CHAPTER 6

One-Variable Data Analysis

Athletes are often compared to one another. Team managers make generalizations about players' abilities to decide what they are worth. What statistical calculations could help the manager

- determine the value of a player's contract?
- decide which player to hire for their team?
- decide how much money to pay a player for a five-year contract?

Popular musicians are compared with artists from the past and present. What statistical calculations could you use to compare

- the number of days on the top 10 charts?
- the most listened to genre of music?
- songs that have lasted the test of time?

In this chapter, you will learn various methods of statistical analysis and assess their strengths and weaknesses.

Key Terms

mean	variance
median	standard deviation
mode	*z*-score
outlier	multiple bar graph
percentile	split bar graph
quartile	relative split bar graph
range	reliable data
interquartile range (IQR)	valid data

Literacy Strategy

A mathematics word wall identifies words and phrases that you need to understand to develop mathematical skills and reasoning. Careful use of visuals helps support understanding of key words. How does this image convey the meaning of the word "median"?

As you work through this chapter, think about some of the key words you encounter. Use visuals and letters to show the meaning of words for a classroom word wall.

Career Link

Risk Assessment Manager

In the financial services industry, many careers involve assessing business risks and taking measures to control or reduce the risk. Victoria is a risk management analyst who assesses the risks related to defaults on loans. It is her job to assess whether borrowers will be able to repay the loans they receive and whether her firm will get enough return on investment.

- Who do you think would be a lower risk, a young employed couple buying a house, or a startup company in the information technology field?
- How do you think risk level affects the type of loan that each could get?

Chapter Problem

Used Car Lot Business Report

Retail sales outlets regularly use statistics to describe their products, and used car lots are no exception. The amount of time a car sits, unsold, on the lot can affect the value of the car. In this exercise, you will analyse time spent on the lot and present a report of your findings.

a) List some pros and cons of buying a used car.
b) List three things car buyers look for when purchasing a used car.
c) A typical used car may take about a month to sell. What do you think would happen to the price of a used car after it has been on the lot for three months?
d) What other statistics will affect the price of the car?

Prerequisite Skills

Measures of Central Tendency

The three most common measures of central tendency are mean, median, and mode.

Mean

Example: To determine the mean of the numbers 15, 21, 12, 4, 13, and 5, add all the data values within the set:

$15 + 21 + 12 + 4 + 13 + 5 = 70$

Divide the total by the number of values.

$\frac{70}{6} \approx 11.7$

Mode

Example: The mode is the number that appears most often. There can be one mode, more than one mode, or no mode. If no numbers repeat, then there is no mode.

In the following data set, the mode is 11.

11 3 4 9 25 8 1 11 3 4 11

Median

Example: To find the median, put the data set in order from least to greatest. Find the value(s) in the middle of the data set.

2 4 5 5 6 7 9 12 15 15 16 17 20 25

There are two middle values. Determine the average of the two values.

$\frac{9 + 12}{2} = 10.5$

1. Determine the mean, median, and mode of each set of data points.

a) 45, 24, 62, 12, 43, 73, 98, 58, 12, 81, 25, 12, 43, 52

b) 6, 14, 3, 14, 21, 20, 14, 16, 19, 6, 7

c) 12.3, 15.8, 9.9, 13.0, 12.7, 16.1, 20.0, 8.3

d) 102, 134, 187, 155, 142, 134, 134, 156, 181

2. The mean of the set of numbers is 15. What is the missing number?

13, 16, 15, 20, 14, ■

Types of Data

Categorical data can be sorted by category. Ordinal data can be ranked. Numerical or quantitative data are in the form of any number.

3. Identify each type of variable as categorical, ordinal, or quantitative.

a) hair colour

b) salary

c) gender (M, F)

d) rating scale (low, medium, high)

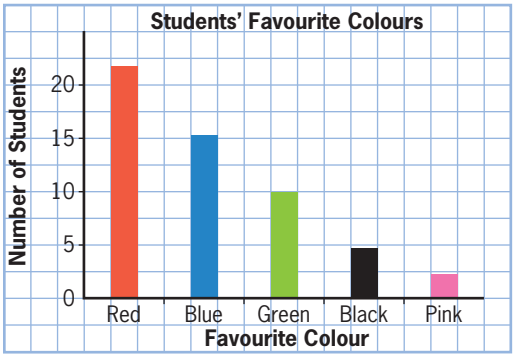
e) level 1, 2, 3, or 4 on a standardized test

f) temperature

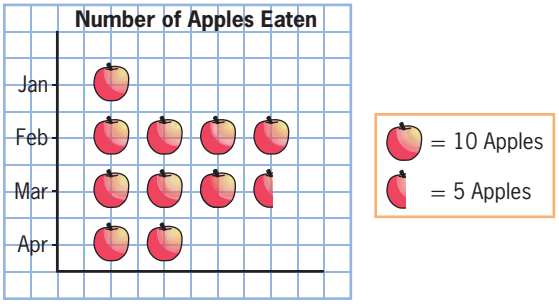
Graphical Summaries

You can use a variety of graphical summaries to display data. Some are better for discrete data, while others are better for continuous data. You will often need to judge which option is the most meaningful way to display your data.

Bar Graph

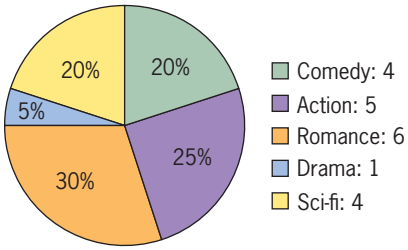


Pictograph



Circle Graph

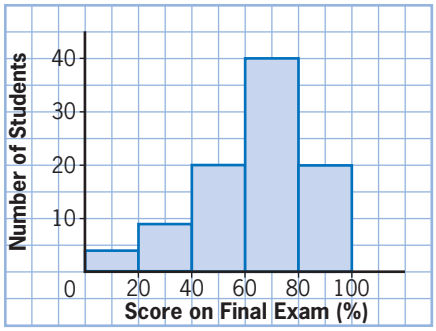
Favourite Type of Movie



Example: To determine the central angle, multiply the percent of the data for each category by 360 degrees. For example, 25% of those surveyed preferred action movies.

$0.25 \times 360° = 90°$

Histogram



Stem and Leaf Plot

Example: Create a stem and leaf plot for the following data:

170 173 173 175 179 179 179 181 181 182 183 187 188 188 189

Represent the hundreds and tens digits as the stem. Represent the ones digits as the leaves. Enter all values in numerical order.

Stem	Leaf
17	0 3 3 5 9 9 9
18	1 1 2 3 7 8 8 9

4. Match each description with the appropriate type of graph. Some graph types can be used more than once.

Description

a) Displays discrete data in separate columns.

b) Organizes data by representing part of each number as a stem and the other part as a leaf.

c) Uses pictures or symbols to represent data.

d) Displays data as a percent of the whole.

e) Uses proportional areas of the bars to show frequencies of the values of the variables.

f) Represents data using a circle that has been divided into sectors.

g) Represents nominal data, such as days of the week.

Graph

Bar graph
Histogram
Circle graph
Stem and leaf plot
Pictograph

5. A survey asked grade 12 students to state their first choice for the university they would like to attend from the list given.

University	Frequency
Queen's University	33
University of Waterloo	32
University of Toronto	41
York University	35
University of Ottawa	40
McMaster University	31
University of Western Ontario	33

Illustrate the data using two graph types of your choice. Justify your choices.

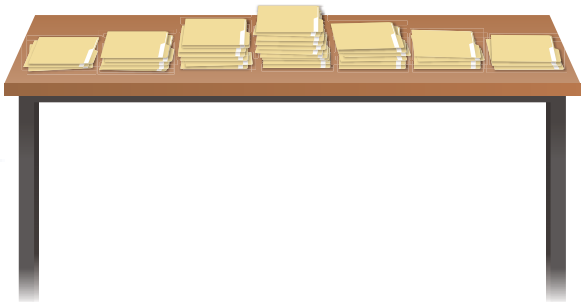
6. Create a stem and leaf plot for each list of student averages. Compare with a histogram.

a) 78, 83, 77, 73, 61, 99, 65, 80, 55, 67, 52, 79, 43, 59

b) 63.2, 71.4, 79.5, 50.0, 93.7, 44.5, 87.6, 65.7, 54.9, 92

6.1

Measures of Central Tendency



Learning Goals

I am learning to

- interpret the mean, median, and mode of a set of data
- choose the measure of central tendency that best describes the data

Minds On...

It is important to collect and organize data in a way that helps you understand where the majority of the numbers are found. This clustering of data is referred to as measures of central tendency. The three most commonly used measures of central tendency are **mean**, **median**, and **mode**.

Think about the following situations:

- A post-secondary admissions board uses students' individual averages to decide who will be most successful in a particular field of study. How might a post-secondary admissions board analyse the cluster of applicants' grades to help its decision-making process?
- How might coaches of a sports team use averages to determine how well a player is performing in the season and how much money she should be paid?
- How might economists use median household income to divide the country into equal income distribution groups?
- How might shoe and clothing store managers use the mode to make decisions about which products to stock?

mean

- the sum of the data entries divided by the number of entries

median

- the middle value of all the data points when the data values are listed in order from least to greatest
- if there is an even number of data points, then the median is the average between the two middle values

mode

- the data value that occurs most often in the list of data points
- it is possible to have no mode, one mode, or more than one mode

Action!

The Greek letter μ, pronounced "mu," is used to represent the population mean. $\bar{x}$, read as "x-bar," is used to represent the sample mean.

In statistics, you can find the mean of a population and the mean of a sample of that population. A sample mean will approximate the actual mean of the population.

Population Mean

$$\mu = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

Sample Mean

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

where N is the size of the population and n is the sample size.

Although the calculations are the same, different symbols are used to indicate whether it represents a population or a sample.

Investigate Where the Money Falls

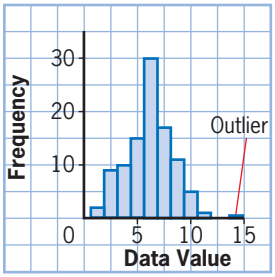
Sport Tech, a sporting goods company, is deciding how it should spend its advertising funds this year. Accountants analyse the company's largest sporting goods accounts for each season. They decide, based on the data presented in the table below, to spend the bulk of this year's budget on advertising winter sports gear.

Season	Sport	Revenue ($)
Winter	Skiing	1 000 000
	Skating	120 000
	Snowboarding	1 200 000
	Snowmobiling	525 658
	Ice fishing	455 200
Spring	Baseball	200 120
	Soccer	450 000
	Tennis	250 000
	Golf	1 000 000
Summer	Surfing	750 000
	Skateboarding	345 200
	Swimming	120 000
	Beach volleyball	120 000
Fall	Football	1 100 000
	Swimming	120 000
	Golf	1 000 000
	Soccer	450 000

1. Refer to the information in the table. Do you agree with the accountants' conclusion? Explain why or why not.
2. What data values may have contributed to the accountants' decision?
3. Determine the mean revenue for each season. Show how you calculated this value.
4. What is the median revenue for each season?
5. Determine the mean and median for the overall sports revenues.
6. **Reflect** What data values may have contributed to the differences in mean and median values between each season and overall? Do these data values have a greater effect on the mean or the median?
7. **Extend Your Understanding** Write a proposal to describe where you think the funds should be distributed. Consider whether you would choose one season or more than one season. What factors might contribute to why a particular sports season generates more revenue than another?

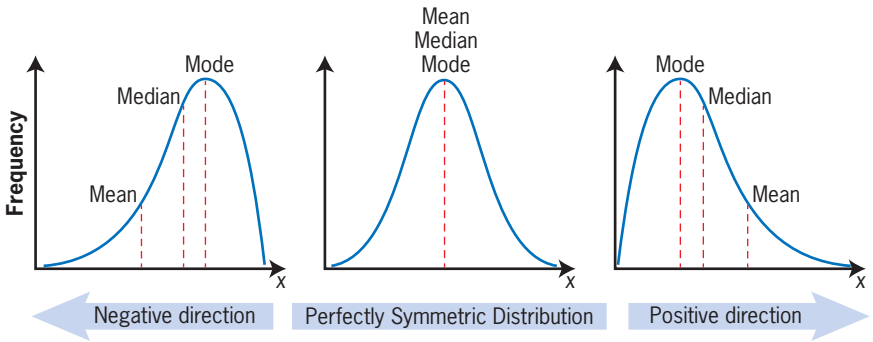
outlier

- an element of the data set that is significantly different from the rest of the data points



The measures of central tendency of a data set can be affected by the presence of **outliers**.

- In a symmetric distribution such as the uniform distribution, the mean, median, and mode will all be equal.
- In a non-symmetric or skewed distribution, the mean, median, and mode will differ.
- In a positively skewed distribution, the mode will be the lowest of the three values and the mean will be the highest.
- In a negatively skewed distribution, the mode will be the highest of the three values and the mean will be the lowest.



Recall that you studied uniform distributions in section 4.2. Why are the mean, median, and mode all equal in a uniform distribution?

Example 1

Evaluating Measures of Central Tendency

You are interviewing for an internship at a risk assessment firm to gain experience for your post-secondary program. The interviewer tells you that the average annual income of the 15 employees at the company is $73 518.27. The chart shows the actual incomes of the 15 employees.

$34 983	$18 980	$12 500	$48 980	$478 320
$17 305	$36 540	$12 500	$250 921	$32 654
$45 678	$33 855	$25 676	$33 450	$20 432

a) Determine the mean, median, and mode of the incomes.

b) Use the measures of central tendency to decide whether the interviewer's statement is accurate.

c) What is the effect of the outliers on the measures of central tendency?

d) Which measure of central tendency best represents the "average" income of the employees?

Solution

a) Method 1: Use Paper and Pencil

Determine the mean of the 15 income values.

Recall that the Greek letter Σ, pronounced "sigma," is used to represent the sum of a series of numbers. $\sum x$ means $x_1 + x_2 + x_3 + \ldots + x_n$.

$$\mu = \frac{\sum x}{n}$$

$$= \frac{34\,983 + 18\,980 + 12\,500 + 48\,980 + \ldots + 20\,432}{15}$$

$$= \frac{1\,102\,774}{15}$$

$$\approx 73\,518.27$$

The mean income is \$73 518.27.

There are an odd number of data points. To determine the median, place the values in order from least to greatest. Then, locate the 8th term because it is the middle of the 15 data points.

\$12 500 \$12 500 \$17 305 \$18 980 \$20 432 \$25 676 \$32 654
\$33 450 \$33 855 \$34 983 \$36 540 \$45 678 \$48 980 \$250 921
\$478 320

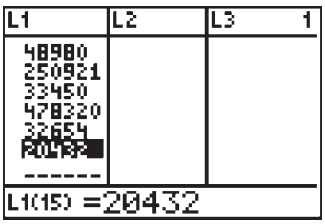
The median is \$33 450.

The mode is \$12 500 because it occurs twice in the set of data points.

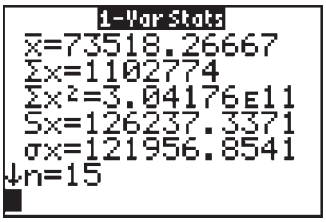
\$12 500 \$12 500 \$17 305 \$18 980 \$20 432 \$25 676 \$32 654
\$33 450 \$33 855 \$34 983 \$36 540 \$45 678 \$48 980 \$250 921
\$478 320

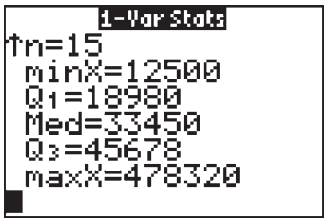
Method 2: Use a Graphing Calculator

- Press STAT, then **1:Edit...**.
  Enter the salaries in list **L1**.
- Press 2ND QUIT.
- Press STAT. Use the arrow keys to choose **CALC**, then **1:1-Var Stats**. Press ENTER.



- Scroll down to read the mean and median, along with numerous other statistics. The calculator will not identify the mode.

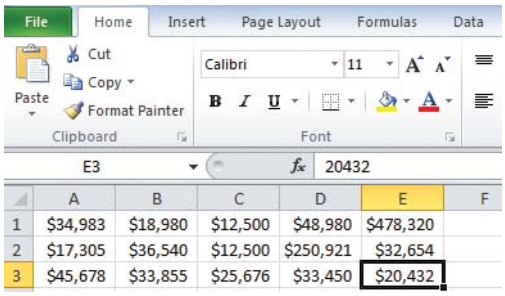




The mean income, $\bar{x}$, is \$73 518.27, and the median income, Med, is \$33 450. You can see the mode of \$12 500 by inspecting the table.

Method 3: Use a Spreadsheet

Open a spreadsheet and enter the data as shown.

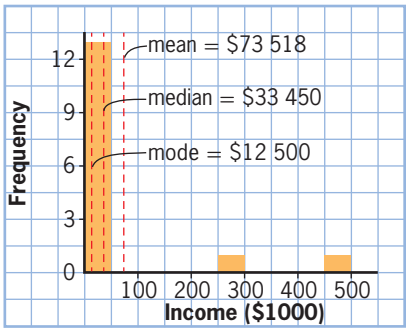


	A	B	C	D	E	F
1	$34,983	$18,980	$12,500	$48,980	$478,320	
2	$17,305	$36,540	$12,500	$250,921	$32,654	
3	$45,678	$33,855	$25,676	$33,450	$20,432	

Remember that commands may vary from one spreadsheet to another. Explore the spreadsheet you are using and record any variations to the commands for measures of central tendency.

To find the mean, click on an empty cell and enter the formula **=average(A1:E3)**. Repeat this step using the command **=median(A1:E3)** and **=mode(A1:E3)** to find the median and mode of your data.

	A	B	C	D	E
1	$34,983	$18,980	$12,500	$48,980	$478,320
2	$17,305	$36,540	$12,500	$250,921	$32,654
3	$45,678	$33,855	$25,676	$33,450	$20,432
4					
5	Mean	$73,518			
6	Median	$33,450			
7	Mode	=Mode(A1:E3)			



This gives a final mean income of $73 518, a median of $33 450, and a mode of $12 500.

b) If you look only at the calculation of the mean, the interviewer's statement is accurate because the average of a set of data is usually defined as the mean. However, when analysing all three measures of central tendency—mean, median, and mode—it is evident that the mean is affected by the two outliers, $250 921 and $478 320, and the mode is the smallest income. So the median is more representative of the data. Therefore, the interviewer's statement is not an accurate reflection of the income of an average employee.

c) The outliers cause the location of the cluster of the data to be skewed. As a result, the mode is the least of the three measures, and the mean is the greatest. The outliers cause the mean to be inflated because the two highest incomes are much larger than the rest of the data set.

Calculate the mean without the outliers of $250 921 and $478 320.

$$\mu = \frac{\sum x}{n}$$
$$= \frac{34\,983 + 18\,980 + 12\,500 + 48\,980 + \ldots + 20\,432}{13}$$
$$= \frac{373\,533}{13}$$
$$\approx 28\,733.31$$

Without the two outliers, the mean income is $28 733.31.

d) The mean is significantly affected by the outliers, the mode is the smallest income of all the employees, and the median is the middle income of all the employees. The median would be the best measure to represent the average income. Looking at the data, this median appears to be fairly close to many data points and provides a description of the data without taking the outliers or repeated values into effect.

Your Turn

Before heading on vacation to Mexico, you observe the actual high temperatures for seven days. The table shows the temperatures.

Day	Temperature (°C)
1	27
2	29
3	32
4	29
5	45
6	29
7	31

a) Determine the mean, median, and mode of the temperatures.

b) The weather report predicts that based on the previous seven-day forecast, the temperature on the day of your arrival should be 36 °C. Use the measures of central tendency in part a) to determine whether the weather report is accurate.

c) Is there an outlier in the data? How does it affect the measures of central tendency?

d) Which measure of central tendency would best represent the temperatures in this Mexican location? Explain.

When the quantity of data is large, you can group the data into intervals to make them easier to analyse. When data are grouped into intervals, you can only approximate the centre of the data. To do this, assume that the data are evenly spaced in each interval, and use the midpoint to represent the values in each interval. Multiply the data values by their respective frequencies. Then, add these products and divide by the total frequency. You can use the following formula to approximate the mean for grouped data.

Mean for Grouped Data

$$\bar{x} = \frac{\sum f_i m_i}{\sum f_i}$$

where m_i is the midpoint of each interval and f_i is the frequency of each interval.

You can use a frequency distribution table to help organize your data.

Example 2

Using Grouped Data

The time taken to complete a chess game was recorded, to the nearest minute. The frequency table shows the data.

Time (min)	10–15	15–20	20–25	25–30	30–35
Frequency	2	20	18	10	5

a) Calculate the estimated mean, median, and mode times, in minutes, to complete a chess game.

b) Describe potential issues with finding the measures of central tendency of grouped data.

c) Graph the data using a histogram. Mark the measures of central tendency on the graph.

d) Discuss any skewing of the data with respect to the measures of central tendency.

Solution

a) Method 1: Use Paper and Pencil

Use the following table.

Number of Minutes	Midpoint, m_i	Number of Games, f_i	$m_i f_i$	Cumulative Frequency
10–15	12.5	2	25	2
15–20	17.5	20	350	22
20–25	22.5	18	405	40
25–30	27.5	10	275	50
30–35	32.5	5	162.5	55

$$\begin{aligned}\sum f_i &= 2 + 20 + 18 + 10 + 5\\ &= 55\end{aligned}$$

$$\begin{aligned}\sum m_i f_i &= 25 + 350 + 405 + 275 + 162.5\\ &= 1217.5\end{aligned}$$

Calculate the grouped mean.

$$\begin{aligned}\bar{x} &= \frac{\sum m_i f_i}{\sum f_i}\\ &= \frac{1217.5}{55}\\ &\approx 22.14\end{aligned}$$

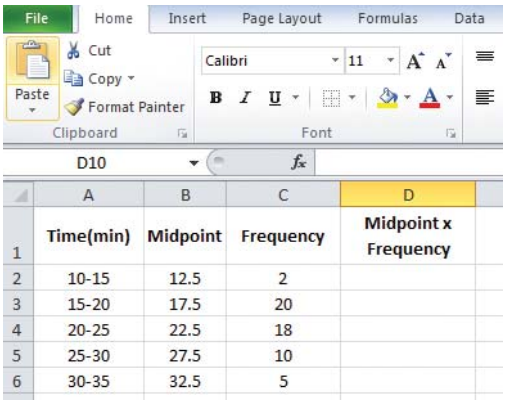
Therefore, the mean number of minutes is approximately 22 min per game.

To determine the median, look for the interval in which the middle value occurs. There are 55 data values in this example, so the median will be the $\frac{55+1}{2} = 28$th term. The frequency table shows that the 28th term occurs within the 20–25 min interval. Therefore, the estimated median is 22.5 min.

When working with grouped data, use a modal interval in place of a mode. The modal interval is the interval with the greatest frequency, namely 15–20 min.

Method 2: Use a Spreadsheet

Enter the data into a spreadsheet.

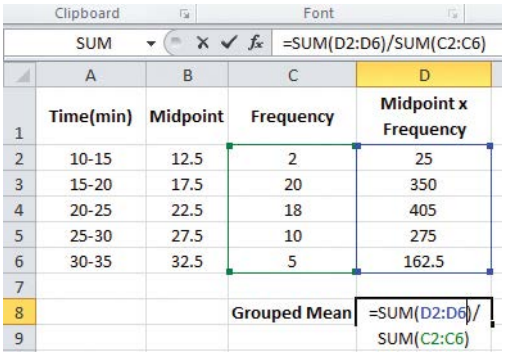


	A	B	C	D
1	Time(min)	Midpoint	Frequency	Midpoint x Frequency
2	10-15	12.5	2	
3	15-20	17.5	20	
4	20-25	22.5	18	
5	25-30	27.5	10	
6	30-35	32.5	5	

When entering the first column of data, change the cell format to text to avoid having an interval (such as 10-15) converted to a date (such as Oct 15).

In cell D2 enter the formula **=B2*C2**, and copy this entry from cell D2 to D6.

To find the mean, divide the sum of column D by the sum of the frequencies in column C. Click on an empty cell. Enter **=Sum(D2:D6)/Sum(C2:C6)**. This yields a grouped mean of approximately 22 min.

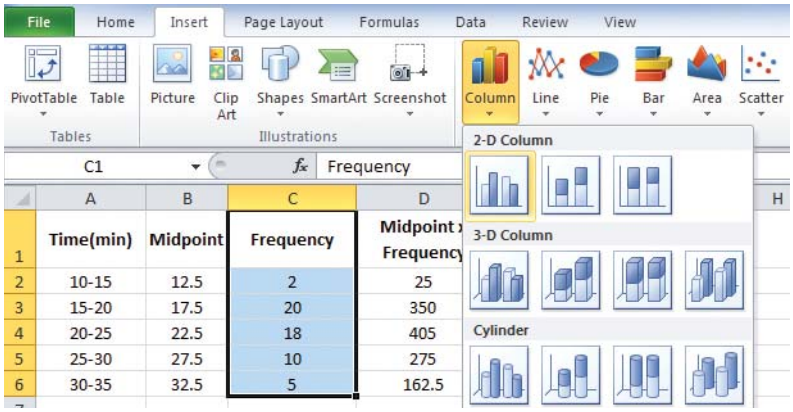


	A	B	C	D
1	Time(min)	Midpoint	Frequency	Midpoint x Frequency
2	10-15	12.5	2	25
3	15-20	17.5	20	350
4	20-25	22.5	18	405
5	25-30	27.5	10	275
6	30-35	32.5	5	162.5
7				
8			Grouped Mean	=SUM(D2:D6)/
9				SUM(C2:C6)

Determine the median and mode by inspection. Since there are 55 data values in this example, the median will be the $\frac{55+1}{2} = 28$th entry, which occurs in the 20–25 min time interval. Its midpoint is 22.5 min. The modal interval is the most frequent interval, which occurs in the 15–20 min time interval.

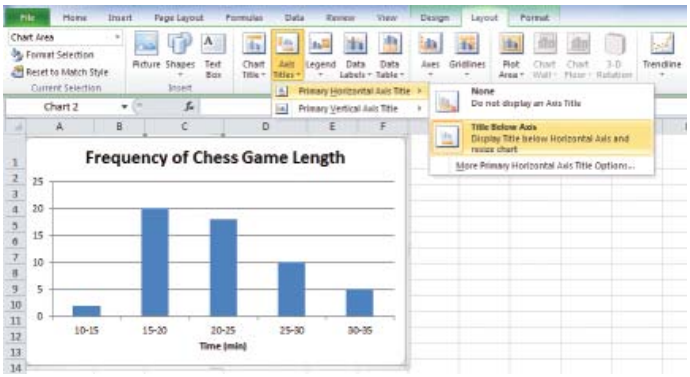
b) Since grouped data use the midpoint of the interval, your calculations could be less accurate when the interval size is quite large. Also, the actual data values in each interval could lie anywhere within the interval. So, the actual values may be closer to the boundaries of the interval and using the midpoint could provide inaccurate results.

c) Enter the data into your spreadsheet as shown. Highlight the frequency column, select the **Insert** tab, and then choose **Column, Clustered Column** from the drop-down menu.



Right click on the x-axis and choose **Select Data...**. Under the options for **Horizontal (Category) Axis Labels**, select **Edit** and highlight the data for your time intervals from A2:A6.

Change the title of your graph to **Frequency of Chess Game Length**. With your graph selected, click on the **Layout** tab and choose **Axis Titles**, then choose **Primary Horizontal Axis Title – Title Below Axis**. Rename the axis title **Time (min)**.



To eliminate the gap between bars, right click on any of the bars and choose Format Data Series. Reduce the Gap Width to 0% and click Close.

d) Since the data are positively skewed, the modal interval is the least appropriate measure of central tendency. The median and mean are very close together, so either one would be an appropriate measure.

Your Turn

A group of children were asked how many hours a day they spend playing video games. The table shows the data.

Number of Hours	Number of Children
0–2	3
2–4	11
4–6	7
6–8	2
8–10	1

a) Determine the estimated mean, median number of hours, and modal interval for the above distribution.

b) Discuss any skewing of the data with respect to the measures of central tendency.

Certain values in a data set are sometimes of greater relative importance than others. In these cases, it is useful to calculate a weighted mean. To do this, multiply the weighting by the corresponding data value, find the sum of these products, and then divide by the total weighting.

Weighted Mean

$$\mu = \frac{\sum x_i w_i}{\sum w_i}$$

Where x_i represents each data value in the data set and w_i represents its weight or frequency.

Example 3

Using a Weighted Mean

A teacher is calculating the marks for the students in her Data Management class. She assigns the following values to each category:

Knowledge: 25%	Thinking: 10%
Application: 20%	Culminating Project: 15%
Communication: 15%	Final Exam: 15%

Kyle has not yet written his final exam, but his marks in the first five categories are 90, 79, 82, 70, and 85.

a) Determine the weighted mean for Kyle before the final exam.

b) How does this weighted mean differ from the unweighted mean?

c) What mark must Kyle receive on the final exam to finish the course with 84%?

Solution

a) Calculate the weighted mean:

$$\mu_w = \frac{90(0.25) + 79(0.20) + 82(0.15) + 70(0.10) + 85(0.15)}{0.85}$$
$$= \frac{70.35}{0.85}$$
$$\approx 82.76$$

b) Calculate the unweighted mean:

$$\mu = \frac{90 + 79 + 82 + 70 + 85}{5}$$
$$= \frac{406}{5}$$
$$= 81.2$$

Without the weighted mean, Kyle would receive a lower average mark because the categories he excels in are worth more when calculated using the weighted mean.

c) The final exam is worth 15%. Determine the final exam score, E, needed for Kyle to receive 84% in the course.

$$84 = \frac{90(0.25) + 79(0.20) + 82(0.15) + 70(0.10) + 85(0.15) + E(0.15)}{1.00}$$

$$84 = 70.35 + E(0.15)$$

$$84 - 70.35 = 0.15E$$

$$13.65 = 0.15E$$

$$91 = E$$

Kyle must receive 91% on his final exam to finish the course with an 84%.

Your Turn

A math department assigns the following weights for each category in its Advanced Functions course:

Knowledge: 25%	Thinking: 10%
Application: 15%	Culminating Project: 10%
Communication: 10%	Final Exam: 30%

Catherine's marks in the course so far are 87, 90, 76, 78, and 84 in each of the first five categories. She still needs to write the final exam.

a) Determine the weighted mean for Catherine before writing her final exam.

b) Is it possible for Catherine to receive a final mark of 90% in the course? Justify your answer.

Consolidate and Debrief

Key Concepts

- Three measures of central tendency are mean, median, and mode.
- The mean represents the average of a set of data.
- The median is the middle number when the numbers are arranged in numerical order.
- The mode is the number that occurs most often; it is possible to have one, more than one, or no mode.
- Outliers have a greater effect on the mean than other measures and either pull the mean up or drag the mean down.
- A weighted mean accounts for the relative importance of each value in the average.
- Grouped data are organized into intervals. Use the interval midpoints and frequencies to estimate the measures of central tendency.

Reflect

R1. Which measure of central tendency is most affected by extreme values? Explain using specific examples to justify your answer.

R2. Describe a situation in which it would be necessary for you to use

- **a)** the mean
- **b)** the weighted mean
- **c)** grouped data

R3. Which measure of central tendency is being used in each situation? Explain.

- **a)** The average person has two hands, two eyes, two ears, and two legs.
- **b)** The average time it takes to get to school is 38 min.
- **c)** Johnny is an above average student.

Practise

Choose the best answer for #3 and #4.

1. Determine the mean, median, and mode for each set of data.

- **a)** 4 6 9 12 15 7 13 4 7 10 3 8 15
- **b)** 9 8 20 23 12 12 9 9 12 9 20 21 9
- **c)** 110 152 112 124 110 134 138 127 118 110 114 162

2. Nina runs the 400-m race for Mustang High School. Her times in the last six track meets were 1.45 min, 1.50 min, 1.42 min, 1.41 min, 1.42 min, and 1.48 min.

- **a)** What are the mean and median for her running times?
- **b)** Which measure of central tendency best describes Nina's average time? Explain.

3. The observation that occurs most frequently in a data sample is the

- **A** mean
- **B** weighted mean
- **C** mode
- **D** median

4. What is the median of the sample 5, 5, 11, 9, 8, 5, 8?

A 9 **B** 6
C 5 **D** 8

Apply

5. The mean of Daniel's marks on five tests was 77.4. His marks on the first four tests were 88, 77, 70, and 72. Calculate Daniel's mark on the fifth test.

6. The average daily snowfall for the first week of December was 2.5 cm. In each of the first two days, 2.5 cm fell. In each of the next four days, 2 cm fell. What was the snowfall for the last day of the week?

7. **Communication** Determine whether the argument is valid for each situation. Explain your thinking.

- **a)** An advertising company has a mean monthly sales record of $16 235. Therefore, half the team members sold more than $16 235.
- **b)** A survey shows that 78% of all salaries are below the mean. Therefore, there must be a mistake.
- **c)** The mean mark of one class is 71, while the mean mark of another class is 76. Therefore, the mean of the two classes is 73.5.
- **d)** My median monthly expenses total $850. Therefore, my total expenses for the year must be $10 200.

8. **Communication** Which measure of central tendency would be best suited for each situation? Explain why you chose the measure that you did.

a) a summary of a class's report card marks

b) an award for the most popular movie of the year

c) an employer budgeting for the average salary of its employees

d) a potential employee looking for the typical salary among current employees

9. **Thinking** Create a data set of at least seven values that satisfies each of the conditions. Use the context of marks, salaries, sports statistics, or choose a context of your own.

a) The mean, median, and mode are all 15.

b) The median is 7.5 and the mean is greater than 15.

c) The mean is 7.5 and the median is greater than 15.

d) Explain why the mean is more affected by outliers than the median.

10. Michael surveyed the grade 12 students at his school to research the number of hours of sleep they got. He asked them how many hours of sleep they got last night. The table shows his results.

Time (h)	4–5	5–6	6–7	7–8	8–9
Frequency	32	50	125	67	108

a) Make a histogram of these data.

b) Estimate the mean, median, and modal interval for the hours of sleep by grade 12 students.

c) Mark the measures of central tendency on your histogram.

d) Discuss any skewing of the data and how it relates to the measures of central tendency.

11. **Application** Your teacher will provide you with a file called **Nobel Winners.csv**, listing the Canadian or Canadian-born Nobel Prize winners up to 2013. Use appropriate technology to answer the questions.

a) Make a histogram of the winners' ages.

b) Calculate the mean, median, and mode ages.

c) Describe the "average" age of a Canadian Nobel Prize winner. Explain why this age would not be younger.

12. **Application** Your teacher will provide you with a file called **Olympics 2014.csv**, listing the medal counts and populations of winning countries in the 2014 Sochi Winter Olympics. The file ranks countries by number of gold medals, followed by silver and bronze. Use appropriate technology to answer the questions.

a) Re-rank the countries by total medals relative to the population.

b) Re-rank the countries using a weighted mean, with each type of medal having a different weighting.

c) Which system do you prefer? Write a paragraph supporting your choice.

13. Karen's term mark is 82%. The term counts for 70% of the final mark. What mark must Karen achieve on the exam to earn a final mark of

a) 80%? **b)** 85%?

c) at least 75%?

d) Can Karen achieve 88%? Explain.

14. **Thinking** Using the data provided in the frequency table, describe a context that the mean, median, and mode could represent.

Age in Years	20–30	30–40	40–50	50–60	60–70
Frequency	14	15	28	19	5

Achievement Check

15. The table shows student absences from Lakeside High School during the first semester. Assume that the absences located exactly on the endpoints of an interval were placed in the lower interval.

Student Absences	Number of Students
0–3	47
3–6	89
6–9	33
9–12	102
12–15	24
15–18	19
18–21	6
21–24	8
24–27	0
27–30	2

a) Calculate the estimated mean, median, and modal interval of student absences.

b) Does there appear to be an outlier? If so, how does it affect the mean and median of the data set?

c) Which would be the most reliable measure of central tendency if you were trying to make a generalization to someone about the data presented? Justify your response with calculations.

16. On a fishing trip with his father, Alex caught eight bass with a mean mass of 1.2 kg and five trout with a mean mass of 2.9 kg. What was the mean mass, in kilograms, of all the fish Alex caught?

17. a) Create a frequency distribution table of the number of vowels and consonants found in the names of the students in your class.

b) Make appropriate graphs of the data.

c) Calculate the measures of central tendency of each type of letter and mark them on your histograms.

d) Decide which measure of central tendency would be best to make an assumption based on the data presented.

e) Summarize your findings.

Extend

18. A trimmed mean removes a small percent of the largest and smallest values before calculating the mean. This is to reduce the effects of outliers. At a diving competition, the marks for Competitor A were 8.7, 8.9, 8.1, 8.6, 8.5, 8.8, and 8.0. The marks for Competitor B were 8.4, 8.6, 8.6, 8.5, 8.5, 8.4, and 8.9.

a) Using the marks as given, which competitor would have the higher mean mark?

b) To reduce the influence of biased judging, the highest and lowest marks are deleted for all competitors. Which competitor has the higher average mark under this system?

19. The harmonic mean is defined as $\frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\frac{1}{x_3}+\cdots+\frac{1}{x_n}}$, where n is the number of values in the set of data. It is also defined as the reciprocal of the mean of the reciprocals.

a) Find the harmonic mean of the numbers 1, 4, and 7.

b) Find the mean of the reciprocals of 1, 4, and 7. What is the reciprocal of the result?

c) The harmonic mean can be used to find the mean of a set of rates. The harmonic mean will give you the average price between two rates. What is the average price between \$1.25/kg and \$1.38/kg?

20. The geometric mean is defined as $\sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdots x_n}$, where n is the number of values in the set of data.

a) Find the geometric mean of the numbers 2, 8, and 9.

b) The geometric mean can be used to calculate the average annual rate of change when rates are compounded on each other. Calculate the average annual rate of change if inflation were at 2%, then 4%, then 3%, respectively, over a three-year period.

6.2

Measures of Spread

Learning Goals

I am learning to

- describe the variability in a sample or population using measures of spread
- calculate the range
- understand how to use quartiles and percentiles to analyse data

Minds On...

The fuel economy of various cars in a given class will vary significantly. Cars can be rated anywhere from 4 L/100 km to 9 L/100 km. Understanding where most of the data occur using the measures of central tendency is sometimes not enough to make a broad generalization of the data. Why is it also important to know the distance associated with each data value and the centre of the data? The measure of how widely the data vary around their central location is referred to as the measure of spread or dispersion.

Action!

Investigate Percentiles

An automotive publication ranked the following 2013 subcompact cars based on their fuel economy in the city. The chart shows the cars' fuel economy in the city and on the highway. Your teacher will provide you with the data in a file called **Fuel Economy.csv**.

Make/Model	L/100 km	
	City	Hwy
TOYOTA PRIUS C	3.5	4.0
SCION iQ	5.5	4.7
SMART FORTWO	5.8	4.2
CHEVROLET SPARK	6.3	5.1
FIAT 500	6.4	4.9
TOYOTA YARIS	6.7	5.5
MINI COOPER	6.8	5.2
KIA RIO	6.8	5.4
FORD FIESTA	6.9	5.1
HONDA FIT	7.1	5.4
MAZDA 2	7.1	5.8
HYUNDAI ACCENT	7.2	5.2
SCION xD	7.4	5.9
NISSAN VERSA	7.4	5.4
CHEVROLET SONIC	7.7	5.5
VOLKSWAGEN BEETLE	9.5	7.1
SUBARU BRZ	9.6	6.6

Source: "Fuel Consumption Ratings," Natural Resources Canada, February 5, 2014.

1. **a)** Identify the car with the median city fuel economy in city driving.
   b) The median is also called the 50th **percentile**. Why would that be?
2. The 80th percentile would be the car whose fuel economy in city driving is greater than 80% of all others. Which car would that be?
3. What percent of the cars have a better fuel economy than a Fiat?
4. **a)** Identify the car with the best fuel economy in the city and the car with the best fuel economy on the highway. What percent of the cars have a worse fuel economy?
   b) Identify the car with the worst fuel economy in the city and the car with the worst fuel economy on the highway. What percent of the cars have a worse fuel economy?
5. **Reflect** Describe how you can use percentiles to compare data.
6. **Extend Your Understanding** If you were comparing your marks to those of your classmates, would you rather be in the 95th percentile or the 5th percentile? Explain.

percentile

- the percent of all the data that are less than or equal to a specific data value

To help analyse the spread of data, you may need to identify the percentile rank or calculate percentiles.

Percentile Rank

$$R = \frac{p}{100}(n + 1)$$

where p is the percentile, n is the size of the population, and R is the whole number rank of the data point. If R is not a whole number, round R down.

Percentile

$$p = 100\frac{(L + 0.5E)}{n}$$

where p is the percentile, L is the number of data less than the data point, E is the number of data equal to the data point, and n is the size of the population.

Example 1

Percentiles

The list shows the marks for 25 students on a recent test out of 40.

31 28 28 30 20 25 38 40 26 28 15 21 28
36 25 16 21 34 37 30 23 24 36 32 25

a) Calculate the 80th percentile.

b) What percentile is a mark of 25?

c) What percentile is a mark of 40?

Solution

Order the data from least to greatest.

15 16 20 21 21 23 24 25 25 25 26 28 28
28 28 30 30 31 32 34 36 36 37 38 40

a) There are 25 data values. To calculate the 80th percentile, use the formula for percentile rank.

$$R = \frac{p}{100}(n + 1)$$
$$= \frac{80}{100}(25 + 1)$$
$$= 20.8$$

Round down to 20. Determine the midpoint of the 20th and 21st measurements.

Why do you round down to 20?

$$\text{80th percentile} = \frac{34 + 36}{2}$$
$$= 35$$

The 80th percentile is a mark of 35. This means that 80% of the data are below 35.

b) A mark of 25 is the 8th ranked mark. It is also the 9th and 10th marks.

There are 7 data values less than 25, so $L = 7$.

There are 3 data values equal to 25, so $E = 3$.

$$p = 100\frac{(L + 0.5E)}{n}$$
$$= 100\frac{(7 + 0.5(3))}{25}$$
$$= 34$$

A mark of 25 is in the 34th percentile. This means that 34% of the data are below 25.

c) A mark of 40 is the 25th ranked mark.

There are 24 data values less than 40, so $L = 24$.

There is 1 data value equal to 40, so $E = 1$.

$$p = 100\frac{(L + 0.5E)}{n}$$
$$= 100\frac{(25 + 0.5(1))}{25}$$
$$= \frac{100(24.5)}{25}$$
$$= 98$$

A mark of 40 is in the 98th percentile.

Your Turn

The mean playing times per game for the 22 hockey players on a team are given.

16.4, 18.3, 21.7, 18.5, 9.2, 17.9, 12.0, 15.2, 23.4, 20.5, 16.7, 13.4, 8.3, 17.9, 22.6, 18.1, 21.7, 14.6, 13.8, 24.3, 12.4, 17.4

a) Determine the 40th and 95th percentiles.

b) Determine the percentile rank of the player who averaged

i) 9.2 min per game

ii) 21.7 min per game

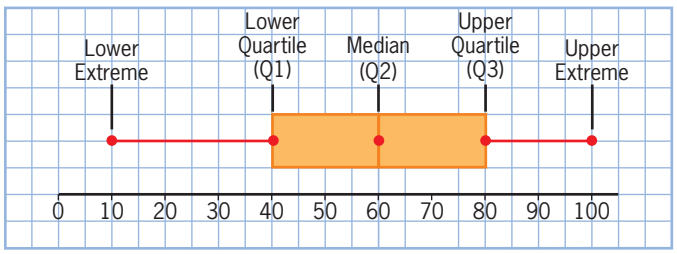
iii) 18.1 min per game

To better understand the variability of a data set, you can use a variety of measures of spread. You can use a box and whisker plot to visually demonstrate the spread of a distribution along a number line.

To construct a box and whisker plot:

- Draw a rectangle whose ends are the first (lower) and third (upper) **quartiles**.
- Draw the median within the rectangle.
- Add "whiskers," which are horizontal line segments connecting the box to the extremes of the data, covering the entire **range**.

Each of the four zones illustrated by a box and whisker plot contains 25% of the data. The difference between the first and the third quartiles is known as the **interquartile range (IQR)**. The interquartile range represents the "middle half" of the data.



Because of the nature of the calculations, box and whisker plots are appropriate only for quantitative data.

quartiles

- three points that divide the data set into four equal groups
- the first quartile (Q1) is the middle number between the smallest number and the median; it is also the 25th percentile
- the second quartile (Q2) is the median of the data set; it is also the 50th percentile
- the third quartile (Q3) is the middle number between the median and the largest number in a data set; it is also the 75th percentile

range

- the difference between the highest value and the lowest value of a data set
- range = highest value – lowest value

interquartile range (IQR)

- the difference between the first and third quartiles
- IQR = Q3 – Q1

Example 2

Interquartile Range and Box and Whisker Plots

The table lists the heights of the 20 girls who signed up to try out for their school basketball team.

Height (cm)	Frequency	Cumulative Frequency
155–160	1	1
160–165	3	4
165–170	4	8
170–175	7	15
175–180	3	18
180–185	1	19
185–190	0	19
190–195	1	20

a) Determine the median, range, first and third quartiles, and interquartile range. Create a box and whisker plot of the data.

b) Describe the data in each zone of the plot.

c) Identify any outliers, if they exist.

Solution

a) Method 1: Use Paper and Pencil

The median is the 50th percentile. Calculate the percentile rank.

$$R = \frac{p}{100}(n + 1)$$
$$= \frac{50}{100}(20 + 1)$$
$$= 10.5$$

The median is the midpoint of the 10th and 11th measurements.
The median lies within the 170–175 cm interval.
The median height is 172.5 cm.

The range is the difference between the highest and lowest values.

$$\text{Range} = 195 - 155$$
$$= 40$$

The range is 40 cm.

The first quartile is the 25th percentile. Use the percentile rank formula.

$$R = \frac{p}{100}(n + 1)$$

$$\text{Q1} = \frac{25}{100}(20 + 1)$$
$$= 5.25$$

Q1 is the midpoint of the 5th and 6th measurement. Looking at the cumulative frequency column, we can see that this lies within the 165–170 cm interval.

Q1 is 167.5 cm.

The third quartile is the 75th percentile. Use the percentile rank formula.

$$R = \frac{p}{100}(n + 1)$$

$$\text{Q3} = \frac{75}{100}(20 + 1)$$

$$= 15.75$$

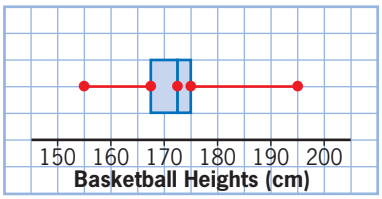
Q3 is the midpoint of the 15th and 16th measurements. Looking at the cumulative frequency column, we can see that this lies on the border between the 170–175 cm and 175–180 cm intervals.

Q3 is 175 cm.

Interquartile range = 175 – 167.5

= 7.5

The interquartile range is 7.5 cm.



Method 2: Use Fathom™

Begin a new collection and name it **Basketball Heights**. Open a table and name the first attribute **Height_cm**. Enter all 20 values from the table, using the midpoint of each interval, and based on the individual frequencies. For example, in the 2nd interval, enter 162.5 a total of 3 times.

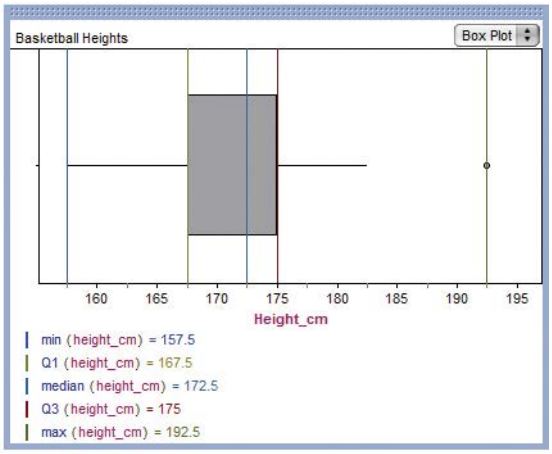
Under the **Graph** menu, select **Graph**. From the upper right, select **Box Plot**.

Making sure the Graph window is selected, under the **Graph** menu, select **Plot Value**. When the calculator window pops up select **Statistical**, then from **One Attribute** select **Median**. Then, from the **Attributes** category, select **Height_cm**. Press **OK**. Repeat the process for Q1, Q3, Max, and Min.

Because Fathom™ only works with raw data, not intervals or grouped data, use the midpoint of the interval for each data point in the interval.

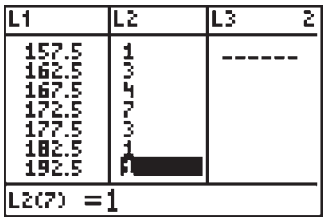
Basketball Heights

	Height_...	<new>
1	157.5	
2	162.5	
3	162.5	
4	162.5	
5	167.5	
6	167.5	
7	167.5	
8	167.5	
9	172.5	
10	172.5	
11	172.5	
12	172.5	
13	172.5	
14	172.5	
15	172.5	
16	177.5	
17	177.5	
18	177.5	
19	182.5	
20	192.5	



Method 3: Use a Graphing Calculator

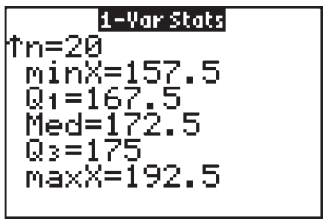
Enter the midpoints in list **L1** and the frequencies in list **L2**.



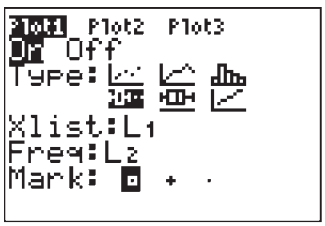
Press STAT, select **CALC**, and select **1:1-Var Stats**. Press ENTER.

Now you need to tell the calculator that the data are in **L1** and the frequency is in **L2**. Select 2ND L1 , 2ND L2 and press ENTER.

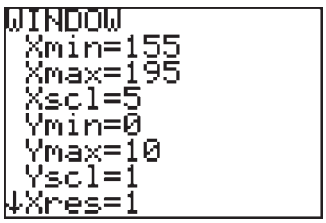
Scroll down to see the quartiles.



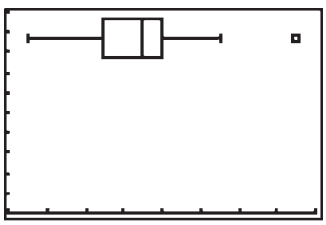
Set up Plot1 as shown.



Use the following WINDOW settings.



Press GRAPH.



Press TRACE and use the arrow keys to scroll through the values on the graph.

b) 25% of the data are contained in each of the intervals 155 to 167.5 cm, 167.5 to 172.5 cm, 172.5 to 175 cm, and 175 to 195 cm.

c) An outlier is identified as being more than 1.5 times the interquartile range (IQR) below Q1 or above Q3.

Lower Extreme:

$$\begin{aligned} &Q1 - 1.5(IQR) \\ &= 167.5 - 1.5(7.5) \\ &= 156.25 \end{aligned}$$

Upper Extreme:

$$\begin{aligned} &Q3 + 1.5(IQR) \\ &= 175 + 1.5(7.5) \\ &= 186.25 \end{aligned}$$

No data point is less than 156.25 cm, but one is greater than 186.25 cm. Therefore, one outlier of approximately 192.5 cm exists in this data set.

Your Turn

A summer camp activity involves measuring the distance travelled by 50 turtles in 15 min. The table shows the results.

Distance (m)	Frequency
0–5	1
5–10	0
10–15	6
15–20	12
20–25	15
25–30	5
30–35	7
35–40	1
40–45	3

a) Determine the median, range, first and third quartiles, and interquartile range. Make a box and whisker plot of the data.

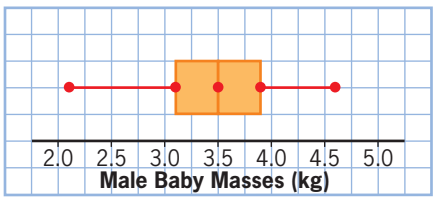
b) Describe the data in each zone of the plot.

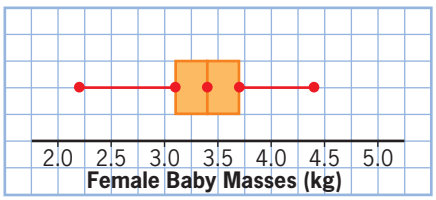
c) Identify any outliers, if they exist.

Example 3

Interpreting Quartiles

The box and whisker plots illustrate the spread of Canadian full-term male and female baby masses, in kilograms, at birth.





Compare the spreads of birth masses for boys and girls.

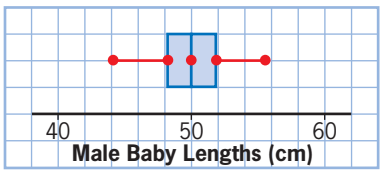
Solution

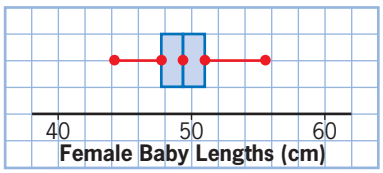
	Boys (kg)	Girls (kg)
Median	3.5	3.4
Range	4.6 – 2.1 = 2.5	4.4 – 2.2 = 2.2
Q1	3.1	3.1
Q3	3.9	3.7
IQR	3.9 – 3.1 = 0.8	3.7 – 3.1 = 0.6

The median birth mass for boys is 0.1 kg greater than the median birth mass for girls. The middle 50% of the birth masses for boys lie between 3.1 kg and 3.9 kg, for an IQR of 0.8 kg. The middle 50% of the birth masses for girls lie between 3.1 kg and 3.7 kg, for an IQR of 0.6 kg. Both the range and the IQR for boys are greater than for girls. So, the birth masses for boys are more spread out.

Your Turn

The box and whisker plots illustrate the spread of Canadian full-term male and female baby lengths, in centimetres, at birth. Compare the spreads of birth lengths for boys and girls.





Consolidate and Debrief

Key Concepts

- A measure of spread helps you understand how closely a set of data is clustered around its centre.
- The range is the difference between the maximum value and minimum value.
- A percentile is the percent of all the data that are less than or equal to the specific data point.
- Quartiles divide the data set into four equal parts. Q1 is the 25th percentile, Q2 is the median (or 50th) percentile, and Q3 is the 75th percentile.
- The interquartile range (IQR) is the distance between the first and third quartiles. To calculate, subtract the value for Q1 from the value for Q3. The interquartile range contains the middle 50% of the data.
- A box and whisker plot uses a rectangle to visually demonstrate the spread of the distribution along a number line by displaying the median, quartiles, and upper and lower extremes.
- An outlier exists if it is less than Q1 – 1.5 × IQR or greater than Q3 + 1.5 × IQR.

Reflect

R1. Consider the information in the table regarding two data sets. Compare their spreads.

	Data Set 1	Data Set 2
Median	56.3	57.1
Min	32.1	24.2
Max	65.9	71.1
Q1	43.2	34.5
Q3	60.2	63.2

R2. What problems can occur if the range is used to measure the spread of a set of data?

R3. What information does the interquartile range provide?

Practise

Choose the best answer for #4 and #5.

1. Calculate the percentile rank for each student's quiz result in a grade 12 math class.

Mark	Frequency
4.0	2
5.0	6
6.0	8
7.0	13
8.0	4
9.0	3
10.0	2

2. What is the range of the data set?

23 56 45 65 59
55 62 54 85 25

3. Calculate the median, range, Q1, Q3, and interquartile range of each set of data. Identify any outliers.

a) 39 51 35 22 28 67 57
42 56 74 51 87 99 48
36 28 57 23 53 74

b) 245 264 222 213 243
215 467 264 276 199
127 216 233

c) 5 7 9 4 5 7 3 2 6
2 8 5 9 1 3 3 0 4
6 8 5 9 2 3 4 5 0
4 2 4 6 2 5 6

d) 6213 7985 3426 5134
7659 3462 5348 6213
5968 7659 3567

4. If each number in a set is increased by two, which of the measures of spread would remain unchanged?

A the range
B the interquartile range
C the percentiles
D all of the above

5. Which is an incorrect statement about the interquartile range?

A It contains the middle 50% of the data.
B An outlier lies more than 1.5 times the IQR from Q1 or Q3.
C To calculate the interquartile range, subtract Q3 – Q1.
D The median always lies at the middle of the interquartile range.

Apply

Use the table for #6 to #8.

The infant mortality rate represents the number of children, per thousand, who die before the age of one year.

Infant Mortality Rates by Province and Territory					
	2007	2008	2009	2010	2011
Newfoundland and Labrador	7.5	5.1	6.3	5.3	6.3
Prince Edward Island	5.0	2.0	3.4	3.6	4.2
Nova Scotia	3.3	3.5	3.4	4.6	4.9
New Brunswick	4.3	3.2	5.8	3.4	3.5
Québec	4.5	4.3	4.4	5.0	4.3
Ontario	5.2	5.3	5.0	5.0	4.6
Manitoba	7.3	6.5	6.3	6.7	7.7
Saskatchewan	5.8	6.2	6.7	5.9	6.7
Alberta	6.0	6.2	5.5	5.9	5.3
British Columbia	4.0	3.7	3.6	3.8	3.8
Yukon	8.5	5.4	7.8	5.2	0.0
Northwest Territories	4.1	9.7	15.5	1.4	7.2
Nunavut	15.1	16.1	14.8	14.5	26.3
Canada	5.1	5.1	4.9	5.0	4.8

Source: Infant mortality rates, by province and territory (both sexes), Statistics Canada.

6. a) Rank the provinces in ascending order by their 2011 infant mortality rates.

b) Determine the percentile ranks for five provinces or territories of your choice.

7. a) Determine the median and interquartile range for the infant mortality rate in each year.

b) Compare these measures across the years.

c) Why would the medians not be the same as the mortality rate for all of Canada?

8. a) Are the 2011 mortality rates for Yukon and Nunavut outliers?

b) Explain the variability of the mortality rates in Yukon, Northwest Territories, and Nunavut as compared to Ontario.

9. Application The table shows the size of each age group in Canada in 2009 and 2013.

Population of Canada by Age Group		
Age Group (years)	2009	2013
All Ages	33 628 571	35 158 304
0 to 10	3 626 272	3 804 924
10 to 20	4 253 528	4 048 205
20 to 30	4 608 623	4 855 939
30 to 40	4 534 301	4 762 084
40 to 50	5 251 373	4 940 356
50 to 60	4 798 598	5 256 870
60 to 70	3 299 618	3 857 403
70 to 80	1 994 853	2 202 364
80 to 90	1 075 522	1 181 124
90 to 100	180 409	242 124
100 or over	5 474	6 911

Source: Estimates of population, by age group and sex for July 1, Canada, provinces and territories, annual, Statistics Canada.

a) Rank the age groups in ascending order by size for each year. Calculate the percentile rank for three different age groups in each year.

b) Describe the changes in population breakdown from 2009 to 2013.

10. Communication The table shows the average net worth of Canadian families, as a percent of total, in 1999 and 2005.

	Net Worth (% of total)	
Quintile	1999	2005
1st	0.1	0.1
2nd	2.5	2.3
3rd	8.8	8.4
4th	20.1	20.2
5th	68.5	69.2

Source: Drummond, Don and David Tulk, "Lifestyles of the Rich and Unequal: An Investigation Into Wealth Inequality in Canada," TD Economics Special Report, December 13, 2006.

a) Describe what is meant by quintile.

b) Describe the change in distribution of incomes from 1999 to 2005.

Achievement Check

11. A consumers group recently tested 100 compact fluorescent light bulbs and recorded their lifetimes. The chart shows the results.

Lifetime (h)	Frequency
5000–6000	7
6000–7000	12
7000–8000	34
8000–9000	27
9000–10 000	10
10 000–11 000	8
11 000–12 000	2

a) Make a box and whisker plot of the data.

b) Make a histogram of the data and mark the quartiles on it.

c) Describe the middle 50% of the data.

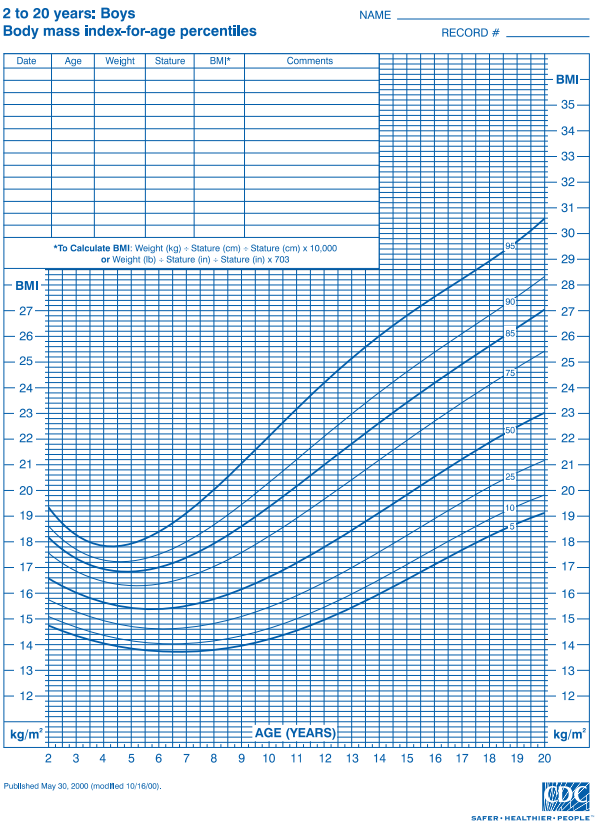
d) Identify any outliers.

12. Agencies track airline flight delays to help consumers compare airlines. The table outlines the number of delayed flights per month for one major airline. Determine whether December's results are an outlier. If so, what might have caused it?

Month	Number of Delayed Flights
Jan	288
Feb	295
Mar	274
Apr	280
May	246
Jun	251
Jul	218
Aug	221
Sep	246
Oct	264
Nov	257
Dec	459

Extend

13. Your teacher will provide you with a file that shows the percentiles for body mass index of boys ages 2 to 20.



Source: National Center for Health Statistics and National Center for Chronic Disease Prevention and Health Promotion, "2 to 20 years: Boys Body mass index–for–age–percentiles," CDC, May 30, 2000.

a) Describe how to use this chart.

b) Make an accompanying table listing the body mass index values for ages 2, 10, and 20, organized by percentiles.

14. An alternate method of calculating Q1 and Q3 is to use the midpoints of the median and the minimum value and the median and the maximum value, respectively. Use any set of data in the questions above to verify that this method works.

6.3

Standard Deviation and z-Scores

Learning Goals

I am learning to

- use technology to calculate the variance and standard deviation of a data set
- calculate and understand the significance of a *z*-score
- relate the positive or negative scores to their locations in a histogram
- develop significant conclusions about a data set

Minds On...

The ozone layer protects Earth's surface from much of the Sun's destructive radiation. Unfortunately, the ozone layer is being destroyed, in part by chlorofluorocarbons such as coolants in old refrigerators. The ozone layer's thickness can vary significantly over periods as short as a week. What other parts of our environment are changing due to pollutants?

Action!

Investigate Standard Deviation

Literacy Link

A *Dobson unit (DU)* measures the thickness of the ozone layer. It is equivalent to the number of molecules of ozone that would be required to create a layer of pure ozone 0.01 mm thick at a temperature of 0 °C and a pressure of 1 atmosphere.

The table shows the thickness of the ozone layer on each day of a given week.

Day	Thickness, x (DU)	Deviation from the Mean $x - \bar{x}$	Squared Deviation $(x - \bar{x})^2$
1	152		
2	158		
3	151		
4	153		
5	159		
6	158		
7	152		
Sum			

1. Calculate the mean thickness, $\bar{x}$.

2. **a)** Calculate the deviation from the mean for each day. Enter the results in the third column.
   b) Enter the sum at the bottom of the column.
   c) Explain the resulting sum.

3. **a)** Calculate the squares of the deviations from the mean. Enter the results in the fourth column.
   b) Enter the sum at the bottom of the column.

4. Divide the sum of the squares by 7. This is called the **variance**.

5. Take the square root of the variance. This is called the **standard deviation**.

6. **Reflect** The standard deviation is the average difference of all the measurements from the mean. What other formula does this resemble?

7. **Extend Your Understanding** For the previous week, the mean thickness was 158.2 DU, with a standard deviation of 4.8 DU. Compare these two weeks' measurements.

Standard deviation is used more commonly than variance as a measure of spread because it is expressed in the same units of measure as the data, whereas variance is expressed in square units.

variance

- the average squared difference of the scores from the mean

standard deviation

- the square root of the variance
- the average distance of the scores from the mean

The variance and standard deviation of a data set allow you to determine how close the values in a distribution are to the middle of the distribution. You can calculate the variance and standard deviation of a data set using the following formulas.

Population Variance

$$\sigma^2 = \frac{\sum(x - \mu)^2}{N}$$

Sample Variance

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum(x - \mu)^2}{N}}$$

Sample Standard Deviation

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

where the population deviation is represented by $(x - \mu)$ and the sample deviation is represented by $(x - \bar{x})$.

The Greek letter σ (lower case), pronounced "sigma," is used to represent the population standard deviation, while the letter s is used to represent the sample standard deviation. Remember that capital sigma, Σ, is used to denote a sum.

Samples rarely contain extreme values when compared to entire populations. As a result, the variance and standard deviation are less than would be expected. To use the sample variance and standard deviation to model a population, divide by $n - 1$ instead of n. This slightly increases their values.

Calculating the variance alone is not a perfect measure of spread. First, because the deviations of each value are squared in the formula, more "weight" is given to extreme values. Therefore, data sets with extreme values or outliers will skew the validity of the result. Second, the variance is calculated in units squared, which is not the same units as the scores in the data set. This means that you cannot show the variance on a frequency distribution and cannot make a direct correlation between its value and the values of your data set. This problem is easily corrected by calculating the standard deviation.

Example 1

Visualizing the Spread of Marks

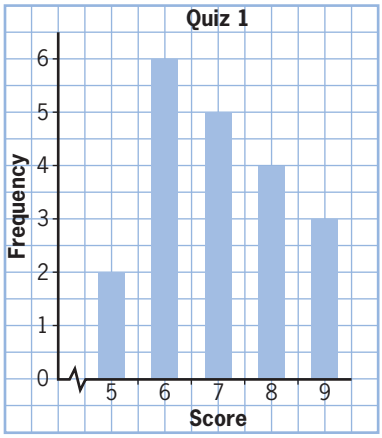
The graphs represent the scores on two quizzes. The mean score for each quiz is 7.0.

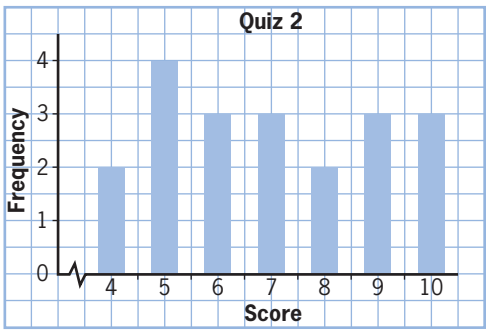
a) Which quiz would have a greater standard deviation? Why?

b) The variance of Quiz 1 is 1.5. What is the standard deviation?

c) What would the Quiz 1 graph look like if the standard deviation were 1.6?

d) What would the graph look like if the standard deviation were 0?





Solution

a) As can be seen on the graphs, Quiz 2 is more spread out than Quiz 1. As a result, its standard deviation will be greater.

b) The standard deviation is the square root of the variance.

$\sigma = \sqrt{1.5}$

≈ 1.22

c) Since $1.6 > 1.22$, a standard deviation of 1.6 would result in a wider spread, such as that in Quiz 2.

d) A standard deviation of 0 means there is no spread. All data would consist of the same value.

Your Turn

Sketch examples of two histograms that show the distribution of two sets of girls' heights with the same mean but with different standard deviations. Indicate which histogram will have a greater standard deviation.

Example 2

Calculating Variance and Standard Deviation

The ages of participants in a school's talent contest are listed below.

16 17 18 16 15 16 17 15 18 14
17 19 18 16 17 17 17 14 15 18

Use technology to answer the questions.

a) Plot a histogram of the data.

b) Calculate the mean and standard deviation.

c) What would happen to the standard deviation if the first person's age were 18?

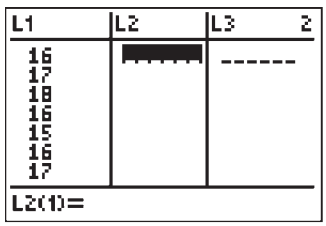
d) What would happen to the standard deviation if the second person's age were 16 instead of 17?

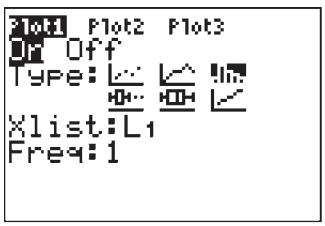
e) What would happen to the standard deviation if each person were one year older?

f) Which ages are more than one standard deviation from the mean?

Solution

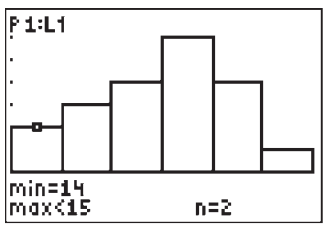
a) and **b)** Enter the data into list **L1**. Set up Plot1 as shown.

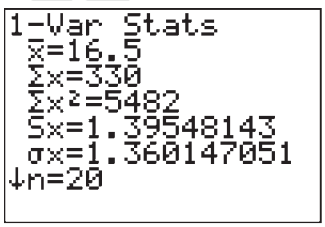




Press ZOOM and then select **9:ZoomStat**.

Press TRACE to view the histogram. Scroll left and right to view the coordinates of each bar. Press STAT CALC and select **1:1-Var Stats**.





The mean, $\overline{x}$, is 16.5 years and the standard deviation, σ, is approximately 1.360 years.

c) $\sigma = 1.429$ years

The standard deviation would increase because age 18 is farther from the mean and the spread of the data would increase.

Population formulas were used here because all participants were included.

d) $\sigma = 1.359\ 23$ years

The standard deviation would decrease because age 16 is closer to the mean and the spread of the data would decrease.

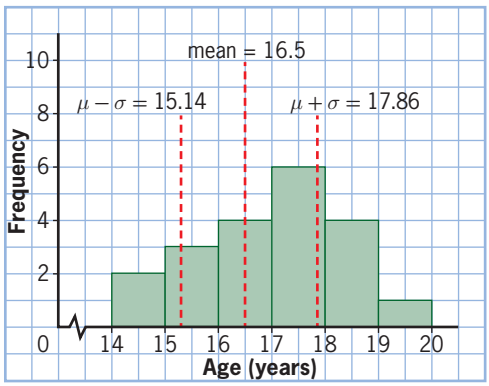
e) $\sigma = 1.395\ 48$ years

Since values increase by the same amount, the spread will not change. The standard deviation would be unchanged.

f) $\mu + \sigma = 17.86$ years

$\mu - \sigma = 15.14$ years

Ages greater than 17 or less than 16 are more than one standard deviation from the mean.



With which type of data (ordinal, quantitative, categorical) is it appropriate to calculate a standard deviation?

Your Turn

The increases in sound volume from a TV program to the advertisements were measured in decibels during a one-hour TV show. The results were as follows:

1.7	1.9	1.5	2.0	2.1	1.8	2.2	1.9	2.0
1.4	1.7	1.8	1.8	2.1	2.7	1.0	0.6	1.8

a) Plot a histogram of the data.

b) Calculate the population mean and standard deviation.

c) Predict what would happen to the standard deviation if the first measurement were 1.5 dB.

d) Predict what would happen to the standard deviation if the second measurement were 1.7 dB.

e) What would happen to the standard deviation if each measurement were 0.5 dB quieter?

f) Which measurements are within one standard deviation of the mean?

A z-score indicates how many standard deviations a data value lies from the mean. In Example 2, part f), the z-score would be 1. You can calculate a z-score using one of the following formulas:

z-score
- the number of standard deviations an observation is from the mean

Population z-Score	Sample z-Score
$z = \frac{x - \mu}{\sigma}$	$z = \frac{x - \bar{x}}{s}$

You can derive computational standard deviation formulas from the given formulas. These formulas simplify the calculations of standard deviation using a scientific calculator.

You will explore the derivation of the computational formulas in Extend #19.

Population Standard Deviation	Sample Standard Deviation
$\sigma = \sqrt{\frac{\sum x^2 - N \cdot \mu^2}{N}}$	$s = \sqrt{\frac{\sum x^2 - n \cdot \bar{x}^2}{n - 1}}$

Example 3

Analysing z-Scores

A food manufacturer makes 2-L jars of pasta sauce. Samples are tested for how close to 2 L the jars are filled. Fifteen samples were taken and their volumes, in litres, were as indicated:

2.11 2.02 2.10 1.99 1.92 2.01 1.89 1.96
2.00 1.96 1.98 2.02 2.08 2.15 2.03

a) Determine the sample mean and standard deviation.

b) Calculate the z-score of the jar that was filled to a volume of 2.02 L. Interpret its meaning.

c) Calculate the z-score of the jar that was filled to a volume of 1.98 L. Compare its distance from the mean to that of 2.02 L.

d) The manufacturer rejects any jars that are filled to less than 1.5 standard deviations below the mean. Which jars would be rejected?

Solution

a) Method 1: Use a Scientific Calculator

Sample mean:

$$\bar{x} = \frac{\sum x}{n}$$
$$= \frac{30.22}{15}$$
$$\approx 2.014\,67$$

The sample mean is approximately 2.014 67 L.

To calculate the sample standard deviation, first add the squares of the measurements:

$$\sum x^2 = 60.955$$

$$s = \sqrt{\frac{\sum x^2 - n \cdot \overline{x}^2}{n-1}}$$

$$= \sqrt{\frac{60.955 - 15(2.014\,67)}{14}}$$

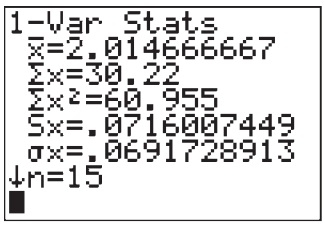
$$\approx 0.0716$$

Most scientific calculators have built-in functions to calculate mean and standard deviation. Refer to your user manual to explore how to calculate the mean and standard deviation on your device.

The sample standard deviation is approximately 0.0716 L.

Method 2: Use a Graphing Calculator

Enter the values into list **L1**. Press STAT CALC and select **1:1-Var Stats**.



Sample mean, $\overline{x} = 2.014\,67$ L.

Sample standard deviation, $s = 0.0716$ L.

b) Use the z-score formula.

$$z = \frac{x - \overline{x}}{s}$$

$$= \frac{2.02 - 2.014\,67}{0.0716}$$

$$\approx 0.0744$$

A jar filled to a volume of 2.02 L is 0.0744 standard deviations greater than the mean. This means that it is very close to the mean.

c) Use the z-score formula.

$$z = \frac{x - \overline{x}}{s}$$

$$= \frac{1.98 - 2.014\,67}{0.0716}$$

$$\approx -0.4842$$

If a positive z-score indicates a standard deviation greater than the mean, what does a negative z-score indicate?

A jar filled to a volume of 1.98 L is 0.4842 standard deviations less than the mean. It is farther from the mean than is 2.02 L.

d) Determine the volume of 1.5 standard deviations below the mean.

$$\overline{x} - 1.5s = 2.014\,67 - 1.5(0.0716)$$

$$= 1.9073$$

Any jars that are filled to less than 1.9073 L would be rejected. Therefore, the jar containing 1.89 L of sauce would be rejected.

Your Turn

A car manufacturer tested the gap between the doors and the body of a car. Eighteen samples were taken. Their gaps, in millimetres, are shown:

1.7 1.9 1.4 1.4 1.5 1.7 1.1 1.6 1.9
1.4 1.5 1.5 1.6 1.5 1.3 1.8 1.6 1.2

a) Determine the sample mean and standard deviation.

b) Calculate the z-score of a door with gap of 1.6 mm. Interpret its meaning.

c) Calculate the z-score of a door with gap of 1.4 mm. Compare its distance from the mean to that of 1.6 mm.

d) The manufacturer rejects any cars with door gaps that are not within two standard deviations of the mean. Which cars would be rejected?

Consolidate and Debrief

Key Concepts

- The variance and standard deviation are measures of spread. The standard deviation is the square root of the variance.

Population variance: $\sigma^2 = \frac{\sum(x-\mu)^2}{N}$

Population standard deviation: $\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$

Sample variance: $s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$

Sample standard deviation: $s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$

- You can use the following computational formulas to calculate standard deviation more easily.

Population standard deviation: $\sigma = \sqrt{\frac{\sum x^2 - N \cdot \mu^2}{N}}$

Sample standard deviation: $s = \sqrt{\frac{\sum x^2 - n \cdot \bar{x}^2}{n-1}}$

- The standard deviation of a set of data determines the average distance of the measurements from the mean. The larger the value, the greater the spread of the data. The units of the standard deviation are the same as for the mean.
- The z-score tells you the number of standard deviations that an observation in a data set is from the mean.

Population z-score: $z = \frac{x-\mu}{\sigma}$ Sample z-score: $z = \frac{x-\bar{x}}{s}$

Reflect

R1. The mean of a set of data is 23.5, with standard deviation of 3.1.

a) What does a z-score of –2 mean for a given data point?

b) What does a z-score of 1.5 mean for a given data point?

R2. Before investing in stocks, you read an analysis that includes the standard deviation of its price over a given period of time. Two stocks have the same mean price of \$15.43 over the past 10 days. Stock A has a standard deviation of \$0.56 and stock B has a standard deviation of \$1.22. What does this mean to you as an investor?

R3. Explain how x relates to the mean if the z-score corresponding to x is

a) positive

b) negative

c) zero

R4. Explain how to decide whether the population or sample formulas need to be used for mean and standard deviation.

Practise

Choose the best answer for #1 and #2.

1. Adam is building a doorway and wants the height of the door to be three standard deviations above the mean Canadian height. How high must the door be if the mean is 210 cm with a standard deviation of 10 cm?

A 230 cm

B 250 cm

C 200 cm

D 240 cm

2. Which is an incorrect statement about standard deviation?

A The variance is the square root of the standard deviation.

B The standard deviation is often called the average distance of the measurements from the mean.

C The standard deviation is expressed in the same units as the data.

D The standard deviation is always a positive quantity.

3. The mean of a data set is 25.3 cm, and the standard deviation is 3.6. Determine the z-score of each of the following and interpret the results.

a) 27.2

b) 24.1

c) 21.9

d) 29.8

4. Calculate the standard deviation for each data set and interpret the results.

a) Lengths, in centimetres, of fish caught on a fishing trip.

15.4 12.3 18.2 9.9
17.4 12.6 16.3 11.8
12.3 12.6 16.7

b) Number of home runs in a season by the players on a team.

3 10 0 12 5 6 10 16 34 11
6 7 21

c) Final scores by the figure skaters in a competition.

168.3 178.2 186.1 134.5
156.7 156.4 167.1 132.0
154.7 149.8 126.2 134.8
154.0 175.2 159.2

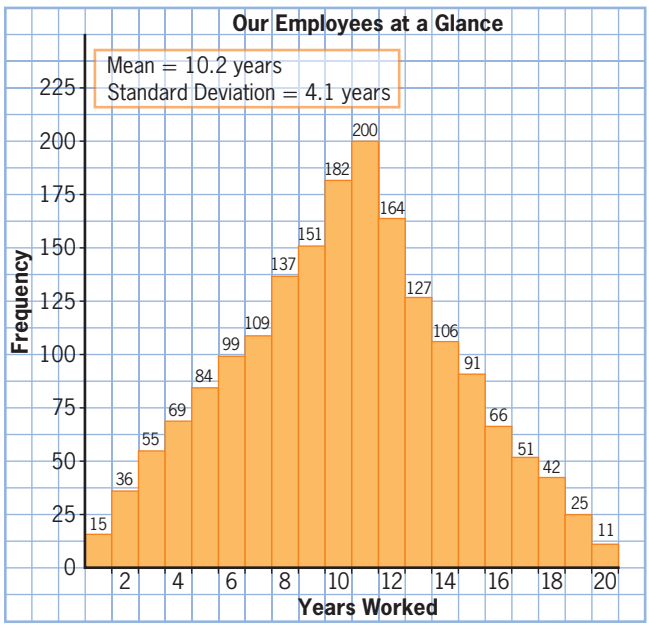
5. For each of the situations, decide whether you would use the sample or population standard deviation formula. Explain your decisions.

a) A researcher recruits females ages 35 to 50 years old for an exercise training study to investigate risk markers for heart disease (e.g., cholesterol).

b) One of the questions on a national survey asks for the respondent's age. Researchers want to describe the variability in all ages received from the survey.

c) A teacher administers a test to her students. The teacher wants to summarize the results the students attained as a mean and standard deviation.

6. As part of a report on its employees, a company published this graph.



a) The standard deviation is given as 4.1 years. Identify which numbers of years worked are within one standard deviation of the mean.

b) What percent of the employees are within two standard deviations of the mean?

c) How does this graph help to explain z-scores?

Apply

7. The chart shows the waiting times for customers while having winter tires installed on their cars.

Time (min)	Frequency
30–35	10
35–40	16
40–45	21
45–50	17
50–55	19
55–60	8
60–65	2

When calculating the standard deviation for data in intervals, use the following formulas:

$$\sigma = \sqrt{\frac{\sum(f_i \cdot m_i^2) - N \cdot \mu^2}{N}} \text{ and}$$

$$s = \sqrt{\frac{\sum(f_i \cdot m_i^2) - n \cdot \bar{x}^2}{n - 1}},$$

where f is the respective frequency of each interval and m is the interval midpoint.

a) Using the midpoints of the intervals as the measurements, estimate the mean and standard deviation of the wait times.

b) Did you use the population or sample formulas? Why?

c) Calculate the z-scores of each of the midpoints.

d) Draw a histogram and mark the information from part c) on the graph.

8. Application The mean size of Canada's 308 electoral districts or ridings is 102 639.28 people, with a standard deviation of 21 855.384. In 2006, Mississauga-Erindale had a population of 143 361. Parkdale-High Park had a population of 102 142.

a) Compare the z-scores for these ridings.

b) What argument could the citizens of Mississauga-Erindale make about their representation in the House of Commons?

9. Maria handed in her final data management project last week. The class mean was 83% with a standard deviation of 8. If Maria's mark produced a z-score of 1.09, what was her grade?

10. The actual volume of milk in 1-L cartons of milk was checked by measuring a selection of 120 cartons. The chart shows the results.

Volume (L)	Frequency
0.98	6
0.99	18
1.00	30
1.01	35
1.02	19
1.03	9
1.04	0
1.05	3

a) Calculate the mean and standard deviation, accurate to three decimal places.

b) Did you use the population or sample formulas? Why?

c) The company has decided that a sample that is within two standard deviations of the mean is acceptable. A random sample was taken and the volume was 0.98 L. Would this be an acceptable sample?

d) On the following day, the mean volume of milk per carton was 1.012 L, with a standard deviation of 0.009 L. Compare the two days' test results.

11. The table shows the lengths of logs shipped to a lumber mill on a particular day.

Length (m)	Frequency
3.5–4.5	3
4.5–5.5	20
5.5–6.5	17
6.5–7.5	38
7.5–8.5	31
8.5–9.5	19
9.5–10.5	15

a) Calculate the mean and standard deviation of the logs, accurate to three decimal places.

b) How does this data set compare to the previous day, with a mean of 8.44 m and standard deviation of 1.836 m?

c) Why would the standard deviation be important to the operators of the lumber mill?

Achievement Check

12. Along with their application to a particular university, students were instructed to submit a 700-word essay. The table shows the lengths of 16 of the essays that were submitted.

Student	Number of Words in Essay
Alex	709
Christian	743
Maria	810
Hasanika	900
Daniel	1112
Barb	568
Brian	804
Cathie	951
Wayne	643
Shaniqua	829
Jiang	674
Bill	769
Mohammed	781
Farah	735
Tim	700
Guido	583

a) Calculate the mean, variance, and standard deviation to the nearest whole number for the data set.

b) Did you use the sample or population formulas? Why?

c) Make an appropriate graph of the data. Mark the interval that is within one standard deviation of the mean.

d) Use z-scores to determine whether Cathie's or Wayne's essay length is closer to the mean.

e) Compare this group's essays to the essays in the previous year, with a mean of 712.1 words and standard deviation of 23.2 words.

13. **Communication** When is it possible for the standard deviation to be larger than the variance?

14. After graduating from university, Yee Ping hopes to get a job in a career with a mean starting salary of \$56 000. Compare the salary ranges for standard deviations of \$15 000 and \$5000, knowing that 95% of the starting salaries are within two standard deviations of the mean.

15. **Communication** When buying an investment such as a mutual fund, investors look at its volatility. Volatility is measured by calculating the standard deviation of the returns over a given period of time.

a) What will the standard deviation show an investor if the mean rate of return of a particular mutual fund unit is 14.37% with a volatility of 6.54%?

b) How would the standard deviation change for a riskier investment?

16. **Thinking** A set of five whole numbers is arranged in order from least to greatest. The fifth number is decreased by one. Would the interquartile range or standard deviation be more affected? Explain.

Extend

17. The mean of a sample of n values is $\overline{x}$ and the standard deviation is s. Suppose you add a constant value a to each observation so that the new data values are

$x_1 + a, x_2 + a, \ldots, x_n + a.$

Determine the new mean and standard deviation.

18. The mean of a sample of n values is $\overline{x}$ and the standard deviation is s. Suppose the observations are multiplied by a constant value c so that the new data values are

$cx_1, cx_2, \ldots, cx_n.$

Determine the new mean and standard deviation.

19. Algebraically derive the computational formula

$$\sigma = \sqrt{\frac{\sum x^2 - N \cdot \mu^2}{N}}$$

from the defined standard deviation formula

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}.$$

6.4

Interpreting Statistical Summaries

Learning Goals

I am learning to

- interpret statistical summaries to describe a one-variable data set and to compare two related one-variable data sets
- understand whether the data presented are valid and reliable
- describe how statistical summaries can misrepresent one-variable data
- make inferences and make and justify conclusions from statistical summaries of one-variable data
- interpret statistics in the media, assess the validity of conclusions made, and explain how statistics are used to promote a certain point of view

Minds On...

Suppose you read this statement online.

- How would you interpret it? Would you agree or disagree?
- Would you like to know who collected the data to be able to make such an inference with confidence?

Action!

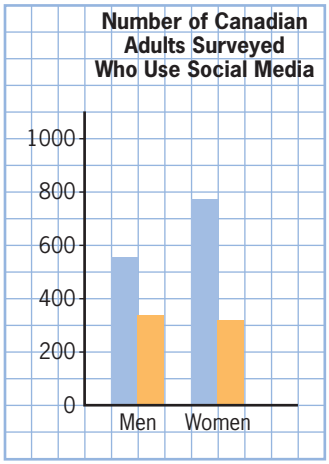
How you collect data is important. How you organize and display data helps you analyse and make conclusions. Finally, how you summarize data determines whether you can make valid generalizations.

Investigate Statistical Claims in the Media

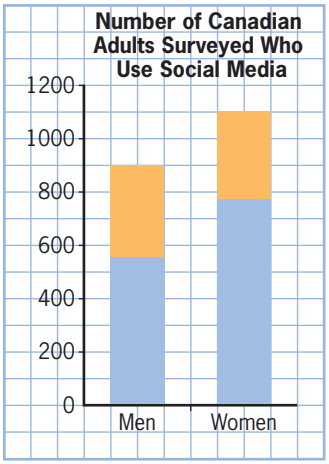
The headline "Significantly more women than men spend time using social media" came from an article about a survey of 900 men and 1100 women on their use of social media. The survey asked the question, "Do you use social media every day?"

The information was compared using a multiple bar graph, a split bar graph, and a relative split bar graph.

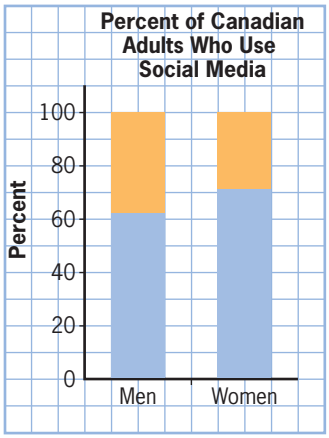
multiple bar graph



split bar graph



relative split bar graph



multiple bar graph

- different quantities are represented by different colours and lengths of bars that are placed side by side

split bar graph

- different quantities are represented by different colours and lengths of bars that are placed one above the other

relative split bar graph

- different percents, totalling 100, are represented by different colours and lengths of bars that are placed one above the other

1. What does the multiple bar graph tell you?
2. What do the split bar graph and relative split bar graph tell you?
3. Which graph helps summarize the information best? Why?
4. Describe any sampling or measurement bias.
5. **Reflect** Does the Facebook post in the introduction accurately reflect the results of the survey? Explain.
6. **Extend Your Understanding** Is it appropriate to extend the results of this survey to the entire population? Explain.

As a society, we collect data to acquire information. As seen in the above investigation, one method of collecting information is through the process of conducting surveys. In chapter 5, you learned that a sample is usually used because it is difficult to survey an entire population.

reliable data

- results of a study that can be duplicated in another study
- repetition of trials will produce more accurate data

valid data

- results that accurately represent the entire population

It is important that the data gathered are both **reliable** and **valid**.

For example, suppose you want to know how many 15- to 18-year-olds in Ontario choose to play Xbox over PlayStation®. You sample students from Aurora, Newmarket, and Stouffville, which are all north of Toronto. Your results are reliable but not necessarily valid, due to sampling bias. Marketing or availability of the units may be different in other parts of Ontario.

Why would these survey results be reliable, but not valid?

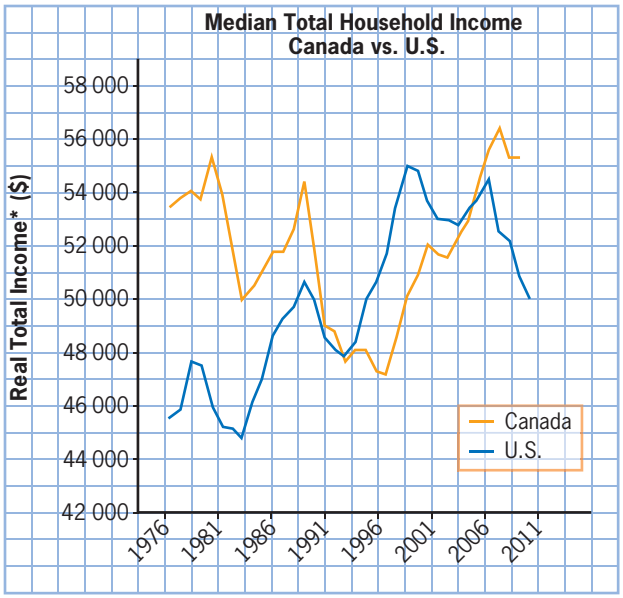
How could you improve the validity of the results?

Example 1

Interpreting Measures of Central Tendency

A recent headline read, "Americans are significantly wealthier than Canadians." The article indicated that, in 2011, the income per capita in Canada was about $38 000, and in the United States was about $42 000.

Another publication showed the following graph. Both cited Statistics Canada and the US Census Bureau as their sources.



*U.S. in $2011, Canada in $2010
Source: U.S. Census Bureau, Statistics Canada

The footnote "*U.S. in $2011, Canada in $2010" indicates that the incomes were adjusted to discount the effects of inflation, using 2011 in the United States and 2010 in Canada as reference years. This is often a challenge when using secondary data to make comparisons. Analysts use whatever statistical summaries are available at the time.

a) Interpret these statements and explain what might cause the discrepancy.

b) Compare the two countries' median household incomes over the years.

c) Discuss the vertical scale of the graph and how it may influence the reader.

Solution

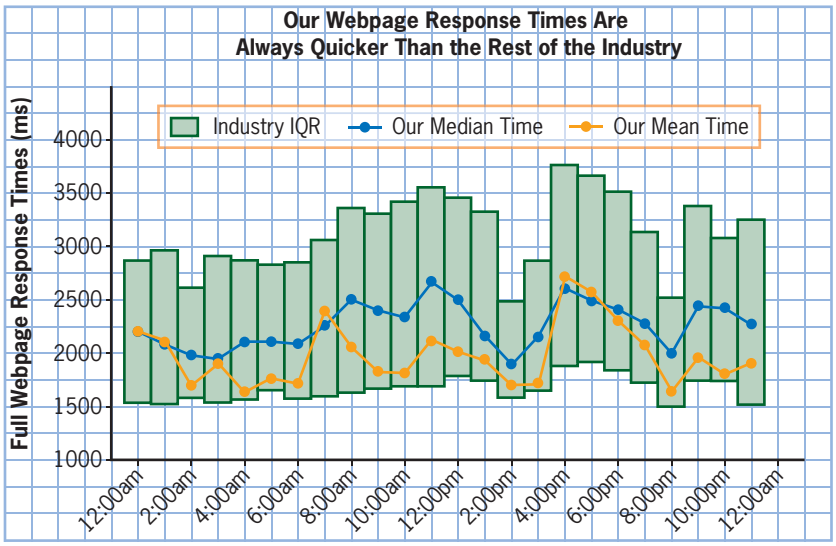
a) The graph indicates that the median income in Canada in 2011 was about $55 500, and in the United States it was about $50 000. Income per capita represents the mean income per person. The mean is influenced by outliers, so extremely wealthy individuals will significantly increase the mean. This results in income per capita being greater in the United States than in Canada. Both articles use reliable sources, but the interpretation in the first article is inaccurate. It is important for the reader to critically analyse claims in the media.

b) The graph shows that the median annual income is greater in Canada than in the United States for most years, with the exception of about 1992 to 2005. For an analysis to be accurate, many years of data must be analysed because the datum for a single year could be an outlier. The use of two different base years could be an issue if there were not such a marked difference between the two countries' results.

c) The vertical scale begins at $42 000. As a result, the differences between the graphs of Canadian and American incomes seem to be greater than in reality. For a more accurate view of the data, the vertical scale should begin at 0.

Your Turn

The graph below was included in a report by an Internet service provider.



a) What statistical measures does the graph show?

b) Which measures of central tendency are used in the Internet service provider's analysis?

c) Discuss whether the Internet service provider's claim is accurate.

d) Discuss the vertical scale of the graph and how it may influence the reader.

Example 2

Comparing Groups Graphically

The Athletic Council of Reliable College wants to know whether males or females play more tennis at their school. They decide to poll the grade 11 and 12 students to organize, display, and draw conclusions about their data.

The following table lists the data they gathered.

	Frequency	
	Male	**Female**
Plays tennis	87	112
Does not play tennis	52	94

a) Create a multiple bar graph, a split bar graph, and a relative split bar graph to display the data.

b) Which graph more clearly shows which gender plays tennis more often?

c) Can you draw any other conclusions by looking at the visual representations of the data?

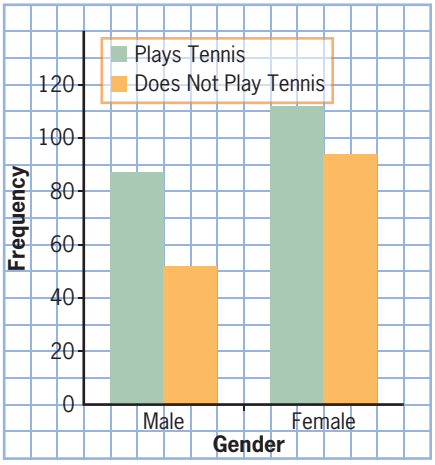
Solution

a) Create a multiple bar graph using the above data.

- Open a spreadsheet.
- Enter the following data:

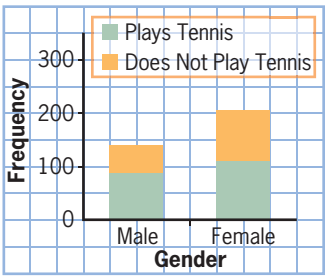
	Male	Female
Plays tennis	87	112
Does not play tennis	52	94

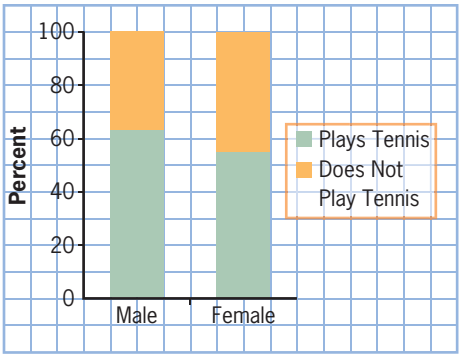
- Highlight the data in the table. Select **Insert**, then select **Column**.
- Choose the first option, **2D Clustered Column**.
- The program will create the chart for you.



Some spreadsheet software refers to a multiple bar graph as a clustered column graph and a stacked bar graph as a stacked column graph. Explore the options in your spreadsheet software to see how these graph types are referred to.

- Create a split bar graph by following the same steps as above, but this time choosing the second option, **2D Stacked Columns**.
- Create a relative frequency stacked graph by selecting **100% Stacked Column**.





Which type of data (categorical, ordinal, quantitative) can be illustrated using these graphs?

b) The multiple bar graph and the split bar graph both show the breakdown of tennis players within the given gender. However, the sample sizes are different, so the relative frequency graph needs to be used to compare how well tennis is liked by males versus females. It indicates that tennis is more popular among males than females in the sample. Looking only at the relative frequency graph, you may not question the sampling method that was used. However, when looking at the multiple bar graph and the split bar graph, you might wonder how likely this gender proportion would be if a good random sample were carried out.

c) The green parts of the relative frequency split bar graph show that more than 50% of both males and females polled play tennis.

Your Turn

You work at StatSmart and your manager wants to determine how the store is performing on sales based on each quarter year, by region. The following sales data were gathered for the store during the year. Organize, display, and draw conclusions about the data to help show the manager how the store is performing.

	StatSmart Sales ($ millions)			
Region	**Quarter 1**	**Quarter 2**	**Quarter 3**	**Quarter 4**
North	100	200	230	185
South	200	137	164	123
East	231	70	110	228
West	125	210	246	166

a) Organize and display the data using a multiple bar graph, a split bar graph, and a relative split bar graph.

b) What can you conclude from the data? Explain.

Example 3

Critical Analysis of Claims in the Media

Source: WorkSmart: 10 Surprising Social Media Statistics That Will Make You Rethink Your Social Strategy; 25% of Smartphone Owners Ages 18–44 Say They Can't Recall the Last Time Their Smartphone Wasn't Next to Them.

a) What message do these statistics convey?

b) What techniques are used to influence the reader with statistics?

c) What questions need to be answered about the data collection to critically analyse the results of the survey?

d) What questions need to be asked to check the reliability of the source?

Solution

a) The statistics suggest that a smartphone is a critical tool for communicating and being in close contact with others all day long.

b) The following techniques were used to influence the reader:

- Coloured highlighting is used for emphasis.
- Circles are used instead of bar graphs, so it is very difficult to see area differences.
- The intervals of time are not equal.
- The sentence "63% of smartphone owners keep their phone with them *for all but an hour of their waking day*…" infers the results of a survey to the entire population.

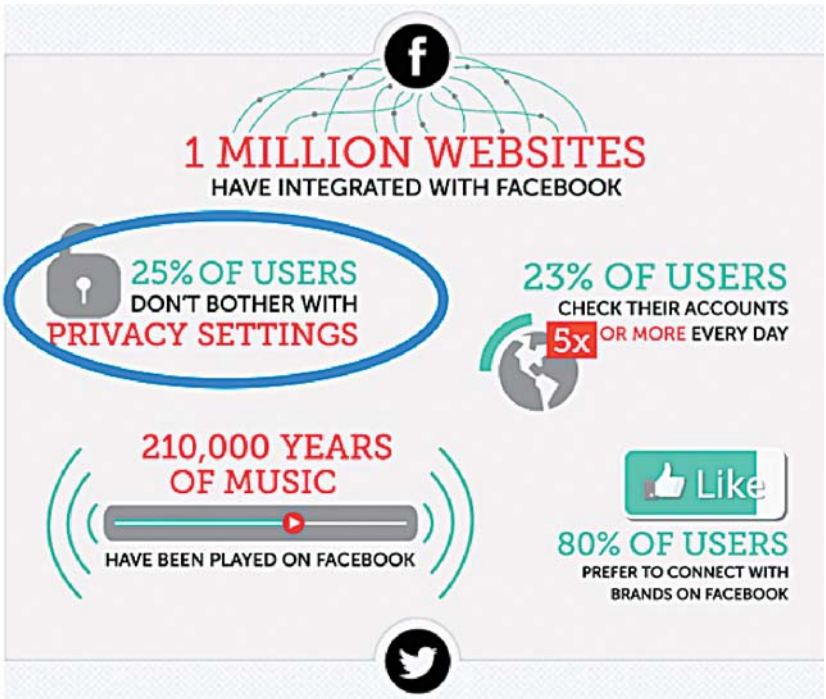
c) The following questions need to be answered to critically analyse the results:

- How large was the sample?
- How was the sample chosen and was it a random selection?
- What questions were asked in the survey?

d) To check the reliability of the source, the following questions should be asked:

- What was the source of the data?
- Were the data primary or secondary?
- Who sponsored the survey?

Your Turn



Source: WorkSmart: 10 Surprising Social Media Statistics That Will Make You Rethink Your Social Strategy; 25% of Facebook Users Don't Bother With Privacy Settings.

a) What messages do these statistics convey?

b) What techniques are used to influence the reader with statistics?

c) What questions need to be answered to critically analyse the results of the survey?

d) What questions need to be asked to check the reliability of the source?

Consolidate and Debrief

Key Concepts

- When you compare data values, it is possible to draw conclusions based on the data set results.
- There may or may not be a relationship between compared values.
- In some instances, graphs provide a stronger visual of the conclusion.
- You can use multiple bar graphs, split bar graphs, and relative split bar graphs to compare two similar data sets.
- Statistics are often used to represent certain points of view by manipulating graph axes, by citing only one measure of central tendency, or through measurement or sampling bias.
- It is key to perform a critical analysis of any statistical report.

Reflect

R1. How are multiple bar graphs, split bar graphs, and relative split bar graphs different? How are they similar?

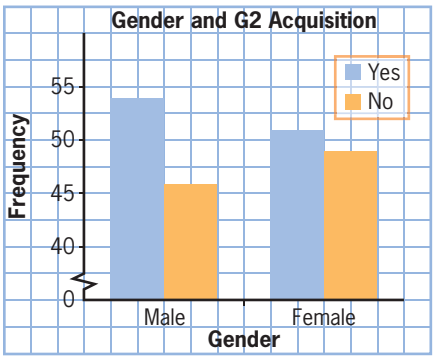
R2. **a)** Why is it important to critically analyse a statistical summary?

b) What are some important questions that would need to be asked?

Practise

Choose the best answer for #3 and #4.

1. Grade 12 students were divided into two groups, male and female, and asked if they have successfully acquired a G2 driver's licence. Based on the results shown below, does gender appear to have an effect on whether or not a grade 12 student has their licence? Explain your reasoning.



2. To decide on a reasonable price for a bottomless cup of hot chocolate, a cafe owner recorded the number of cups each customer ordered on a winter morning.

2	1	2	3	0
1	1	1	2	2
1	3	1	4	2
0	1	2	3	1

a) Will these data offer a reliable data set? Explain.

b) How might the cafe owner ensure a reliable and valid source of data?

3. In a split bar graph,

A the parts are compared to the whole

B the bars are divided into categories

C each bar displays a total

D all of the above

4. Which of the following best describes reliable data?

A represents the entire population

B can be duplicated

C you can trust the sample responses

D none of the above

5. Identify the information provided in the following statistical summary:

Total length of rainbow trout ($n = 128$) averaged 50.4 cm ($s = 12.4$ cm) in June 2014 samples from Lake Ontario.

6. Students were asked whether they ate their lunch in the school cafeteria. The table summarizes their responses.

	No	Yes
Grade 9	21	79
Grade 10	36	54
Grade 11	59	41
Grade 12	84	16

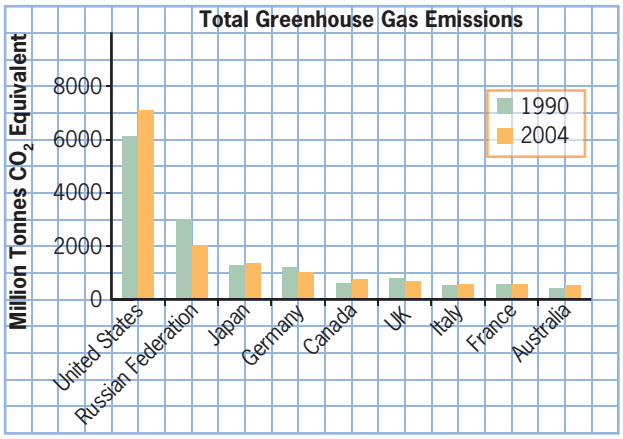
a) Create a multiple bar graph, a split bar graph, and a relative split bar graph to represent the data.

b) Which graph is more informative? Why?

Apply

7. The United Nations published a statement that 2% of the world's population has more than half the world's wealth, whereas half the world's population has only 1% of the world's wealth. In 2013, the world's population reached 7 100 000 people, while the world's total wealth reached $231 trillion. Analyse the United Nations' announcement in context.

8. Application The graph below had the headline, "Is climate change really happening?"



Source: UNFCCC

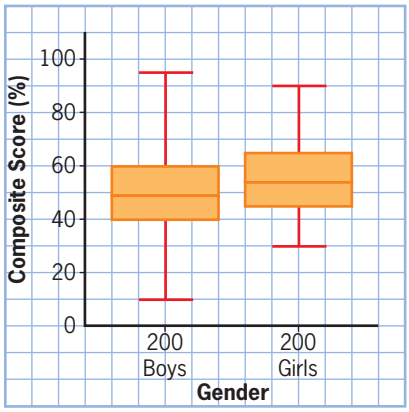
a) What do you think the author is trying to illustrate with this graph?

b) Is there enough evidence to suggest Russia is one of the most polluting countries on the list?

c) Is there enough evidence to suggest Russia has done the most to reduce greenhouse gases?

d) Critically analyse the information in the graph and the headline.

9. Thinking The side-by-side box plots were provided as part of a report on the results of a mathematics aptitude test involving 200 boys and 200 girls.



a) Write a paragraph summarizing the results, using proper mathematical terminology.

b) Evaluate the strength of the evidence that girls are better than boys at math.

10. An article on child development included the statement that, on the height-for-age distribution for a 36-month-old boy, the mean height is 96 cm and the difference between a z-score of –2 and a z-score of –1 is 3.8 cm. The same difference is found between a z-score of 0 and a z-score of +1 on the same distribution. Interpret the meaning of this statement.

11. Application A group of grade 12 male and female students were asked how many minutes they think they will spend getting ready for the school prom. The chart lists their responses.

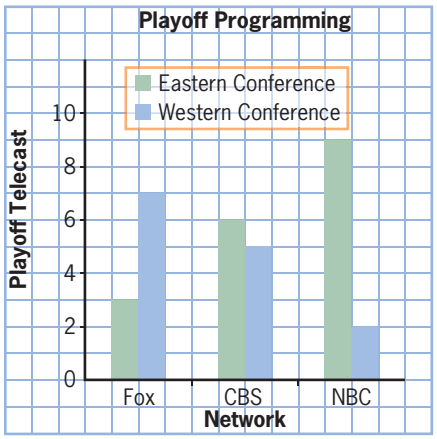
a) Calculate the measures of central tendency.

b) Display the data with an appropriate graph.

c) Write a headline and a one-paragraph article for the school newspaper based on parts a) and b).

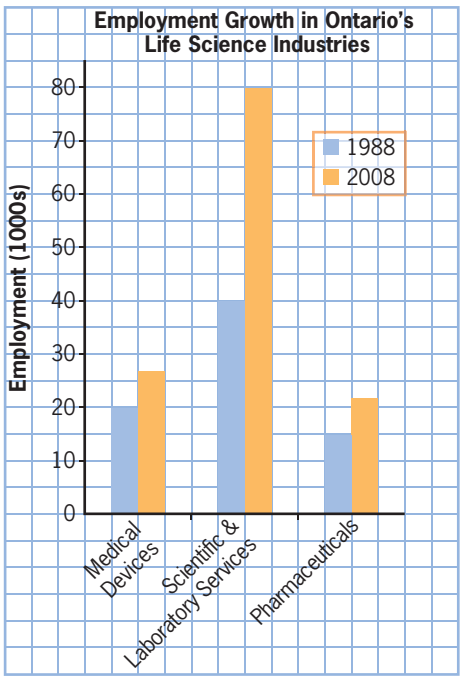
Number of Minutes	
Male	**Female**
60	94
64	128
68	54
59	102
74	108
66	120
88	79
97	87
54	111
51	51

12. Communication The graph shows the number of NFL football playoff telecasts for each US network, broken down by the league's conferences. Write a brief summary of the information provided in the graph from the point of view of CBS.



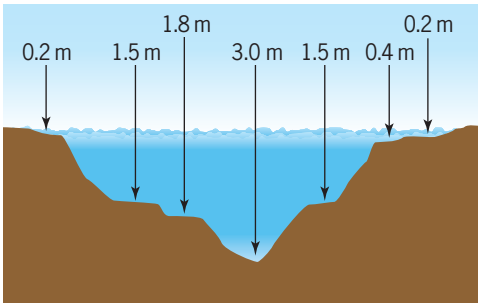
13. a) What is implied by this graph?

b) Where would you look if you were to try to validate the data?



Source: Statistics Canada (Labour Force Survey) and Ontario Ministry of Finance.

14. This diagram was used in an article to illustrate the cross section of a river. Perform appropriate calculations and provide a statistical summary of the depth of the river.



15. A common belief among fitness trainers is that you should do cardiovascular exercises to warm up before weight training. Trainers believe that a warm-up helps the elasticity of the arteries, which is beneficial before resistance training. One study among 15 male power lifters showed that cardio warm-ups caused fatigue among the athletes. The report concluded that cardiovascular training should not occur before weight training. Outline the steps that would be needed to validate this claim.

16. A store manager wants to show head office that sales are increasing dramatically, given that the sales for six months are $321 587, $335 892, $336 998, $340 256, $348 640, and $359 361.

a) Draw a graph to accurately represent the data.

b) Draw a graph to show a dramatic increase in sales.

Achievement Check

17. Your teacher will provide you with a file called **WorldHealthStatistics2013.pdf**. It is a summary of a report by the World Health Organization.

a) Describe the main points of the article.

b) What are the sources of the data? Are they reliable? Explain.

c) What types of graphs are used to present the data? Do the graphs distort the information? If yes, how would you change them to make them more accurate?

d) Would you consider the data and sources to be reliable? Explain.

e) Is there sufficient evidence to conclude that most people have access to clean drinking water, and that access is continuing to improve? Explain.

18. The rate of return on investments is very important to investors. Comparisons to other similar investments are often done. Instead of providing detailed statistics, they are usually summarized by indicating in which quartile the rate of return falls. However, in financial circles, quartiles are listed in reverse, so Q1 would mean the 75th percentile and Q4 the 25th percentile. The following table compares the returns of three mutual funds, by quartile.

Fund	2014	2013	2012	2011	2010
Global Science & Technology Fund	Q4	Q2	Q3	Q3	Q1
Canadian Mineral Resource Fund	Q1	Q2	Q3	Q1	Q2
North American Growth Fund	Q3	Q2	Q2	Q3	Q3

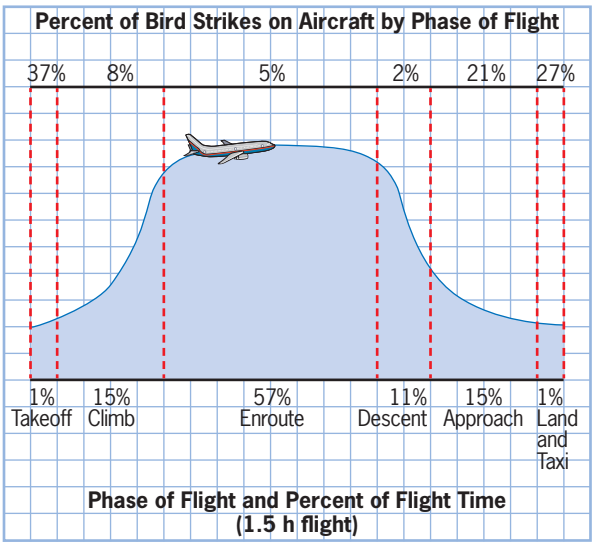
a) Which fund had the most consistent ratings? Explain.

b) Which fund would you recommend as an investment? Explain.

19. Do a search on the Internet for examples of misrepresentations of statistics in the media. Print out three or more of them and describe how statistics are misrepresented.

Extend

20. The diagram shows the percent of bird strikes on Canadian aircraft by phase of flight.



a) Rank the bird strike percents from least to greatest.

b) Consider the number of bird strikes in 100 flights. Rewrite the bird strike percents as percentiles.

c) Recalculate the bird strike percents as bird strike risk, relative to the flight phase. Then re-rank them and change to percentiles.

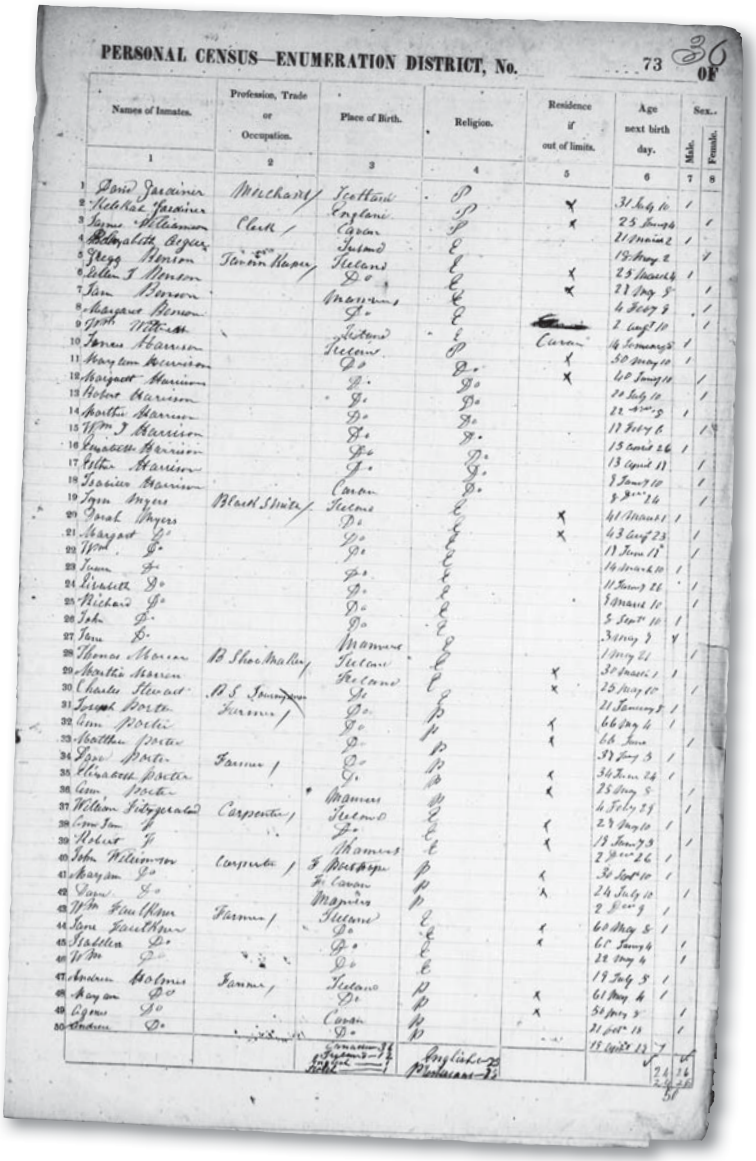
6.5

Analysing Data from Statistics Canada

Learning Goals

I am learning to

- collect data through secondary sources
- generate, using technology, the relevant graphical summaries of one-variable data
- interpret statistical summaries
- assess the validity of conclusions presented in the media
- draw conclusions from the analysis of data and evaluate the strength of the evidence


PERSONAL CENSUS—ENUMERATION DISTRICT, No. ... 73

Names of Inmates.	Profession, Trade or Occupation.	Place of Birth.	Religion.	Residence if out of limits.	Age next birth day.	Sex. Male.	Sex. Female.
1	2	3	4	5	6	7	8
1 David Jamieson	Merchant	Scotland	P		31 July 10	/	
2 Rebekah Jamieson		England	P	X	25 Jany 4		/
3 James Williamson	Clerk	Canada	P	X	21 March 2	/	
4 Elizabeth Bryce		Ireland			19 May 2		/
5 Hugh Monson	Farmer Rapids	Ireland	C		25 March 4	/	
6 Ellen J Monson		Do	C	X	21 May 5		/
7 Sam Monson		Canada	C	X	4 Feby 8		/
8 Margaret Monson		Do	C		2 Sept 10		/
9 Wm Williamson		Ireland			4 Jany 3	/	
10 James Harrison		Ireland	P	Canada	30 May 10	/	
11 Mary Ann Harrison		Do	Do	X	40 Jany 11		/
12 Margaret Harrison		Do	Do	X	20 July 12		/
13 Robert Harrison		Do	Do		22 Aug 3	/	
14 Agatha Harrison		Do	Do		17 Feby 6		/
15 Wm J Harrison		Do	Do		15 Sept 26	/	
16 Eugenia Harrison		Do	Do		13 April 11		/
17 Peter Harrison		Do	Do		3 Jany 10		/
18 Isabella Harrison		Canada	Do		8 Dec 24		/
19 Tom Myers	Blacksmith	Ireland	C	X	41 March 1	/	
20 Sarah Myers		Do	C	X	43 Aug 23	/	
21 Margaret Do		Do	C		13 June 12		/
22 Wm Do		Do	C		14 March 10	/	
23 Tom Do		Do	C		11 Jany 26	/	
24 Russell Do		Do	C		8 March 10		/
25 Richard Do		Do	C		5 Sept 10	/	
26 John Do		Do	C		3 May 7	X	
27 Jane Do		Do	C		1 May 21		/
28 Thomas Warren	B Shoemaking	Ireland	C		30 March 1	/	
29 Martha Warren		Ireland	C	X	28 May 10		/
30 Charles Stewart	B S Journeyman	Do	P	X	21 January 7	/	
31 Joseph Porter	Farmer	Do	P		66 May 4	/	
32 Wm Porter		Do	P	X	66 June		/
33 Matthew Porter		Do	P	X	31 Feby 5	/	
34 Jane Porter	Farmer	Do	P		36 June 24	/	
35 Elizabeth Porter		Do	P	X	25 May 5		/
36 Wm Porter		Do	P	X	4 Feby 23		/
37 William Fitzgerald	Carpenter	Ireland	C		2 May 10	/	
38 Wm Jas Do		Do	C	X	19 June 7		/
39 Robert Do		Do	C	X	2 Dec 26	/	
40 John Williamson	Carpenter	Thomas	P		30 Sept 10	/	
41 Mary Do		Fr Canada	P	X	24 July 10		/
42 Jane Do		Do	P	X	2 Dec 9	/	
43 Wm Faulkner	Farmer	Ireland	C		60 May 5	/	
44 Jane Faulkner		Do	C	X	60 June 4		/
45 Stratton Do		Do	C		28 May 4		/
46 Wm Do		Do	C		19 July 5	/	
47 Andrew Holmes	Farmer	Ireland	P		61 May 4	/	
48 Mary Ann Do		Do	P	X	54 May 5		/
49 Agnes Do		Canada	P	X	31 Oct 10		/
50 Andrew Do		Do	P		11 Oct 12	/	

Minds On...

Statistics Canada is the country's national statistics agency. It collects statistics on Canadians and Canadian issues through the national census and through regular surveys. The data tables are published online at CANSIM (Canadian Socio-Economic Information Management System), Statistics Canada's key socioeconomics database. Once the data are analysed, Statistics Canada produces a report, which is announced in its bulletin, *The Daily*.

The first national census, completed in 1851, asked questions about family, dwellings, religion, and origin. Census takers filled in this census by hand. Over the years, the questions changed relative to the times and the surveys are now fully computerized. What kind of data would it be important for Statistics Canada to gather today to help various government agencies with their planning?

Action!

Investigate 1 Explore Census Profiles

A census is a survey taken of every household in the country. It is mandatory that all households answer the questions in the census. The census occurs every five years, in the years ending in the digits 1 and 6.

Your teacher will direct you to the Statistics Canada website, where you can find results from the most recent census.

1. List which census results are available to view.
2. Census profiles are available for all cities across Canada. Explore this page and describe the steps you need to take to find the census profile of Kitchener, Ontario.
3. Select two cities of your choice and download the census profile for their census metropolitan areas.
4. Give a brief summary of the characteristics that are profiled.
5. Click on the **Download** tab. Select **CSV** format.
6. **Reflect** Select three characteristics and make a graphical and written comparison of your chosen cities.
7. **Extend Your Understanding** How could the city councils use this information in their planning?

Investigate 2 Use CANSIM Tables

Statistics Canada summarizes high-income earners by census metropolitan area (CMA). Your teacher will direct you to the Statistics Canada website, where you can find results from the most recent census.

1. Describe how Statistics Canada categorizes its data.
2. Search for **High Income Trends**. Click on the table number. Describe the data in the table.
3. Identify the source.
4. Identify the sample and sampling method.
5. Identify any possible sources of bias. How reliable would the data be?
6. Select **Add/Remove Data**. Describe the options available.
7. Select **Manipulate**. Describe the options available.
8. Select **Download**. Describe the options available.
9. If you wanted to use data from this table, describe how you would select the items you wish to include and how to download into a spreadsheet.
10. Download the appropriate items to compare the median after-tax income for the top 5% income group in each province and territory for the latest year.

Project Prep

You may decide to use the Statistics Canada website as a source of data for your culminating project. Remember to write down the table number and URL for future reference.

11. **Reflect** Use a spreadsheet or dynamic statistical software to produce graphs that would be useful in a report on the top 5% of income earners.

12. **Reflect** Rank the provinces and territories based on the top 5% of income earners.

13. **Extend Your Understanding** Explain why the rankings might be as they are.

Investigate 3 Analyse a Statistics Canada Report

Select one of the following articles. Your teacher will provide you with a file for either **Gender Differences in STEM Programs at University** or **Differences in Life Expectancy, Inuit vs. Rest of Canada**.

Read the article and perform a critical analysis.

1. What are the major findings of this study?

2. What types of graphs are used? How are the data displayed in graphs? Is there any bias in the graphs?

3. How are the data organized in charts? Is there any bias in how they are presented?

4. What are the sources of the data?

5. How big is the sample? Is it large enough to make inferences to the population?

6. How recent are the data? Are they recent enough for current use?

7. What methods were used to generate or obtain the data? Do the methods show any bias?

8. Does the article make reference to other sources?

9. Who conducted the study? Is there potential for bias due to influences by special interest groups?

10. **Reflect** Consider your answers to all of these questions. Describe your level of confidence that the findings of this study can be applied to the entire population.

11. **Extend Your Understanding** Describe some follow-up questions that might need to be answered if further studies were to be conducted.

Consolidate and Debrief

Key Concepts

- Statistics Canada collects statistics on Canadians and Canadian issues through the national census and through regular surveys.
- Statistics Canada holds a national census every five years.
- Data are published online at CANSIM. The data are available in table form, and often in graphical form.
- Statistical reports are available online on the Statistics Canada website, and are summarized in their bulletin, *The Daily*.
- When reading a statistical report, it is important to perform a critical analysis.

Reflect

R1. A census profile includes such information as population, growth rates, population density, level of education, income, number of immigrants, and age distribution. Who could make use of this information?

R2. Statistics Canada is considered by many organizations to be one of the most reliable sources of data in the world. What factors would be considered in this rating?

R3. Why is it important to perform a critical analysis of a statistical report?

Practise

Choose the best answer for #1 and #2.

1. The national census is taken

A every year

B every 5 years

C every 7 years

D every 10 years

2. CANSIM stands for

A Canadian Simulated Industrial Methodology

B Canadian Association of National Statistical and Informational Marketers

C Canadian Socio-Economic Information Management System

D Canadian Annual National Study of Inferential Media

3. When reading a CANSIM data table,

a) what does the **Add/Remove Data** tab allow you to do?

b) what does the **Manipulate** tab allow you to do?

4. What kind of information is provided in a census profile of a particular city?

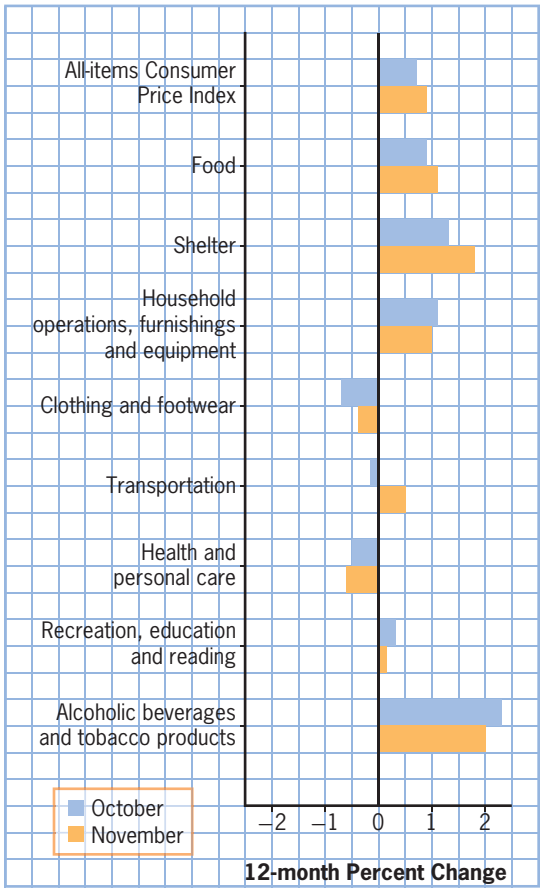
Apply

5. **Communication** Go to the Statistics Canada website. Select the **2006 Census of Population**. Select **Data Products**, then **Census Tract Profiles**. You will be able to enter a postal code and see its statistical profile.

a) How is income related to educational attainment?

b) Collect data on median income, number of post-secondary degrees and diplomas, and size of the population age 15 or older for 5 to 10 different Ontario postal codes. Compare the profiles for these areas and state a hypothesis relating income to education.

6. Consider this Statistics Canada graph with the title **Prices increase in six of eight major components**.



Source: Statistics Canada. *The Daily*, Friday, December 20, 2013: Consumer Price Index, November 2013: Chart 2

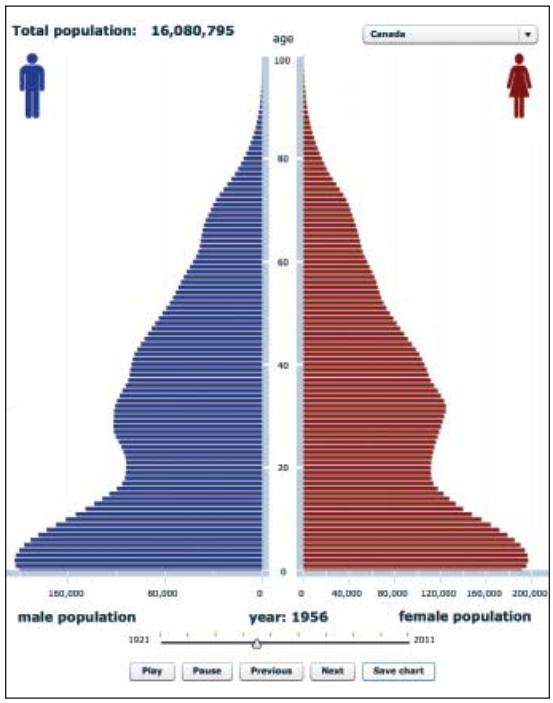
a) Which components of the Consumer Price Index increased in November? Which ones decreased?

b) Which components of the Consumer Price Index increased in October? Which ones decreased?

c) Describe the time frame for the percent change.

7. Communication Go to the Statistics Canada website. Select the **2011 Census**. Select **Data Products**, then **Historical Age Pyramid**.

Historical Age Pyramid for the Population of Canada, 1921 to 2011



Source: Statistics Canada. Censuses of Population and Population Estimates Program, 1921 to 2011. December 23, 2013.

a) Describe the age pyramid in 1921.

b) Describe the age pyramid in 2011.

c) Describe the age pyramid in your year of birth.

d) The post-World War II baby boom occurred between 1946 and 1965. Use the animation feature to follow the baby boom. What does it look like on the age pyramid after 1965?

e) Estimate the male and female populations aged 10 and 50 in 2011.

8. **Application** Go to CANSIM on the Statistics Canada website. Select **Education, Training and Learning**, then **Students**. Select the table that provides data described as, **Weighted average tuition fee for full-time Canadian undergraduate students, by field of study**. Download data that allow you to compare tuition fees in Ontario to any two other provinces, by field of study, for the current and previous school years. Use the data to make appropriate graphs. Calculate measures of central tendency. Make appropriate comparisons between the selected provinces.

9. **Open Question** Choose a topic in the CANSIM category of the environment. Identify the table name and number. Collect data for all provinces and territories. Select the appropriate data, provide graphs, and calculate measures of central tendency and measures of spread where appropriate. Make comparisons to these measures for at least three provinces and territories.

10. Collect and summarize data from the latest edition of the National Household Survey (available on CANSIM) to compare the income distributions for males and females. Write a brief newspaper article on this issue, supported by graphs and summary statistics. Include an appropriate headline.

11. **Thinking** Do a critical analysis on the Statistics Canada article on the Consumer Price Index. Your teacher can provide you with the document entitled **Consumer Price Index**. Alternatively, go to the Statistics Canada website and search for Consumer Price Index. Refer to Investigation 3 for the steps involved in a critical analysis.

Literacy Link

The *Consumer Price Index (CPI)* is a weighted average of a "basket" of goods and services purchased by consumers. The weighting is relative to the level of use in the Canadian economy. The CPI is used as an indicator of changes in consumer prices over time.

12. **Thinking** Do a critical analysis of your choice of articles. Select a topic of interest from the Statistics Canada website. Select
    - English
    - Information for…
    - Analysts and Researchers
    - Click on the html or pdf link of the topic of interest to you.

    Refer to Investigation 3 for the steps involved in a critical analysis.

Extend

13. The first national census was taken in 1851, and occurred every 10 years until 1951. After 1951, a census was held every 5 years. Search for the 1851 census and select the file that compares the 1851 census with the most recent available. Use the download link in the top left corner of the page to download the data. Then use technology to answer the following questions.
    a) Calculate the growth rate for each decade.
    b) Make a time series graph of the population. Describe the graph.
    c) Make a time series graph of the population growth rates. Describe the graph and how it relates to the graph in part b).

Literacy Link

A *time series graph* is a line graph where time is measured on the horizontal axis and the variable being observed is measured on the vertical axis. The variable being observed is usually measured at successive points in time spaced at uniform time intervals.

Chapter 6 Review

Learning Goals

Section	After this section, I can
6.1	• interpret the mean, median, and mode of a set of data • choose the measure of central tendency that best describes the data
6.2	• describe the variability in a sample or population using measures of spread • calculate the range • understand how to use quartiles and percentiles to analyse data
6.3	• use technology to calculate the variance and standard deviation of a data set • calculate and understand the significance of a *z*-score • relate the positive or negative scores to their locations in a histogram • develop significant conclusions about a data set
6.4	• interpret statistical summaries to describe a one-variable data set and to compare two related one-variable data sets • understand whether the data presented are valid and reliable • describe how statistical summaries can misrepresent one-variable data • make inferences and make and justify conclusions from statistical summaries of one-variable data • interpret statistics in the media, assess the validity of conclusions made, and explain how statistics are used to promote a certain point of view
6.5	• collect data through secondary sources • generate, using technology, the relevant graphical summaries of one-variable data • interpret statistical summaries • assess the validity of conclusions presented in the media • draw conclusions from the analysis of data and evaluate the strength of the evidence

6.1 Measures of Central Tendency, pages 252–265

1. **a)** Define the three measures of central tendency.

b) Explain how each measure is determined.

c) Provide a real-life example of where each measure is most appropriate.

2. Calculate the mean, median, and mode of the data sets. Express your answers to one decimal place.

a) 75 989 54 76 675 45 242 54
85 342 12 931 2 37 675

b) 7 19 21 5 17 31 62 7 50 10 7 34

c) 1856 6754 2346 5200
6754 9564 2346 1880

3. A softball player's slugging average is calculated using the formula

$$SLG = \frac{S + 2D + 3T + 4H}{B},$$

where S is the number of singles, D is the number of doubles, T is the number of triples, H is the number of home runs, and B is the number of times batting. Calculate each baseball player's slugging average.

a) Jane, with 85 singles, 15 doubles, 1 triple, and 20 home runs, in 308 times at bat.

b) Tonya, with 56 singles, 25 doubles, 0 triples, and 38 home runs, in 294 times at bat.

c) Monique, with 112 singles, 10 doubles, 9 triples, and 6 home runs, in 315 times at bat.

4. **a)** Determine the mean, median, and modal interval of the data set.

b) Graph the data with a histogram and mark the measures of central tendency on the graph.

Salary Range ($ thousands)	Number of Employees
30–40	10
40–50	18
50–60	31
60–70	14
70–80	5
80–90	0
90–100	2
100–110	5

6.2 Measures of Spread, pages 266–277

5. **a)** Describe what is meant by percentiles and quartiles.

b) Explain how quartiles would be useful for a store ordering shoe sizes.

6. The table provides the number of Facebook friends for a sample of 50 people aged 18 to 25.

a) Determine the percentiles for each of the Number of Friends intervals.

b) Determine the quartiles and the interquartile range.

c) Make a box and whisker plot.

d) Determine whether there are any outliers.

Number of Friends	Frequency
0–25	3
25–50	18
50–75	16
75–100	35
100–125	62
125–150	23
150–175	14
175–200	0
200–225	5
225–250	2

6.3 Standard Deviation and *z*-Scores, pages 278–289

7. The table provides the full-time enrollments of Ontario universities for 2012–2013.

Selected Universities	Total Full-Time Enrollment
Algoma University	1 427
Brock University	15 678
Carleton University	21 988
University of Guelph	20 692
Lakehead University	7 046
Laurentian University	6 635
McMaster University	24 798
Nipissing University	3 757
OCAD University	3 570
University of Ontario Institute of Technology	8 469
University of Ottawa	33 581
Queen's University	19 901
Ryerson University	22 194
University of Toronto	69 081
Trent University	6 760
University of Waterloo	31 611
University of Western Ontario	29 108
Wilfrid Laurier University	15 984
University of Windsor	13 557
York University	44 492

a) Calculate the mean, variance, and standard deviation.

b) What is the z-score for York University?

c) Which universities have a z-score of –2 or less?

8. What does it mean to have a z-score of 1.5?

9. The quality control department of Cool Cola tested the bottle fillers and found them to fill 500 mL bottles to a mean of 501.1 mL, with a standard deviation of 0.48 mL. The company's standard is set at test results being within two standard deviations of the mean.

a) What is the acceptable range of fills?

b) Why would the company want to overfill the bottles?

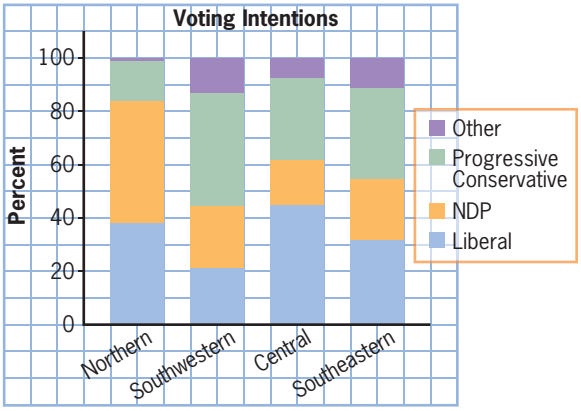
c) Three bottles of Cool Cola were tested for fill volume. Which are acceptable?

i) 501.0 mL **ii)** 502.1 mL

iii) 500 mL

6.4 Interpreting Statistical Summaries, pages 290–301

10. The graph shows the voting intentions in four regions of Ontario, taken from a poll of 2000 voters three days before an election.



a) Identify three pieces of information that you can read from the graph.

b) Would you consider the graph to be a valid predictor of the outcome of the election? Explain.

11. The following information was collected by a large marketing firm interested in attracting individuals from Generation Y to apply for a position at the company. The firm surveyed its managers to develop a list of qualities and benefits that Generation Y employees bring to the workplace. The table shows the number of responses to the question, "What are the most important characteristics of a Generation Y employee?" Visually represent the data and provide a conclusion based on the inferences that can be made.

Characteristic	Frequency
Technological productivity	50
Global mindedness	35
Networking potential	35
Motivation through rewarding experiences	21
Innovative thinking	15
Openness	32
Flexibility	26
Confidence	6
Nothing	12

6.5 Analysing Data from Statistics Canada, pages 302–307

12. Go to CANSIM on the Statistics Canada website. Select **Travel and Tourism**, then **Domestic Travel**. Select the latest version of the file **Air passenger traffic and flights, annual**. Download the data that provide the total number of passengers enplaned and deplaned for each province for the latest year. Make a report that compares the number of airline passengers in each province. Include graphs, the mean, the standard deviation, and comparisons to the mean.

13. Your teacher will provide you with the Statistics Canada summary **Wage Growth Over the Past 30 Years: Changing Wages by Age and Education**. Perform a critical analysis by answering the following questions:

a) What are the major findings of this study?

b) How are the data displayed in graphs? Is there any bias in the graphs?

c) How are the data organized in charts? Is there any bias in how the data are presented?

d) What are the sources of the data?

e) How big is the sample? Is it large enough to make inferences to the population?

f) How recent are the data? Are they recent enough for current use?

g) What methods were used to generate or obtain the data? Do the methods show any bias?

h) Does the article make reference to other sources?

i) Who conducted the study? Is there potential for bias due to influences by special interest groups?

j) Is enough evidence presented to agree with the newspaper headline, "Wages have steadily increased over the past 30 years"? Explain.

Chapter 6 Test Yourself

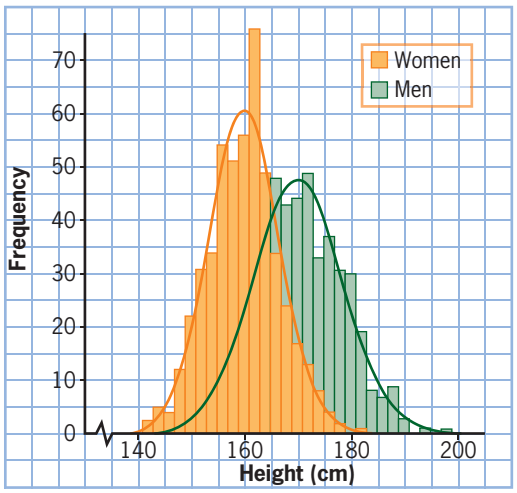
Achievement Chart

Category	Knowledge/ Understanding	Thinking	Communication	Application
Questions	1, 2, 3, 4, 5, 8, 11	6	6, 7	8, 9, 10, 11

Multiple Choice

Choose the best answer for #1 to #4.

1. The graph shows histograms of men's and women's heights in centimetres on the same set of axes.



If the data sets were combined, the distribution of heights would have

A no measures of central tendency

B two modes

C only one set of measures of central tendency

D none of the above

2. Final marks in Maria's Data Management course are based on 70% for term work, 15% for the exam, and 15% for the final course project. What term mark did Maria receive if her final mark was 87 and she received 84 on the exam and 95 on her final project?

A 87% **B** 83%

C 85% **D** 86%

3. Find Q3 for the following masses of students in kilograms:

70 74 78 80 81 84 90 92 94

A 90

B 87

C 91

D 92

4. What measure of central tendency is most appropriate to announce the most used bridge, on a daily basis, in Canada?

A mean

B mode

C median

D weighted mean

5. A set of nine different masses of pet cats are arranged in numerical order. The fifth mass is then increased by one. Which measure of spread for the data set could this change?

A the range

B the standard deviation

C the interquartile range

D all of the above

Short Answer

6. If you are given the data listed below and are asked to use the interquartile range, could you successfully determine which baseball player's home run season totals are more consistent? Explain why or why not.

Ron: 20 21 23 25 18 19
Joshua: 20 20 23 24 19 22

7. Explain why sampling bias is not a major concern for the national census conducted by Statistics Canada.

8. The mean daily temperature during January was –12.1 °C, with a standard deviation of 5.6 °C. Use z-scores to indicate which of the following daily mean temperatures is closest to the monthly mean.

a) –17.4 °C

b) –3.6 °C

c) 0 °C

d) –6.4 °C

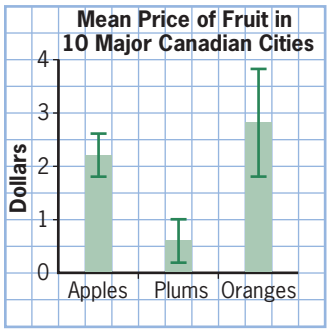
9. For his culminating project, Asim referenced two sources as shown below. What information do these citations tell the reader?

a) Statistics Canada. (2011). Table 109-0300 - Census indicator profile, Canada, provinces, territories, health regions (2011 boundaries) and peer groups, every 5 years, CANSIM [Database]. Retrieved from: http://cansim2.statcan.gc.ca/

b) Statistics Canada. (2012). 2011 Census of Canada visual census, population change by broad age groups, Canada, 1996 to 2001, 2001 to 2006 and 2006 to 2011 [Graph]. Retrieved from http://www12.statcan.gc.ca/census-recensement/2011/dp-pd/vc-rv/index.cfm?Lang=ENG&TOPIC_ID=2&GEOCODE=01

Extended Response

10. The graph illustrates price fluctuations for three types of fruit. Each bar shows the mean price, with plus and minus one standard deviation superimposed. State the mean and interpret the standard deviation for each type of fruit.



11. For a data management project, Ryan sent a survey to the teachers in his school, asking them how many years they have taught. Thirty teachers responded. Here are their responses:

3, 12, 2, 2, 18, 27, 19, 0, 14, 15, 3, 17, 12, 37, 25, 17, 22, 1, 5, 5, 18, 13, 18, 6, 1, 10, 10, 4, 9, 28

a) Calculate the mean, interquartile range, and standard deviation.

b) Organize the data into a frequency distribution with five intervals.

c) Estimate the mean, interquartile range, and standard deviation using the frequency distribution in part b). How do they compare to the true values?

d) Illustrate all the calculations on appropriate graphs.

e) What percentile rank is associated with 10 years of teaching?

f) How many years of teaching are represented by the 90th percentile?

g) Determine whether there are outliers. Identify any that are present.

h) Analyse the validity of Ryan's sampling method.

Chapter Problem

Used Car Lot Business Report

The manager of a used car lot asks you to write a report on the average number of days the cars have been on the lot, along with the spread of the time. He provides you with the data in the table at the right, which represent the length of time, in days, the cars have been on the lot. *Note*: Measurements on an interval boundary are placed into the lower interval.

Time on the Lot (days)	Frequency
0–20	9
20–40	18
40–60	6
60–80	2
80–100	0
100–120	1

1. **a)** Determine the measures of central tendency.

b) Plot a histogram of the data.

c) Determine the interquartile range and make a box and whisker plot.

d) Calculate the variance and standard deviation. Illustrate the standard deviation, along with the mean, on the histogram.

e) How would this set of data compare to another set of data with a mean of 33.9 days and standard deviation of 18.9 days?

2. The original raw data for number of days on the lot are shown below.

17	22	27	41	34	118
45	19	32	8	12	49
12	22	29	53	28	29
31	25	50	21	38	2
20	27	45	1	24	74
30	33	21	61	12	38

a) Calculate the measures of central tendency, the IQR, and the standard deviation.

b) Compare these measures to those of the grouped data. Explain any differences.

c) Identify any outliers.

d) How would the mean, median, IQR, and standard deviation be affected if the outliers were removed from the data set?

3. Write a one-page report to the manager of the used car lot describing the time the cars spent on the lot. Make appropriate use of the measures of central tendency and spread, and support your findings with appropriate graphs.

Chapters 4 to 6 Cumulative Review

1. Two dice are rolled and the product of the upper faces is recorded. Show the probability distribution in table form and graphically.

2. A set of cards with the numbers 200 to 299 is used in a game. The cards are shuffled and the top card is turned up. Calculate the expectation and explain its meaning.

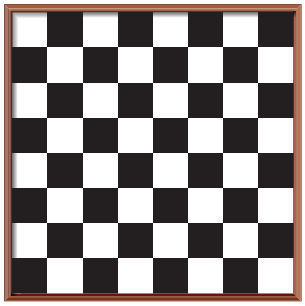
3. The serial numbers on $5 bills include three letters followed by seven digits. Assuming the digits are assigned at random, what is the probability that a serial number will contain

a) exactly two 5s?

b) at least four 5s?

c) all 5s?

4. Five checkers are randomly placed on a checkerboard. What is the probability that three checkers are on squares of one colour and two checkers are on another colour?



5. a) Show the probability distributions, in table form and graphically, for the following distributions:

i) Selecting a card four times, with replacement, from a standard deck, and recording the number of diamonds.

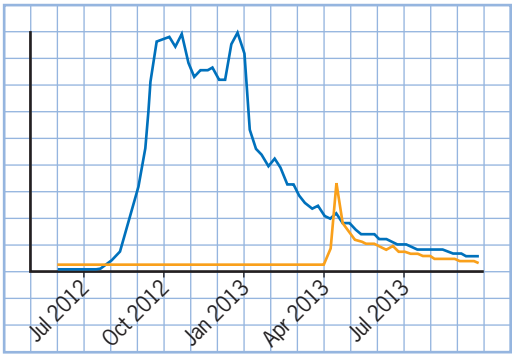
ii) Selecting four cards at the same time, from a standard deck, and recording the number of diamonds.

b) Compare the resulting graphs.

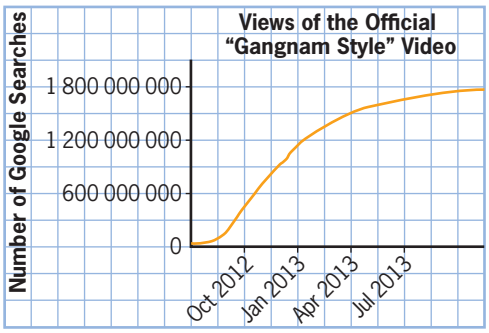
c) Compare the expectations and comment on your findings.

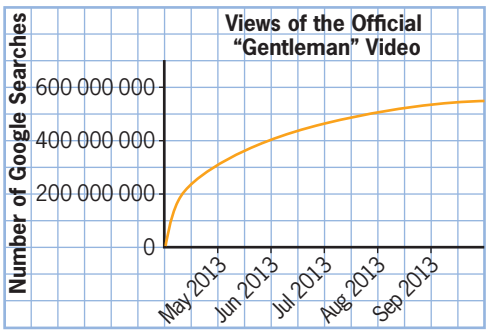
6. Thinking In 2012, the musician Psy brought Korean pop music (K-Pop) to the world with his hit song "Gangnam Style." In 2013, he released another big hit, "Gentleman."

a) The graph shows the number of Google searches related to "Gangnam Style" and "Gentleman." Which colour represents which search? Give reasons for your answer.



b) The graphs below show the number of YouTube views of the videos for each song as of September 2013. Based on these data, do you think that "Gentleman" is doing better or worse than "Gangnam Style"? Justify your answer.





7. You are conducting a survey about one of the following topics: entertainment, sports, the environment, school, or technology.

a) Discuss the methods you will use to conduct your survey.

b) What questions will you ask?

c) What types of data will you collect?

d) How will you keep the data free of bias?

8. The table shows some data from *Romeo and Juliet*.

Character	Number of Words Spoken	Appearances
Romeo	4690	163
Juliet	4314	118
Friar	2624	52
Nurse	2223	91
Capulet	2156	50
Mercutio	2112	62
Benvolio	1157	64
Lady Capulet	874	45
Prince	590	16
Paris	542	23
Montague	319	10
Tybalt	263	17
Sampson	256	20
Peter	248	13
Balthasar	233	12

a) How is the title of the play reflected in the data?

b) Does it appear that the number of words spoken is related to the number of appearances? Justify your answer using the data.

c) Who has the most spoken words in relation to their number of appearances?

d) Create a graphical summary of the data. Your teacher may provide you with a file called **RomeoAndJuliet.csv**.

Use the following information to answer #9 to #13.

The chart shows the temperature, in degrees Celsius, of coffee in 30 recently tested coffee makers.

76	68	72	73	70	69	68	73	81	72
66	85	72	72	69	72	67	74	73	69
75	65	70	71	71	71	68	74	79	73

9. Calculate

a) the mean, median, and mode

b) the range, standard deviation, and variance

c) the quartiles and interquartile range

10. a) Is there any value(s) in the data set that could potentially change the outcome of the measures of central tendency? Explain.

b) Remove the value(s) identified in part a) and recalculate the mean, median, and mode. Which measure of central tendency is most appropriate to describe the distribution of the temperatures? Explain why.

c) What makes the other two measures less appropriate? Explain why.

11. a) Create a frequency table by grouping the data into intervals.

b) Create a histogram and a box and whisker plot of the data.

12. Coffee makers below the 5th and above the 95th percentiles are not recommended.

a) How many of these coffee makers will this include?

b) What are the temperatures of the coffee in the non-approved coffee makers?

13. You want to make a generalization about variability of coffee temperatures in coffee makers. Do you have enough information to make this claim? If not, explain what other pieces of information you would need.