create a new graph to demonstrate your point.

Extend

10. **Thinking** Go to the Gapminder website. From the **Menu** choose **Data** and in the search enter "life." From the options for the indicator named "Life Expectancy (Years)," choose **Visualize** . Slide the slider back to 1800 and click on the Play button.

a) Describe what happens to the graph.

b) Explain what the size and colour of the circles represents. What else can the size of the circle represent?

c) What kind of story is told by the data? How is Canada part of the story?

d) Go back to the Data page and find a different data set. What story does your new set tell?

5.5

Bias

Learning Goals

I am learning to

- distinguish among types of bias when sampling data
- analyse and interpret statistics presented by the media to judge their validity
- identify different ways that graphical data can be misleading

Minds On...

On competitive reality shows, the audience often votes for contestants. But phone voting systems are often overwhelmed. In one instance, two contestants both had about 24 million votes, but over 200 million calls did not get through. What does this say about the reliability of the voting system?

Action!

response bias
- when respondents change their answers to influence the results, to avoid embarrassment, or to give the answer they think the questioner wants

sampling bias
- when the sample does not closely represent the population

measurement bias
- when the collection method is such that the characteristics are consistently over- or under-represented

non-response bias
- when the opinions of respondents differ in meaningful ways from those of non-respondents

Certain types of bias make it difficult to get a representative sample or truthful responses.

Type of Bias	Example
response bias	A teacher asks students to raise their hand if they cheated on last week's test. Students will not want to admit to cheating on a test so it is unlikely that many will raise their hand.
sampling bias	A politician goes back to the farming community she grew up in to ask for opinions on her latest initiative for the agriculture industry. It is likely that a larger proportion of the people she speaks to would support the initiative, both because it would benefit them and because she grew up in the area. This would not accurately represent the entire population.
measurement bias	A survey question asks, "A lot of people do not like math. How would you feel being referred to as a math geek?" This is a leading question; the wording of the question can affect the outcome by influencing someone's answer. Other types of measurement bias can occur when the collection method affects the results, for example when the options in a multiple choice question are too limited for an honest response.
non-response bias	A mail-in survey asked respondents about their drinking habits. Only 3% of the surveys were returned. Such a small return rate would likely not yield a representative sample. In fact, those who respond often have very strong opinions about the subject matter and so the results could easily over- or under-estimate the feelings of the population.

Example 1

Identifying Bias

Project Prep

When collecting your own data or using a secondary source of data, it is important to make sure the data are free of bias. If you fail to do this, the integrity of your project will be compromised.

A large sample does not guarantee good data. This is especially true if the method of data collection has some bias associated with it. Identify the type of bias that may occur in each situation.

a) Families in a neighbourhood are told they are part of a study about healthy eating habits.

b) You ask only students on sports teams how to spend the school fundraising money.

c) A survey question asks, "Who is the best basketball player of all time, Michael Jordan or LeBron James?"

d) A radio call-in show asks callers to answer the question, "Are you in favour of a law that would ban pitbulls from the city?"

Solution

a) This is response bias. Because the families are told the study is looking at healthy eating habits, it is likely they will start eating better because they are being watched and want to give a good impression.

b) Since you are only asking students who play sports, it is likely the respondents will overwhelmingly want to spend the money on sports-related things. This is sampling bias.

c) This is measurement bias, because there are only two choices.

d) Usually, the people who call in to radio shows have extreme opinions. People who are indifferent or do not think the topic is important may not vote, so a large proportion of listeners choose not to respond. This is non-response bias. Also, since only listeners of the radio show will call in, it is also sampling bias since the respondents likely would not properly represent the population.

Your Turn

Identify the type of bias that may occur in the following situations.

a) A survey question asks, "How many words per minute can you read?"

b) A survey is sent to parents of school-age children that asks whether bus safety lanes should be installed.

c) A phone company surveys its customers via text message about which services people like the best.

d) A survey asks, "Now that the city is in debt, do you think the current mayor will win the next election?"

Example 2

Misleading Statistics

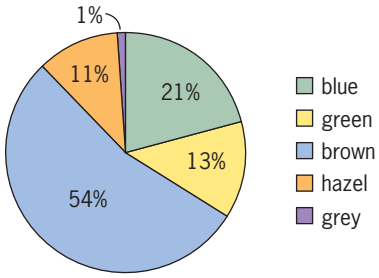
Explain how each case uses misleading or questionable statistics.

a) The *Daily Mirror* and *The Times* retweeted this tweet about a British high-speed rail project (HS2) that is expected to take 20 years to complete.

> Interesting stat: 93% of today's UK population will be dead by the time HS2 is finished. Wow. #bbcqt

b) In 2005, the British Cheese Board conducted a study to show that eating cheese does not give you nightmares.

c) **Classmates' Eye Colour**



d) A newscaster reports about a survey that asked, "Did scientists falsify research to support their own theories on climate change?" and shows the following summary:

59% Somewhat likely
35% Very likely
26% Not very likely

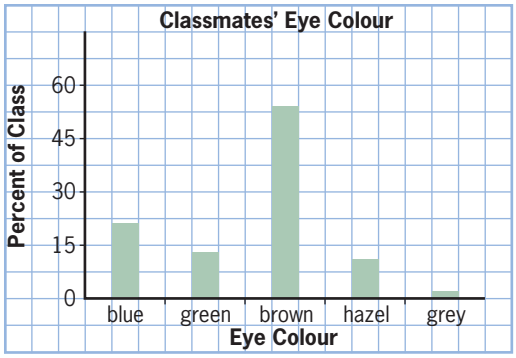
e) Research shows that if you eat a sausage a day, your risk of a certain type of cancer goes up by 20%.

Solution

a) The original figure was mathematically implausible since it is unlikely that such a high proportion of the population will die in the next 20 years. The media outlets did not check whether the statistic was true before they retweeted it.

b) When a company or organization conducts research to disprove negative effects, the method of collection and validity of the data should be studied closely. In this case, the subject of the study is nightmares, which are intangible and hard to quantify and verify.

c) Circle graphs are hard to visually interpret if the sizes of the wedges are similar because judging the size of areas is difficult. In this graph, you know how big each section is because the percents are shown. A bar graph would allow people to more easily compare the sizes of the bars even without the percents labelled.



d) The values in this summary add up to more than 100%. This mistake was made by a major news service. When you refer to the original survey, you can see the results were different.

Newscast	Actual Survey
59% Somewhat likely	35% Very likely
35% Very likely	24% Somewhat likely
26% Not very likely	21% Not very likely
	5% Not at all likely
	15% Not sure

The newscast combined the top two percents and combined the third and fourth responses. But they also included the data for "Very likely" again, skewing the data further.

e) When dealing with percents, it is often important to have more information to determine whether the number is significant. In this case, 20% sounds like it could be a significant increase. However, you need to know the current rate of this type of cancer. If about five people in 100 will have it, a 20% increase is actually only going from five out of 100 to six out of 100.

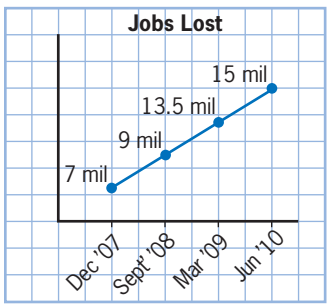
Literacy Link

An *increase of 20%* can sometimes be misinterpreted as, "The current rate is 5%; if the rate goes up 20%, the new rate becomes 25%." However, 20% is a proportion, so it is referring to 20% of the 5%. This is 1 percentage point. So, the increase could be stated in two ways: as it was originally, "a 20% increase," or less ambiguously, "a 1 percentage point increase."

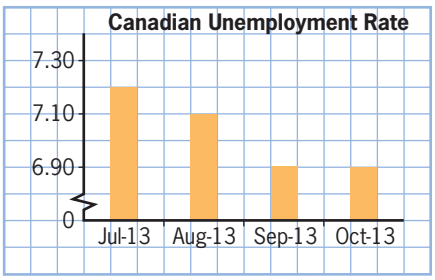
Your Turn

Explain how each case uses misleading or questionable statistics.

a) A news report on jobs lost shows this graph.



b) A newspaper headline states "Unemployment Dropping" and includes this graph.



c) A yogurt manufacturer states that 67% of customers surveyed claimed they lost weight while eating its brand of yogurt.

Example 3

Some Numbers Are Bigger Than Others

Sometimes numbers that seem big are really small, and other times numbers that seem big truly are. When you see big numbers, try to put them into perspective.

a) A recent settlement of $52.5 million was awarded to 970 000 households from Québec affected by the 1998 ice storm. Is this a big settlement?

b) Switzerland has the highest consumption of chocolate per person. The population is 8.0 million people and they eat about 82 million kg of chocolate each year. Is that a lot of chocolate?

Solution

a) Determine the amount of money awarded to each household.

$\frac{52\ 500\ 000}{970\ 000} = 54.1237\ldots$

Each household will get $54.12.

At first $52.5 million sounds like a lot of money, but when you consider all the households that received money, it seems much less significant.

b) First, determine the amount of chocolate eaten per person.

$\frac{82\ 000\ 000}{8\ 000\ 000} = 10.25$

Each person in Switzerland eats 10.25 kg of chocolate per year.

Next, get a sense of how much that is. One chocolate bar is about 50 g.

$\frac{10\ 250\ \text{g}}{50\ \text{g}} = 205\ \text{bars}$

$\frac{205\ \text{bars}}{52\ \text{weeks}} = 3.9\ \text{bars/week}$

The average person in Switzerland eats approximately four chocolate bars per week. Someone who does not eat chocolate regularly would consider this a lot of chocolate, but someone who eats chocolate every day would not.

Your Turn

a) Every year, Canadians and Americans spend over $8 billion on Halloween. Do you think this is a lot of money? Justify your response.

b) In 2011, Apple Corporation's profit grew by almost $19 000 000 000 from 2010. Do you think this is a large increase? Justify your response.

Example 4

Interpreting Infographics

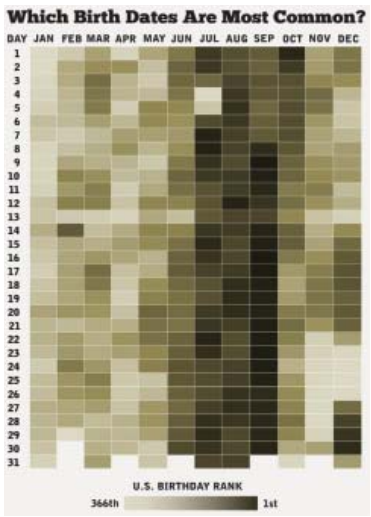
Infographics combine both visual and text-based representations of data. In some cases the representations are standard, while others are more unique or stylish.

a) Describe how this infographic was constructed.

b) Why do you think this representation is called a heat map?

c) The scale represents the ranking. How does this show that ranking is a problem (especially if that is all you know about the data)?

d) What kind of story is told by the data?



Source: NYTimes, Amitabh Chandra, Harvard University. Photo from the Daily Viz.

Solution

a) Each rectangle represents a day of the year. The data are based on the number of birthdays each day. The day with the most birthdays is ranked 1st and has the darkest colour. The day with the fewest birthdays is ranked 366th and has the lightest colour.

b) This is called a heat map because colour coding is often used to represent temperature. The data are represented by the shade of colour. Instead of representing heat or temperature, these data show rank on a list.

c) This method only ranks the data, it does not show how significant the differences are between the data. Consider the two fictional classes below.

Rank		1	2	3	4	5	6	7	8	9	10
CLASS A	Student	T	A	D	F	S	G	H	U	E	O
	Mark	93	89	88	87	87	85	84	82	81	80
CLASS B	Student	Z	B	R	M	K	N	L	J	X	W
	Mark	95	87	83	77	72	68	67	60	59	51

Both classes have 10 students ranked from highest to lowest. In Class A, all of the marks are 80 or above. In Class B, student Z has the highest mark of both classes, but student W has the lowest mark of both classes. Student Z's mark of 51 is significantly lower than the lowest mark in Class A of 80. If you consider only rank, you see only the order of the marks and do not know that even though Student O is ranked last in Class A, the mark of 80 is actually a fairly high mark.

d) Even though you don't know the actual number of birthdays on each day, you can see that the highest numbers of birthdays are around September.

Your Turn

Consider the infographic from CBC's *Hockey Night in Canada*.

a) How are Canadian teams distinguished from other teams?

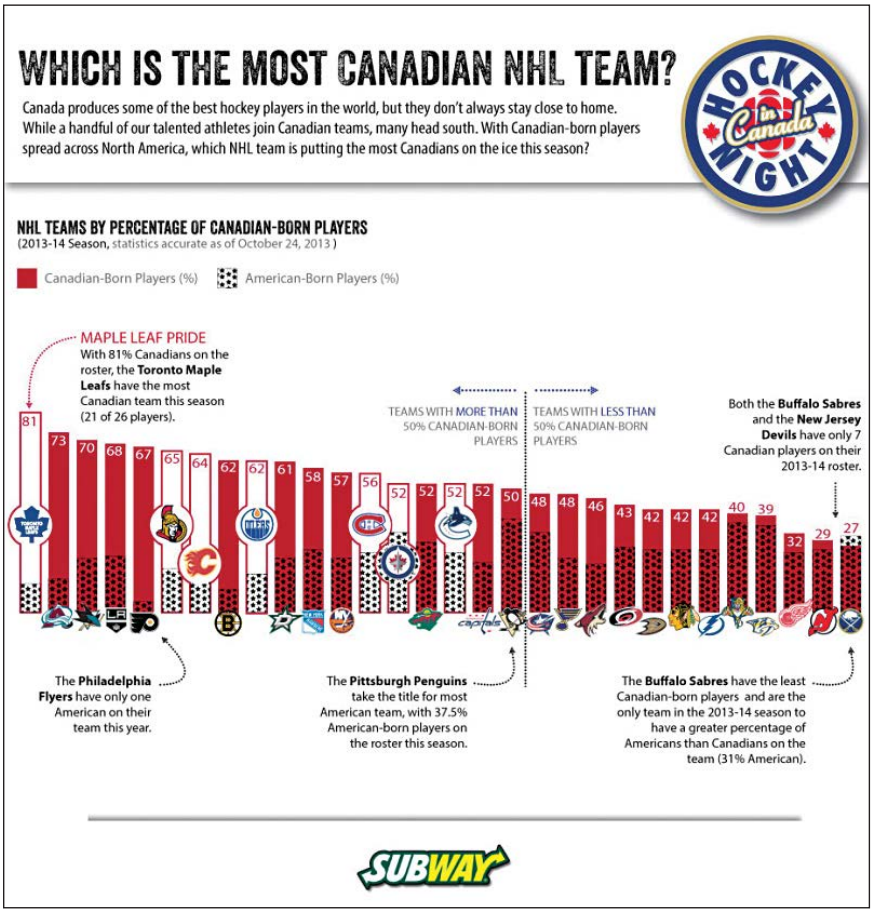
b) How could you determine which team had the fewest North American players?

c) How could you use the data to argue that hockey is Canada's pastime, but not the pastime of the United States?

d) What information is difficult to determine from this representation?

e) Is this infographic free of bias? Explain.

f) Write a tweet (140 characters or less) that is related to this infographic.



Source: "Infographic: The Most Canadian NHL Team," CBC, October 25, 2013

Consolidate and Debrief

Key Concepts

- For data to be valid, collection methods must be free from bias.
- The data can be affected if the collection methods suffer from sampling, measurement, response, or non-response bias.
- Different ways of displaying data can distort it and make it biased.
- Large numbers should always be put in context.
- Infographics can be dense with information or convey an idea with unique methods.

Reflect

R1. Companies often want to back up their product with scientific claims. Is it possible to make misleading claims even if the data are free from bias?

R2. How is it possible for someone to "lie" with statistics yet still be telling the truth?

R3. In recent years, many soft drink companies have developed vitamin-enriched drinks, claiming they have health benefits. When the owners of vitaminwater® were sued for making false health claims, part of their defence was that

"no consumer could reasonably be misled into thinking vitaminwater was a healthy beverage or was composed only of vitamins and water because the sweet taste of vitamin water puts consumers on notice that the product contains sugar."

Source: Gleeson, John, "Memorandum and Order," CV-09-0395 (JG) (RML).

What is the company claiming and why should it make you wary as a consumer?

Practise

Choose the best answer for #1.

1. Which of the following is not a leading question?

A We have recently upgraded SurveyMakers to become a first-class tool. What are your thoughts on this first-class site?

B Should the mayor fix the dirty, potholed streets in the city?

C As a good patriotic citizen, do you think you should buy an imported car?

D How important is health care compared to other social issues?

2. Identify the type of bias and suggest how the same data could be researched without bias.

a) Your teacher asks you to raise your hand if you understood the lesson.

b) City council conducts a survey about the police force by asking everyone who comes to the next council meeting.

c) The local news posts a poll on its website asking if climate change is real.

d) An application to join a gym asks how much you like exercising.

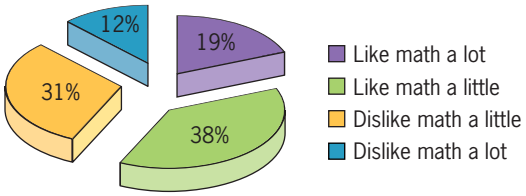
e) A store that sells hunting gear petitions to lower the cost of registering a gun.

3. **Open Question** Choose one of the four types of bias and develop a situation that would result in biased data. Trade with a classmate and identify each other's type of bias.

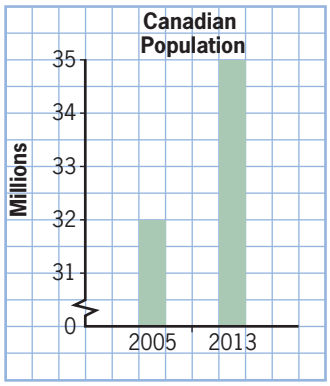
4. Identify the problem with the graph or situation.

a) A drug company states that in its clinical trials, patients had 20% fewer headaches.

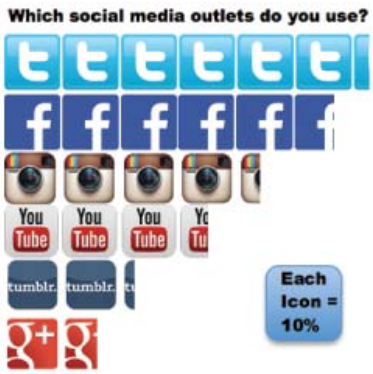
b)



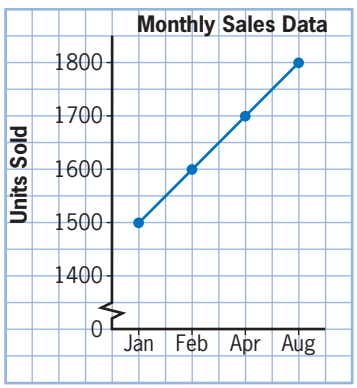
c)



d) Teens were surveyed to see which social media sites they use. The following results were given:



e)



Apply

Achievement Check

5. Elections are integral to democracy.

a) In an election, the polls are open for a particular period of time. In some cases, the party in power drops the number of polling booths in areas where they have low support and increases the number of polling booths in areas where they have more support. Describe how this may affect the results and how it is biased.

b) In 2011, Canada held a federal election. The results were:
Conservative Party: 39.6%
NDP: 30.6%
Liberal Party: 18.9%
Bloc Québécois: 6.0%
Green Party: 3.9%

If you wanted to make the Conservative win sound bigger, would you use percents or percentage points to compare the results? Explain.

c) Voter turnout for the 2011 election was 61.1%. The population eligible to vote was 24 257 592. What percent of Canadians actually voted for each party? How does that compare to the percent that did not vote? What type of bias could this represent?

d) In the 1898 election only 44.6% of the population voted. What implications might this have had on the actual results of the election?

e) **Thinking** Your teacher will provide you with a file called **VoterTurnout.csv**. It contains data from all of Canada's past elections. Create an infographic that displays some aspects of the data.

6. Determine if the number is significant.

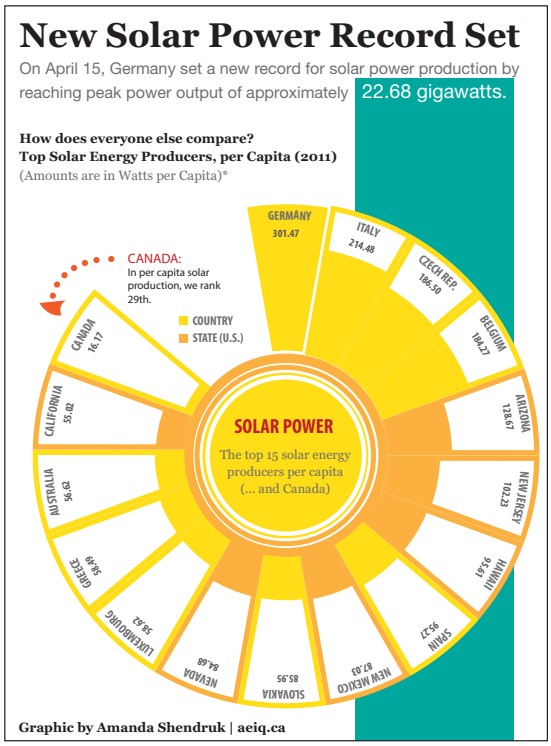
a) Canada Post delivers 9.8 billion pieces of mail each year.

b) Over 70 billion hours of video are watched on YouTube each year.

c) Clayton Kershaw earns $30 000 000 a year to play baseball.

d) In 2013, the Canadian national debt was over $650 billion.

7. *Maclean's* magazine published an infographic about solar energy.



a) What makes this infographic appealing?

b) How does Canada rank as a solar-energy producing nation? Why might this be?

c) How does the way the data are displayed visually connect them to the topic of solar power?

d) What do you think the editors are trying to convey with the statement in the centre of the circle: "The top 15 solar energy producers per capita (… and Canada)."

8. Communication In 2013, the Canadian Cancer Society reported a 1.6% rate of a certain type of cancer in women. If this were to rise to 2% in 2014, create two headlines to relay this information. Make one headline shocking and one neutral.

9. Communication Research "push polling" and explain how it affects elections.

10. Research the Hawthorne effect and give an example of when it might occur.

11 Communication There are types of bias other than the four types described in this chapter. Research the following types of bias and describe whether you have experienced any of them.

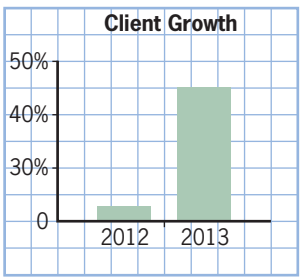
a) confirmation bias

b) observational selection bias

12. An agricultural research firm conducts a survey of local farmers.

a) One of the questions on the survey asks, "The use of high-tech farming methods increases yields by up to 20%. How important is it for you to use high-tech farming methods?" What type of bias does this question have? Rewrite it so it is free of bias.

b) The brochure includes this graph. Why do you think the firm presents the information this way?

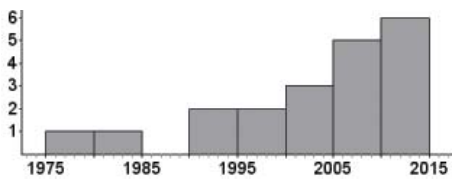


13. The table shows the all-time North American box office Top 20 up to 2013.

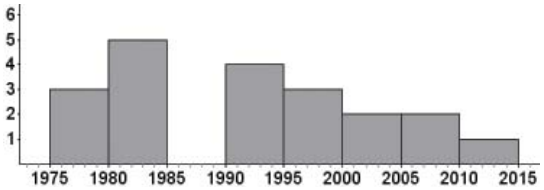
Rank	Released	Film Name	Total Box Office
1	2009	*Avatar*	$760 507 625
2	1997	*Titanic*	$658 672 302
3	2012	*Marvel's The Avengers*	$623 279 547
4	2008	*The Dark Knight*	$533 345 358
5	1999	*Star Wars Ep. I: The Phantom Menace*	$474 544 677
6	1977	*Star Wars Ep. IV: A New Hope*	$460 998 007
7	2012	*The Dark Knight Rises*	$448 139 099
8	2004	*Shrek 2*	$441 226 247
9	1982	*ET: The Extra-Terrestrial*	$435 110 554
10	2006	*Pirates of the Caribbean: Dead Man's Chest*	$423 315 812
11	1994	*The Lion King*	$422 780 140
12	2010	*Toy Story 3*	$415 004 880
13	2013	*Iron Man 3*	$408 992 272
14	2012	*The Hunger Games*	$408 010 692
15	2002	*Spider-Man*	$403 706 375
16	2009	*Transformers: Revenge of the Fallen*	$402 111 870
17	1993	*Jurassic Park*	$395 708 305
18	2011	*Harry Potter and the Deathly Hallows: Part II*	$381 011 219
19	2003	*Finding Nemo*	$380 529 370
20	2005	*Star Wars Ep. III: Revenge of the Sith*	$380 270 577

a) What is surprising about the data?

b) The graph shows the movies organized by the year of their release. What does it say about the fairness of this comparison?



c) This graph is adjusted for inflation. How is it different? Why does it make more sense?



Literacy Link

Inflation is an increase in prices over time. When comparing dollar amounts from different time periods, it is important to *adjust for inflation*.

d) Your teacher will provide you with a file called **Top20Movies.csv**. It contains both the adjusted for inflation and non-adjusted Top 20 movies. How does the list change when comparing the unadjusted to the adjusted movies?

e) **Thinking** Create an infographic that displays the information provided above.

Extend

14. **Thinking** Even though celebrities may or may not be experts, they can often be influential because they are visible in the media. Vaccines have had some controversy associated with them.

a) Research how these two facts are connected.

b) Why is there such caution about vaccines and why might this be unfounded?

15. **Communication** When medication is being developed, pharmaceutical companies use complex procedures to verify that the drugs actually work. To be completely sure, they use a procedure called a double blind study. Research double blind studies and explain why they are so important.

Chapter 5 Review

Learning Goals

Section	After this section, I can
5.1	• show how data are used and misused in statistical studies • identify different types of data • understand that there is variability in data • see that you can analyse single sources of data or related sources
5.2	• distinguish between a population and a sample • understand why sampling a population can give information about that population • understand that when sampling data the results can vary • sample data in various ways
5.3	• collect primary data by designing surveys and experiments • describe the characteristics of an effective survey
5.4	• distinguish between primary and secondary sources of data • distinguish between microdata and aggregate data • collect and analyse data from primary and secondary sources • collect and analyse data obtained through experimentation
5.5	• distinguish among types of bias when sampling data • analyse and interpret statistics presented by the media to judge their validity • identify different ways that graphical data can be misleading

5.1 Data Concepts and Graphical Summaries, pages 196–205

1. Give an example of a variable that could be measured using each type of data.

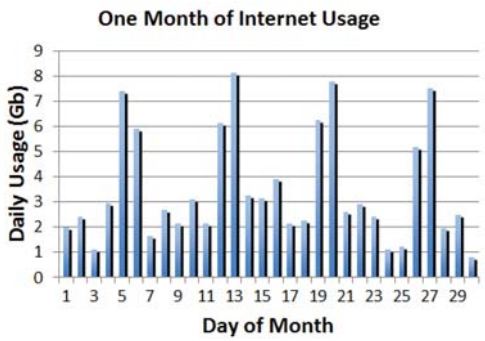
a) numerical, discrete

b) numerical, continuous

c) categorical, ordinal

d) categorical, nominal

2. The graph shows the Internet usage of a family of five for one month.



a) Account for the peaks in the data. Justify your answer.

b) The family's Internet plan has a 100 Gb monthly data limit. Estimate whether they went over their limit.

c) Your teacher will provide you with a file called **InternetUsage.csv**. Check your answer from part b).

d) Determine the average daily Internet usage. Do you think this number is useful for the family? Explain why or why not.

5.2 Principles of Data Collection, pages 206–211

3. Determine the type of sampling.

a) Your favourite social networking site asks what your favourite band is.

b) The Pelee Island Bird Observatory does its annual bird count every December. Researchers set up traps in the trees in three locations to gather the birds.

5.3 Collecting Data, pages 212–221

4. For each survey question, identify the problem with the question and rewrite it.

a) How old are you?

■ 15 and below ■ 15–20
■ 20–35 ■ 35–60 ■ Above 60

b) Violence in video games could cause people to be violent in real life. How do you feel about a violence rating system for video games?

■ Strongly agree ■ Agree
■ Agree a little ■ Don't agree

c) Do you like the new logo for the school?

■ Yes ■ No

5. Determine whether each study is observational or experimental.

a) In 2006, a research team asked nearly 5000 households for charitable donations to see which methods produced higher amounts. When the solicitor was an attractive female, the average donation increased by 50–135%, especially if a male answered the door.

b) A teacher looks at class averages over the last 10 years and sees that the smaller the class size, the higher the class average.

5.4 Interpreting and Analysing Data, pages 222–232

6. In each case, determine whether the data are from a primary or secondary source.

a)

Percent of Students Achieving at or Above Provincial Standard in Grade 9 Math				
	School		School Board	
Year	Applied	Academic	Applied	Academic
2011	66	92	45	85
2012	43	86	49	86
2013	56	91	58	87

b) Your classmates fill out a survey that asks for their age, height, gender, eye colour, and favourite food.

c)

Late Night TV Hosts' Yearly Salaries (millions)	
Jon Stewart (*The Daily Show*)	$35
Jimmy Fallon (*The Tonight Show*)	$12
Jimmy Kimmel (*Jimmy Kimmel Live*)	$10

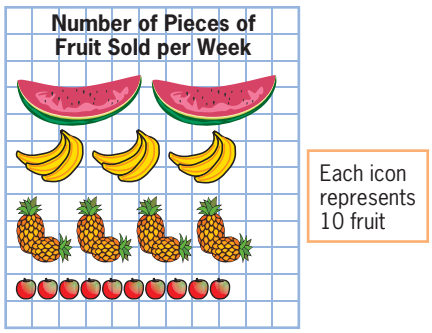
5.5 Bias, pages 233–243

7. Give an example to show each type of bias.

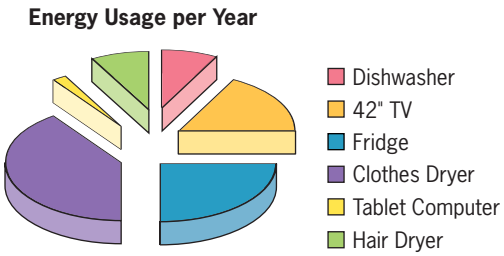
a) sampling **b)** measurement
c) response **d)** non-response

8. In each case, indicate how the representation is misleading.

a)



b)



9. Create an infographic of the Canadian data.

Total area = 9 985 000 000 km^2
Land area = 9 094 000 000 km^2
Population = 35 million people
Population density = 3.95 people per km^2
Population growth = 1.05% per year
Life expectancy = 81.2 years
Median age = 40.4 years
Internet usage = 30.26 million people
Mobile phone usage = 26.26 million people
Number of students = 6.188 million people
Number of teachers = 278 000 people
Health spending = $3895 per person per year

Chapter 5 Test Yourself

Achievement Chart

Category	Knowledge/ Understanding	Thinking	Communication	Application
Questions	1, 2, 3, 4, 5	9, 10	5, 6, 7, 8, 9, 10	6, 7, 8

Multiple Choice

Choose the best answer for #1 to #4.

1. Which of these are significant numbers?

A Facebook surpasses 1 trillion page views per month.

B Health care spending in Ontario will go up by $2.1 million.

C On average, a person will breathe about 500 million times in their life.

D Approximately $135 million is donated to the Canadian Cancer Society each year.

2. Which of the following is not a valid sampling method?

A Order data from highest to lowest, randomly select one item, and then choose every fifth value after that.

B Put all the data into a bin and randomly select as many as you wish.

C Arrange the data into demographic groups and pick 100 from each group.

D Send out a survey and ask people to fill it out and send it back.

3. Which of these is a good survey question?

A Where do you get most of your information on video games and movies?
■ Friends ■ Game store ■ Internet ■ TV

B Should formula sheets be allowed on math exams?
■ Strongly agree ■ Agree ■ Don't know
■ Disagree ■ Strongly disagree

C Which uniforms do you prefer for the football team, the new modern style or the old style?
■ New ■ Old

D What is your favourite sport?
■ Soccer ■ Tennis ■ Volleyball ■ Golf

4. Which situation comprises microdata?

A the average height of students in each class in your school

B the height of every student in your school

C the total number of recycled cans collected by each school

D the average annual salary of citizens in different cities

Short Answer

5. Describe why these situations have bias and what you could do to eliminate it.

a) A company comes door-to-door selling rooftop solar panels. The sales representative claims that by the end of the third year you will be making more than $500 per month.

b) Your teacher explains a difficult topic and then asks the students if they understand.

6. Your teacher will provide you with a file called **Waste.csv**.

a) Create a graph of these data.

b) What story is told by the data?

7. Create a short survey that you could give to your classmates on one of the following topics:
- Internet usage
- climate change
- poverty

8. The table shows data on energy production in Ontario.

Type	Nuclear	Gas	Hydro	Coal	Wind	Other
Energy Produced (MW)	12 998	9987	7939	3293	1725	122

a) Create a graphical representation of the data that is not misleading.

b) Create a graphical representation of the data that is misleading. Explain why it is misleading.

Extended Response

9. Researchers surveyed 881 people to find out what province each person was born in.

NFLD	PEI	NS	NB	QUE
18	10	36	16	225

ONT	MAN	SASK	ALTA	BC
301	29	33	80	133

a) Graph these data.

b) What type of sampling method was likely used? Why?

c) The researchers also collected data from a much larger group about the number of trips people made.

NFLD	PEI	NS	NB	QUE
3500	1091	8115	5458	60 169

ONT	MAN	SASK	ALTA	BC
90 174	7984	8624	21 558	22 380

A newspaper published the headline "Ontarians travel much more than people from other provinces." Why might this headline be incorrect?

d) How could the data be displayed more accurately?

10. a) Collect data on a topic of your choice.

b) Consider sampling methods, non-biased techniques, and good survey questions to develop a strategy that would lead to collecting data that are
- highly credible
- not very credible

Chapter Problem

Food Production

It is estimated that by 2050 there will be over 9 billion people living on the planet. Some estimates suggest that by then we will need 30–70% more food and 40% more water. As the population continues to increase, why is it important for farmers to use data management techniques to help them grow their crops? To justify your answer, think about ways farmers can use
- primary data collection
- experimental design
- interpretation of data from the media or advertising
- sampling techniques