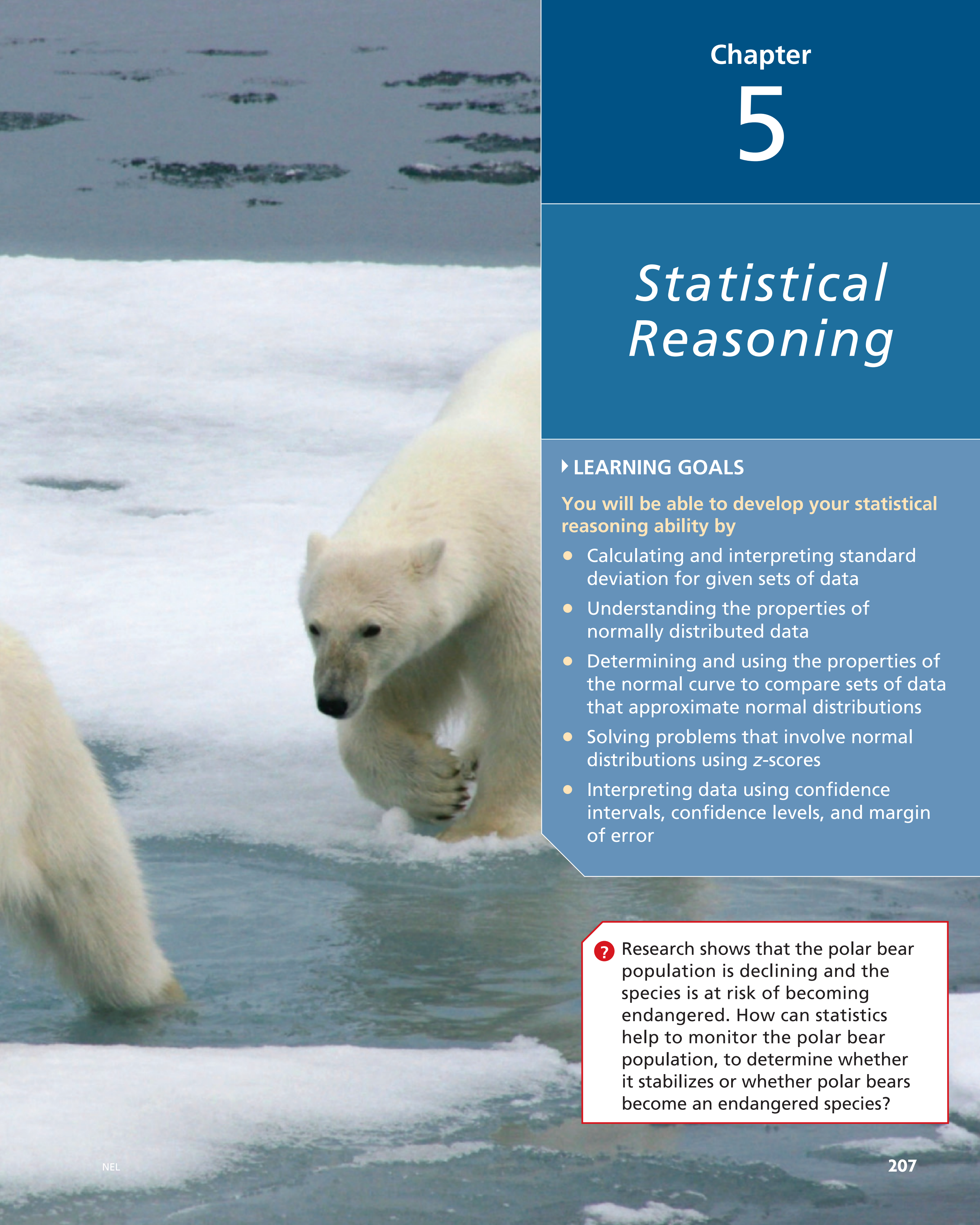


Chapter 5

Statistical Reasoning

LEARNING GOALS

You will be able to develop your statistical reasoning ability by

- Calculating and interpreting standard deviation for given sets of data
- Understanding the properties of normally distributed data
- Determining and using the properties of the normal curve to compare sets of data that approximate normal distributions
- Solving problems that involve normal distributions using *z*-scores
- Interpreting data using confidence intervals, confidence levels, and margin of error

? Research shows that the polar bear population is declining and the species is at risk of becoming endangered. How can statistics help to monitor the polar bear population, to determine whether it stabilizes or whether polar bears become an endangered species?

5 Getting Started

YOU WILL NEED

- calculator
- grid paper

Comparing Salaries

The payrolls for three small companies are shown in the table. Figures include year-end bonuses. Each company has 15 employees. Sanela wonders if the companies have similar "average" salaries.

Employee Payrolls ($)		
Media Focus Advertising	**Computer Rescue**	**Auto Value Sales**
245 000	362 000	97 500
162 000	112 000	66 900
86 000	96 500	64 400
71 000	96 500	63 800
65 000	63 000	62 800
61 000	62 500	62 300
61 000	59 200	61 500
57 500	59 000	58 900
47 400	56 500	58 300
42 500	55 900	58 200
39 500	55 200	57 900
36 200	53 800	57 300
33 400	53 100	56 900
28 500	52 700	55 250
27 300	52 300	55 250

? What is the best indicator of an "average" salary for each company?

A. What is the **range** of salaries for each company?

B. Examine the data. Which companies have data that would be considered **outliers**? Tell how you know.

C. Determine the measures of central tendency (**mean**, **median**, and **mode**) for the salaries for each company.

D. Which measure of central tendency is most affected by outliers? Explain.

E. Create a **line plot** for each of the three companies. Look for outliers, measures of central tendency, and the range on your line plots. Which of these features are easily visible?

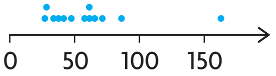
F. Which measure of central tendency best illustrates the "average" salary for each company? Why?

outlier

A value in a data set that is very different from other values in the set.

line plot

A graph that records each data value in a data set as a point above a number line.



WHAT DO You Think?

Decide whether you agree or disagree with each statement. Explain your decision.

1. To compare two sets of data, you need only the mean, the median, and the mode.
2. Most sets of data are evenly distributed about their mean.
3. By looking at the data for a survey, you can decide if the results for the sample that was surveyed closely match the results you would get if you surveyed the whole population.

5.1 Exploring Data

YOU WILL NEED

- calculator
- grid paper

GOAL

Explore the similarities and differences between two sets of data.

EXPLORE the Math

Paulo needs a new battery for his car. He is trying to decide between two different brands. Both brands are the same price. He obtains data for the lifespan, in years, of 30 batteries of each brand, as shown below.

Measured Lifespans of 30 Car Batteries (years)

Brand X					Brand Y				
5.1	7.3	6.9	4.7	5.0	5.4	6.3	4.8	5.9	5.5
6.2	6.4	5.5	5.7	6.8	4.7	6.0	4.5	6.6	6.0
6.0	4.8	4.1	5.2	8.1	5.0	6.5	5.8	5.4	5.1
6.3	7.5	5.0	5.7	8.2	5.7	6.8	5.6	4.9	6.1
3.3	3.1	4.3	5.9	6.6	4.9	5.7	6.2	7.0	5.8
5.8	6.4	6.1	4.6	5.7	6.8	5.9	5.3	5.6	5.9

? How can you compare the data to help Paulo decide which brand of battery to buy?

Reflecting

A. Describe how the data in each set is distributed. Describe any similarities and differences between the two sets of data.

B. Explain why the mean and median do not fully describe the difference between these two brands of batteries. Consider the range, which is one measure of **dispersion** for data. Explain what additional information can be learned from the range of the data.

C. Is the mode useful to compare in this situation? Explain.

D. Suppose that one battery included in the set of data for brand Y is defective, and its lifespan is 0.5 years instead of 5.9 years. Discuss how this would or would not affect Paulo's decision.

dispersion

A measure that varies by the spread among the data in a set; dispersion has a value of zero if all the data in a set is identical, and it increases in value as the data becomes more spread out.

In Summary

Key Ideas

- Measures of central tendency (mean, median, mode) are not always sufficient to represent or compare sets of data.
- You can draw inferences from numerical data by examining how the data is distributed around the mean or the median.

Need to Know

- To compare sets of data, the data must be organized in a systematic way.
- When analyzing two sets of data, it is important to look at both similarities and differences in the data.

FURTHER Your Understanding

1. a) Construct a graph to illustrate the average daily temperatures in Langley, British Columbia, and Windsor, Ontario.

Average Daily Temperatures in Langley, BC												
Month	**Jan.**	**Feb.**	**Mar.**	**Apr.**	**May**	**Jun.**	**Jul.**	**Aug.**	**Sep.**	**Oct.**	**Nov.**	**Dec.**
average daily temperature (°C)	2.2	4.4	6.3	8.6	11.8	14.2	16.7	17.0	14.2	9.8	5.1	2.7

Average Daily Temperatures in Windsor, ON												
Month	**Jan.**	**Feb.**	**Mar.**	**Apr.**	**May**	**Jun.**	**Jul.**	**Aug.**	**Sep.**	**Oct.**	**Nov.**	**Dec.**
average daily temperature (°C)	−4.5	−3.2	2.0	8.2	14.9	20.1	22.7	21.6	17.4	11.0	4.6	−1.5

Environment Canada

b) Determine the range, mean, and median for the average daily temperatures in the two cities.

c) Use your graph and your results from part b) to compare the temperatures in the two cities.

d) Why might a comparison of the two sets of data be useful?

2. a) Use the range and measures of central tendency (mean, median, and mode) to compare the results for two geography tests given by the same teacher to the same class in the same semester.

Unit 1 Test				
81	76	73	71	64
80	75	73	71	63
79	75	73	68	61
79	74	73	67	58
78	73	72	66	57

Unit 2 Test				
98	84	73	71	57
95	81	73	69	53
93	79	73	64	44
89	79	73	59	41
87	76	73	59	37

b) Did the class perform better on the Unit 1 test or Unit 2 test? Justify your decision.

c) Were the modes useful to compare in this situation? Explain.

3. a) Describe the distribution of data for average housing prices in 11 major Canadian cities in 1996, 1998, and 2000. Then compare the three sets of data.

Average Housing Prices ($)			
City	**1996**	**1998**	**2000**
St. John's	116 443	118 519	137 665
Halifax	117 990	141 353	156 988
Toronto	206 738	220 049	224 246
Winnipeg	144 858	161 337	166 761
Regina	147 889	152 784	152 114
Calgary	157 768	180 258	193 275
Edmonton	146 280	164 808	172 503
Vancouver	212 010	218 025	236 617
Victoria	208 400	246 135	228 983
Whitehorse	157 677	167 396	170 986
Yellowknife	181 790	175 646	221 632

Statistics Canada

b) Why might a comparison of the three sets of data be useful?

5.2 Frequency Tables, Histograms, and Frequency Polygons

GOAL

Create frequency tables and graphs from a set of data.

YOU WILL NEED

- calculator
- grid paper

LEARN ABOUT the Math

Flooding is a regular occurrence in the Red River basin. During the second half of the 20th century, there have been nine notable floods, four of which have been severe, occurring in 1950, 1979, 1996, and 1997. The flood that occurred in 1997 is known as the "flood of the century" in Manitoba and North Dakota.

The following data represents the flow rates of the Red River from 1950 to 1999, as recorded at the Redwood Bridge in Winnipeg, Manitoba.

Maximum Water Flow Rates for the Red River, from 1950 to 1999, Measured at Redwood Bridge*

Year	Flow Rate (m^3/s)	Year	Flow Rate (m^3/s)	Year	Flow Rate (m^3/s)	Year	Flow Rate (m^3/s)	Year	Flow Rate (m^3/s)
1950	3058	1960	1965	1970	2280	1980	881	1990	396
1951	1065	1961	481	1971	1526	1981	159	1991	280
1952	1008	1962	1688	1972	1589	1982	1458	1992	1399
1953	357	1963	660	1973	530	1983	1393	1993	946
1954	524	1964	1002	1974	2718	1984	1048	1994	1121
1955	1521	1965	1809	1975	1671	1985	991	1995	1877
1956	1974	1966	2498	1976	1807	1986	1812	1996	3058
1957	654	1967	1727	1977	187	1987	2339	1997	4587
1958	524	1968	510	1978	1750	1988	564	1998	1557
1959	991	1969	2209	1979	3030	1989	1390	1999	2180

National Research Council Canada

(*assumes NO flood protection works in place, for data after 1969 when the floodway was in use)

? How can you approximate the water flow rate that is associated with serious flooding in Winnipeg?

EXPLORE...

- Margaret inherited her grandfather's coin collection. How can she organize a catalogue of the coins to see how many of each type of coin she has?

EXAMPLE 1 Creating a frequency distribution

Determine the water flow rate that is associated with serious flooding by creating a **frequency distribution**.

frequency distribution
A set of intervals (table or graph), usually of equal width, into which raw data is organized; each interval is associated with a frequency that indicates the number of measurements in this interval.

Francine's Solution: Creating a frequency distribution table

Highest water flow rate: 4587 m^3/s, in 1997
Lowest water flow rate: 159 m^3/s, in 1981

I decided to organize the data from the table on page 213 into a frequency distribution table because there are too many numbers to order easily. A table would allow me to see, at a glance, the frequency of various flow rates.

Range: 4587 − 159, or 4428

I determined the range of the data so that I could choose a suitable interval.

$$\frac{4428}{10} = 442.8$$

If the interval width is 500, the intervals will end at 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, and 5000.

Most tables have between 5 and 12 intervals, so I decided to use 10 equally sized intervals to sort the data. Since the range is 4428, I needed an interval width of at least 442.8. I decided to round this value to 500, since 500 will be easier to work with.

Flow Rate (m^3/s)	Tally	Frequency (number of years)
0–500	𝍸 |	6
500–1000	𝍸 𝍸 |	11
1000–1500	𝍸 ||||	9
1500–2000	|𝍸 𝍸 ||||	14
2000–2500	𝍸	5
2500–3000	|	1
3000–3500	|||	3
3500–4000		0
4000–4500		0
4500–5000	|	1

Each interval begins just after the first value in the row and includes all the numbers up to and including the last number. The first interval begins at 0.01 and goes up to 500. This ensures that a number like 500 is counted in only one row.

Communication *Tip*

In frequency tables in this resource, the upper limit of each interval includes that number. For example, in the Flow Rate table to the left, 2000 is included in the interval 1500–2000 and not in the interval 2000–2500.

Flow Rate (m^3/s)	Tally	Frequency (number of years)
0–500	𝍸 |	6
500–1000	𝍸 𝍸 |	11
1000–1500	𝍸 ||||	9
1500–2000	𝍸 𝍸 ||||	14
2000–2500	𝍸	5
2500–3000	|	1
3000–3500	|||	3 floods
3500–4000		0
4000–4500		0
4500–5000	|	1 flood

I could see that 17 of the 50 years had a water flow less than or equal to 1000 m^3/s. There would have been no flooding in those years.

From the data, I knew that the "flood of the century" had a flow rate of 4587 m^3/s. Looking at the last row in my frequency table, I noticed that this was significantly higher than all the other flow rates.

I knew that there were nine floods, and four were severe. Since there were floods in the four years when flow rates were greater than 3000 m^3/s, flow rate and flooding are likely connected. Four of the six flow rates in the 2000–3000 interval would probably have caused floods. The minimum flow rate that results in a flood should be in the 2000–2500 interval.

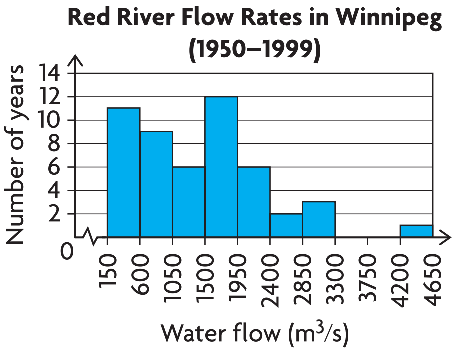
I predict that water flow rates that result in serious flooding are greater than 2000 m^3/s.

I can check my prediction by comparing the years that had flow rates from 2000 to 3000 m^3/s with the historical records of flooding.

Tasha's Solution: Creating a histogram

Flow Rate (m^3/s)	Tally	Frequency (number of years)
150–600	𝍸 𝍸 |	11
600–1050	𝍸 ||||	9
1050–1500	𝍸 |	6
1500–1950	𝍸 𝍸 ||	12
1950–2400	𝍸 |	6
2400–2850	||	2
2850–3300	|||	3
3300–3750		0
3750–4200		0
4200–4650	|	1

I created a frequency table using an interval width of 450 for the Red River water flow data from page 213.



I drew a **histogram** to represent the data in the frequency distribution, since the data is grouped into intervals.

The interval width for my histogram is 450. I labelled the intervals on the horizontal axis. I labelled the frequency on the vertical axis "Number of years."

In general, with the exception of the interval 1500–1950, as the maximum flow rate increases, the number of data points in each interval decreases. Low maximum flow rates have been more common than high maximum flow rates.

histogram

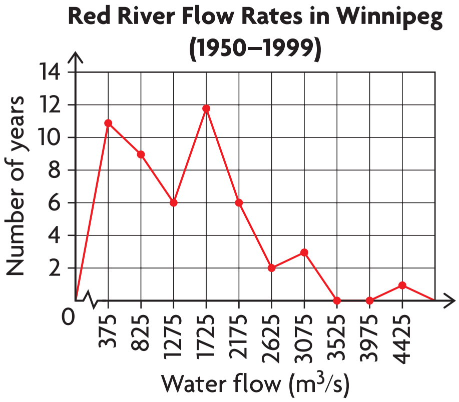
The graph of a frequency distribution, in which equal intervals of values are marked on a horizontal axis and the frequencies associated with these intervals are indicated by the areas of the rectangles drawn for these intervals.

There were nine floods. Based on my histogram, the flow rate was greater than 1950 m^3/s in only 12 years. These 12 years must include the flood years. I predict that floods occur when the flow rate is greater than 1950 m^3/s.

Monique's Solution: Creating a frequency polygon

Flow Rate (m^3/s)	Midpoint	Frequency (number of years)
150–600	375	11
600–1050	825	9
1050–1500	1275	6
1500–1950	1725	12
1950–2400	2175	6
2400–2850	2625	2
2850–3300	3075	3
3300–3750	3525	0
3750–4200	3975	0
4200–4650	4425	1

I created a frequency table using an interval width of 450 for the Red River water flow data from page 213. I determined the midpoint of each interval by adding the boundaries of each interval and dividing by 2.



Next, I labelled the axes on my graph.

Finally, I plotted the midpoints and joined them to form a **frequency polygon**. I included one interval above the highest value, with a frequency of 0, and connected the first midpoint to 0 to close the polygon.

The graph made it easy to see that the flow rate of 4425 m^3/s was unusual.

Most of the data is in the first four intervals, and the most common water flow is between 1500 and 2000 m^3/s. After this, the frequencies drop off dramatically.

There were six years where the flow rate was around 2625, 3075, or 4425 m^3/s. These must have been flood years. The other three floods should have occurred when the flow rate was around 2175 m^3/s. According to my frequency polygon, there were flows around that midpoint in six years. Assuming that the flow rate in three of those years was 2175 m^3/s or greater, floods should occur when the flow rate is 2175 m^3/s or greater.

frequency polygon
The graph of a frequency distribution, produced by joining the midpoints of the intervals using straight lines.

Reflecting

A. Identify similarities and differences in the frequency distributions created by Francine, Tasha, and Monique.

B. Francine used a different interval width than Tasha and Monique. How did this affect the distribution of the data? Explain.

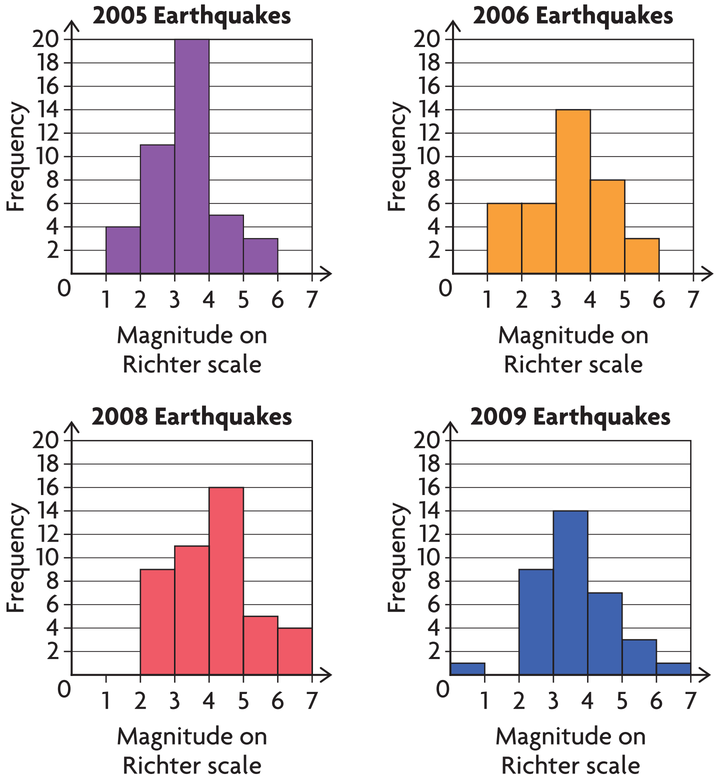
C. Suppose that Francine created a frequency distribution using an interval width of 200. Do you think this interval would make it easier to see which water flow rates result in flooding? Why or why not?

D. Would 2000 be a good interval width for determining critical water flow rates? Explain.

APPLY the Math

EXAMPLE 2 Comparing data using histograms

The magnitude of an earthquake is measured using the Richter scale. Examine the histograms for the frequency of earthquake magnitudes in Canada from 2005 to 2009. Which of these years could have had the most damage from earthquakes?



National Research Council Canada

Understanding the Richter Scale*	
Magnitude	**Effects**
less than 3.0	recorded by seismographs; not felt
3.0–3.9	feels like a passing truck; no damage
4.0–4.9	felt by nearly everyone; movement of unstable objects
5.0–5.9	felt by all; considerable damage to weak buildings
6.0–6.9	difficult to stand; partial collapse of ordinary buildings
7.0–7.9	loss of life; destruction of ordinary buildings
more than 7.9	widespread loss of life and destruction

*Every unit increase on the Richter scale represents an earthquake 10 times more powerful. For example, an earthquake measuring 5.6 is 10 times more powerful than an earthquake measuring 4.6.

Bilyana's Solution: Using a frequency table

2005 had the most earthquakes in any one category: 20 earthquakes with a magnitude from 3.0 to 3.9.

At first glance, it seemed that 2005 was the worst year, because it had the highest bar. However, an earthquake of magnitude 3 on the Richter scale does not cause much damage.

Year	2005	2006	2007	2008	2009
Frequency of Earthquakes from 4.0 to 4.9	5	8	9	16	7

I knew that some minor damage will occur if the magnitude of an earthquake is from 4.0 to 4.9 on the Richter scale. I decided to examine the frequency of these earthquakes for 2005 to 2009.

2008 had the greatest number of earthquakes with the potential for minor damage.

Four of the years had three earthquakes with magnitudes from 5.0 to 5.9, while 2008 had five earthquakes with these magnitudes.

Year	Magnitude on Richter Scale			Total
	4.0–4.9	5.0–5.9	6.0–6.9	
2008	16	5	4	25
2009	7	3	1	11

Only 2008 and 2009 had earthquakes with magnitudes from 6.0 to 6.9, with the potential to cause moderate damage. I decided to examine these years more carefully.

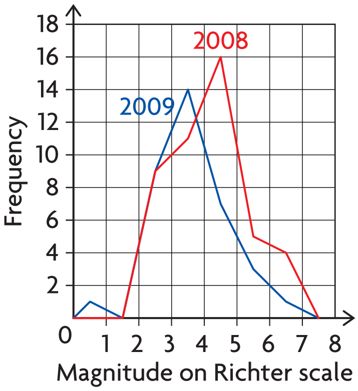
I created a table for earthquakes from 4.0 to 6.9 on the Richter scale for these years.

Therefore, 2008 could have had the most damage from earthquakes.

Miguel's Solution: Comparing frequency polygons

Both 2008 and 2009 had the strongest earthquakes, registering from 6.0 to 6.9 on the Richter scale.

I examined the histograms. There were only two histograms with earthquakes that registered more than 6 on the Richter scale. These years could have had the most damage, because an earthquake registering from 6.0 to 6.9 will result in moderate damage.



I decided to draw frequency polygons, instead of histograms, for these two years. I drew both polygons on the same graph to compare them.

The number of earthquakes in the three highest intervals was greater in 2008 than in 2009, so 2008 could have had the most damage from earthquakes.

I compared the shapes of the frequency polygons.

Your Turn

a) Compare Bilyana's solution with Miguel's solution.

b) What other factors should be considered when determining which year could have had the most damage from earthquakes?

In Summary

Key Ideas

- Large sets of data can be difficult to interpret. Organizing the data into intervals and tabulating the frequency of the data in each interval can make it easier to interpret the data and draw conclusions about how the data is distributed.
- A frequency distribution is a set of intervals and can be displayed as a table, a histogram, or a frequency polygon.

Need to Know

- A frequency distribution should have a minimum of 5 intervals and a maximum of 12 intervals, although any number of intervals is possible. Too many or too few intervals will result in a table or a graph that may not effectively show how the data is distributed.
- The interval width can be determined by dividing the range of the data by the desired number of intervals and then rounding to a suitable interval width.
- The height of each bar in a histogram corresponds to the frequency of the interval it represents.
- Because the individual pieces of data are not visible in a frequency distribution, the minimum and maximum values and the median cannot be determined directly.
- Frequency polygons serve the same purpose as histograms. However, they are especially helpful for comparing multiple sets of data because they can be graphed on top of each other.

CHECK Your Understanding

1. The numbers of earthquakes in the world during two five-year periods are shown in the frequency table.

	Years	
Magnitude	2000–04	2005–09
0.1–0.9	253	5
1.0–1.9	6 957	133
2.0–2.9	28 391	19 120
3.0–3.9	33 717	43 701
4.0–4.9	43 890	58 100
5.0–5.9	6 487	8 948
6.0–6.9	675	770
7.0–7.9	70	61
8.0–9.9	5	8

a) Draw a frequency polygon for the numbers of earthquakes during each five-year period on the same graph.

b) Use your graph to compare the earthquakes in the world during the two five-year periods.

2. Emmanuella walks her golden retriever regularly. She kept track of the lengths of her walks for one month and grouped the data in a frequency table.

a) The first walk was 15 min long. In which interval did she place this piece of data?

b) Draw a frequency polygon to represent the data in the table. Describe how the data is distributed.

Length of Walk (min)	Frequency
5–10	1
10–15	3
15–20	7
20–25	10
25–30	6
30–35	11
35–40	8
40–45	5
45–50	4
50–55	2
55–60	3

PRACTISING

3. A cherry orchard has 30 trees with these heights, given in inches.

78	70	83	79	74	81	80	65	66	76
85	82	74	63	75	76	86	80	72	72
80	69	71	80	77	81	75	75	64	87

a) Make a frequency table with six intervals to organize the heights.
b) Construct a histogram of the data.
c) Which range of heights occurs most frequently? Which occurs least frequently?

4. The amounts withdrawn from an ATM, in dollars, are recorded for a single Wednesday.

20	120	50	70	60	80	140	120	80	160
80	60	110	100	100	80	180	160	40	100
50	80	200	140	160	60	40	80	60	140
100	140	160	200	140	20	80	20	100	70
40	20	120	40	140	100	40	50	180	60

a) What interval width will give a good representation of how the data is distributed?
b) Sort the amounts in a frequency distribution table.
c) Construct a histogram to represent the table in part b).
d) Describe how the data is distributed.

5. The final scores for the 30 women who competed in the women's figure skating competition at the Vancouver 2010 Olympics are shown. Canadian Joannie Rochette captured the bronze medal.

78.50	63.76	61.02	53.16	50.74	43.84
73.78	63.02	59.22	52.96	49.74	43.80
71.36	62.14	57.46	52.16	49.04	41.94
64.76	61.92	57.16	51.74	49.02	40.64
64.64	61.36	56.70	50.80	46.10	36.10

a) Make a frequency table to organize the scores.
b) Draw a histogram of the data.
c) Does your histogram help you see the range of scores that corresponded to a top-five placement? Explain.

Joannie Rochette trains in St-Leonard, Quebec.

6. **a)** On the same graph, draw frequency polygons to show the populations of males and females in Canada for the year 2009.

b) Examine your graph. Describe any differences you notice in the populations of the two sexes.

Population by Gender and Age Group in 2009		
Age Group	**Male (%)**	**Female (%)**
0–4	5.6	5.3
5–9	5.5	5.1
10–14	6.0	5.7
15–19	6.9	6.5
20–24	7.1	6.6
25–29	7.1	6.8
30–34	6.8	6.6
35–39	6.9	6.7
40–44	7.5	7.2
45–49	8.4	8.2
50–54	7.7	7.6
55–59	6.5	6.6
60–64	5.5	5.7
65–69	4.1	4.3
70–74	3.0	3.4
75–79	2.4	2.9
80–84	1.6	2.4
85–89	0.9	1.6
90+	0.3	0.9
Total	99.8%	100.1%

Statistics Canada

7. The following frequency table shows the number of production errors in vehicles coming off an assembly line during the first, second, third, and fourth hour of the day shift.

Number of Errors		1	2	3	4	5	6	7	8
Frequency	**1st Hour**	17	13	8	5	4	3	1	1
	2nd Hour	26	10	6	5	4	2	0	0
	3rd Hour	30	10	5	3	1	1	0	0
	4th Hour	35	8	3	2	1	0	0	0

a) Draw all the frequency polygons for the data on the same graph.

b) What conclusions can you make, based on your graph?

8. Holly and Jason have 14-week training programs to prepare them to run a marathon. On different days during their programs, they run different distances. Holly plans to run the half marathon (21.1 km). Jason plans to run the full marathon (42.2 km). The distances that they run on various training days are shown below.
   a) Construct a frequency distribution table for each training program. Explain the size of interval that you choose.
   b) Use your frequency distribution table to graph a frequency polygon for each runner on the same graph.
   c) Compare the two training programs.

	Holly's Program (km)					Jason's Program (km)					
Week	**Tues.**	**Wed.**	**Thurs.**	**Fri.**	**Sun.**	**Mon.**	**Tues.**	**Wed.**	**Thurs.**	**Fri.**	**Sun.**
1	8	5	8	5	10	off	13	5	13	6	16
2	8	5	10	5	10	off	13	6	13	6	19
3	10	5	13	5	13	off	13	6	16	6	22
4	10	5	13	6	13	5	13	6	16	6	26
5	10	6	13	6	13	6	16	6	16	6	19
6	13	6	13	6	16	6	16	6	16	6	29
7	13	6	13	6	16	6	16	6	19	6	29
8	13	6	13	6	13	6	16	8	19	8	29
9	13	6	16	6	16	6	16	8	19	10	32
10	13	8	13	8	13	8	16	8	16	13	35
11	13	8	16	8	19	6	16	8	16	8	22
12	16	6	16	6	22	10	16	10	19	6	35
13	10	5	13	5	10	5	13	5	10	5	16
14	8	off	8	3	race day	5	10	5	8	3	race day

Serious Injuries in Car Accidents		
Age Group	**Driver (%)**	**Passenger (%)**
0–4	0	5.2
5–14	0.4	10.8
15–24	24.6	33.5
25–34	20.6	12.5
35–44	18.9	9.7
45–54	15.7	8.1
55–64	9.5	6.7
65+	9.9	9.5
not stated	0.4	3.9
Total	100	99.9

9. Examine the given data for injuries, taken from the Canadian Motor Vehicle Traffic Collision Statistics. On the same graph, draw two frequency polygons to illustrate the percent of serious injuries for drivers and for passengers, by age group. Compare the data distributions. What conclusions can you make?

10. For a histogram to display the distribution of data accurately, intervals of equal width must be used. Explain why, using examples.

Closing

11. a) What are the advantages of grouping raw data into intervals?

b) How does a histogram differ from a frequency polygon? When would using frequency polygons be better than using histograms?

Extending

12. This table gives the Aboriginal population in 151 metropolitan areas across Canada, based on statistics from the 2006 census. In 2006, the total Aboriginal population was 1 172 785.

a) Use this table to estimate the mean population. Explain what you did.

b) Use this table to estimate the median population. Explain what you did.

c) Using the actual data, the value for the median population is 1700 and the value for the mean is 4275. How do your estimates compare with these values? Explain any discrepancies.

Aboriginal Populations in Cities, 2006

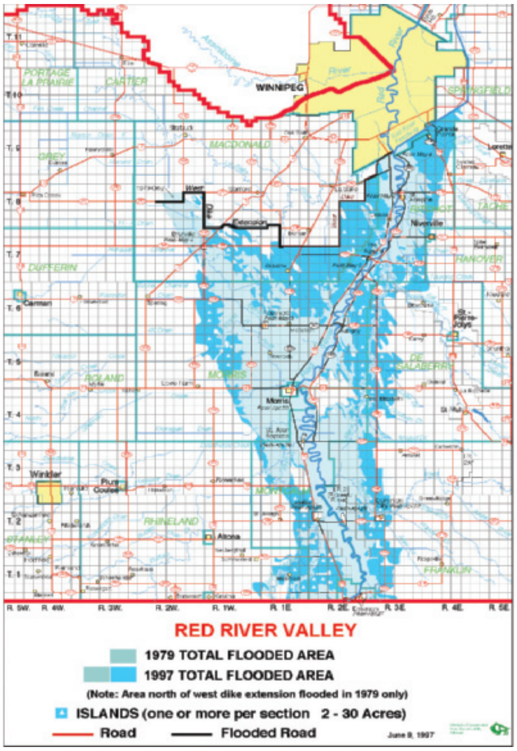
Population	Frequency
0–5 000	122
5 000–10 000	16
10 000–15 000	4
15 000–20 000	2
20 000–25 000	2
25 000–30 000	2
30 000–35 000	0
35 000–40 000	0
40 000–45 000	1
45 000–50 000	0
50 000–55 000	1
55 000–60 000	0
60 000–65 000	0
65 000–70 000	1

Statistics Canada

History Connection

Duff's Ditch

Winnipeg's Red River Floodway was constructed between 1962 and 1968 to protect the city from severe flooding. Affectionately known as Duff's Ditch (after Premier Duff Roblin, who insisted that the project go forward), the floodway is a 47 km channel that diverts water around the city to allow river levels in Winnipeg to remain below flood level. The gates to the floodway are opened whenever the city is threatened by flooding.


WINNIPEG
RED RIVER VALLEY
1979 TOTAL FLOODED AREA
1997 TOTAL FLOODED AREA
(Note: Area north of west dike extension flooded in 1979 only)
ISLANDS (one or more per section 2 - 30 Acres)
Road
Flooded Road
June 9, 1997

A. Flooding starts to happen in Winnipeg when the water level is 5.5 m. This is equivalent to a flow rate of about 1470 m^3/s. When the level gets close to 4.6 m, sandbagging is needed in low-lying areas and the floodway is opened. How many times between its first use in 1969 and 1999 was the floodway opened? (Use the data tables on page 213.)

B. Do you think that high flow rates are becoming the norm rather than the exception? Refer to the data tables on page 213, and do research to find more recent data. Justify your answer.

5.3 Standard Deviation

YOU WILL NEED

- calculator OR computer with spreadsheet software

GOAL

Determine the standard deviation for sets of data, and use it to solve problems and make decisions.

EXPLORE…

- A teacher has two chemistry classes. She gives the same tests to both classes. Examine the mean mark for each of the first five tests given to both classes. Compare the results for the two classes.

Test	Class A (%)	Class B (%)
1	94	84
2	56	77
3	89	76
4	67	81
5	84	74

INVESTIGATE the Math

The coach of a varsity girls' basketball team keeps statistics on all the players. Near the end of one game, the score is tied and the best starting guard has fouled out. The coach needs to make a substitution. The coach examines the field goal stats for five guards on the bench in the last 10 games.

Player	Field Goal Percent in Last 10 Basketball Games									
Anna	36	41	43	39	45	27	40	37	31	28
Patrice	36	39	36	38	35	37	35	36	38	34
Morgan	34	41	38	37	48	19	33	43	21	44
Paige	34	35	33	35	33	34	33	35	34	33
Star	41	33	39	36	38	36	29	34	38	39

? How can the coach use the data to determine which player should be substituted into the game?

A. Which player seems to be the most consistent shooter? Explain.

B. Analyze the data for Paige using a table like the one shown on the next page. Determine the mean of the data, $\bar{x}$, for Paige, and record this value in the first column.

C. Complete the second column for the **deviation** of each field goal percent: $(x - \bar{x})$

D. Complete the third column for the squares of the deviations.

Communication ***Tip***

The symbol $\bar{x}$ (read as "x bar") represents the mean of the data.

deviation

The difference between a data value and the mean for the same set of data.

Paige's Field Goal (%)	Deviation $(x - \overline{x})$	Square of Deviation $(x - \overline{x})^2$
34	0.1	0.01
35	1.1	1.21
33		

E. Determine the **standard deviation** of Paige's data by following these steps:

Step 1: Determine the mean of the squares of the deviations.

Step 2: Determine the square root of the mean from Step 1. This number is the standard deviation.

standard deviation

A measure of the dispersion or scatter of data values in relation to the mean; a low standard deviation indicates that most data values are close to the mean, and a high standard deviation indicates that most data values are scattered farther from the mean.

F. Analyze all the data using a spreadsheet like the one below. Enter the field goal percent for Anna in row 2, and for Patrice, Morgan, Paige, and Star in rows 3, 4, 5, and 6, as shown.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Game Player	1	2	3	4	5	6	7	8	9	10	Mean	Standard Deviation
2	Anna	36	41	43	39	45	27	40	37	31	28		
3	Patrice	36	39	36	38	35	37	35	36	38	34		
4	Morgan	34	41	38	37	48	19	33	43	21	44		
5	Paige	34	35	33	35	33	34	33	35	34	33		
6	Star	41	33	39	36	38	36	29	34	38	39		

G. Using the features of the spreadsheet software, determine the mean and standard deviation for each set of data.

H. Compare your result from part E with your results for Paige from part G. What do you notice?

I. Examine the means of the players. Would you use the most consistent player identified in part A as a substitute? Explain.

J. Which player has the greatest percent range? Which player has the least? How do the standard deviations of these players compare?

K. Compare the means and standard deviations of the data sets for all the players. Which player's data has the lowest standard deviation? What does this imply about her shooting consistency?

L. Based on past performance, which player has the potential to shoot most poorly? Which player has the potential to shoot most successfully?

M. If you were the coach, which player would you substitute into the game? Explain why.

Reflecting

N. The mean, $\overline{x}$, can be expressed using symbols:

$$\overline{x} = \frac{\Sigma x}{n}$$

Based on your understanding of the mean, what does the symbol Σ represent?

Communication | ***Tip***

The symbol $\boldsymbol{\sigma}$ (read as "sigma") represents the standard deviation of the data.

O. The standard deviation, $\boldsymbol{\sigma}$, can also be expressed using symbols:

$$\sigma = \sqrt{\frac{\Sigma(x - \overline{x})^2}{n}}$$

Interpret this expression verbally.

P. Standard deviation is a measure of dispersion, of how the data in a set is distributed. How would a set of data with a low standard deviation differ from a set of data with a high standard deviation?

APPLY the Math

EXAMPLE 1 Using standard deviation to compare sets of data

Brendan works part-time in the canteen at his local community centre. One of his tasks is to unload delivery trucks. He wondered about the accuracy of the mass measurements given on two cartons that contained sunflower seeds. He decided to measure the masses of the 20 bags in the two cartons. One carton contained 227 g bags, and the other carton contained 454 g bags.

Masses of 227 g Bags (g)			
228	220	233	227
230	227	221	229
224	235	224	231
226	232	218	218
229	232	236	223

Masses of 454 g Bags (g)			
458	445	457	458
452	457	445	452
463	455	451	460
455	453	456	459
451	455	456	450

How can measures of dispersion be used to determine if the accuracy of measurement is the same for both bag sizes?

Brendan's Solution

227 g bags:
$\text{Range} = 236\text{ g} - 218\text{ g}$
$\text{Range} = 18\text{ g}$

454 g bags:
$\text{Range} = 463\text{ g} - 445\text{ g}$
$\text{Range} = 18\text{ g}$

I examined each set of data and found the greatest mass and least mass. Then I determined the range. The range was the same for both cartons.

227 g bags:
$\bar{x} = 227.15\text{ g}$
$\sigma = 5.227...\text{ g}$

454 g bags:
$\bar{x} = 454.4\text{ g}$
$\sigma = 4.498...\text{ g}$

I used my graphing calculator to determine the mean and standard deviation for each set of data.

Both means were above the mass measurements given on the two cartons.

The accuracy of measurement is not the same for both sizes of bag.

The standard deviation for the 454 g bags is less than the standard deviation for the 227 g bags.

The difference in the standard deviations indicates that the masses of the larger bags were closer to their mean mass.

Therefore, the 454 g bags of sunflower seeds have a more consistent mass.

Your Turn

a) Explain why the standard deviations for the masses of the two sizes of bag are different, even though the ranges of the masses are the same.
b) How might standard deviation be used by the company that sells the sunflower seeds for quality control in the packaging process?

EXAMPLE 2 Determining the mean and standard deviation of grouped data

Angèle conducted a survey to determine the number of hours per week that Grade 11 males in her school play video games. She determined that the mean was 12.84 h, with a standard deviation of 2.16 h.

Janessa conducted a similar survey of Grade 11 females in her school. She organized her results in this frequency table. Compare the results of the two surveys.

Gaming Hours per Week for Grade 11 Females

Hours	Frequency
3–5	7
5–7	11
7–9	16
9–11	19
11–13	12
13–15	5

Cole's Solution: Determining $\bar{x}$ and σ manually

Note: The purpose of Cole's Solution is to provide an understanding of what technology does to calculate the mean and standard deviation when working with grouped data. Students are not expected to determine mean and standard deviation manually.

An estimate for the mean of the gaming hours for Grade 11 females is 9.

I predicted that the mean is about 9 h because most of the data is in the 7 to 11 intervals. However, I need to verify my estimate.

A	B	C	D
Hours	Frequency (f)	Midpoint of Interval (x)	$f \cdot x$
3–5	7	4	28
5–7	11	6	66
7–9	16	8	128
9–11	19	10	190
11–13	12	12	144
13–15	5	14	70
	70		626

I didn't know the actual values in each interval, so I determined the midpoint of each interval. I knew that some values would be greater than the midpoint and some values would be less, but I thought that the midpoint could represent all the values in each interval.

I multiplied the frequency by the midpoint for each interval to determine the number of hours in each interval.

$$\bar{x} = \frac{\Sigma(f)(x)}{n}$$

$$\bar{x} = \frac{626}{70}$$

$$\bar{x} = 8.942\ldots \text{ h}$$

Next, I determined the mean for the data set. I divided the total number of hours by the total number of data values.

C	D	E	F
Midpoint of Interval (x)	$f \cdot x$	$(x - \bar{x})^2$	$f \cdot (x - \bar{x})^2$
4	28	24.431...	171.022...
6	66	8.660...	95.264...
8	128	0.888...	14.223...
10	190	1.117...	21.233...
12	144	9.346...	112.153...
14	70	25.574...	127.873...
	626		541.771...
$\bar{x}$ = 8.942... h			

In column E, I squared the deviation from the midpoint for each interval.

In column F, I multiplied each squared value by the frequency, and I added these products to estimate the total square deviations for all the data.

$$\sigma = \sqrt{\frac{\Sigma f(x - \overline{x})^2}{n}}$$

$$\sigma = \sqrt{\frac{541.771...}{70}}$$

$$\sigma = 2.782... \text{ h}$$

I determined the standard deviation by dividing the sum of the squares of the deviations by the number of data values and then taking the square root.

Males:	Females:
$\overline{x} = 12.84$ h	$\overline{x} = 8.942...$ h
$\sigma = 2.16$ h	$\sigma = 2.782...$ h

I compared the results for the two groups.

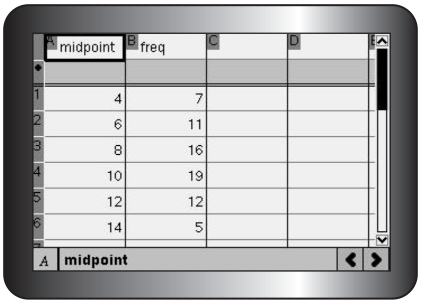
The males played nearly 4 h more per week than the females, on average.

The standard deviation for males is lower than the standard deviation for females. Therefore, the males' playing times are closer to their mean (almost 13 h) and don't vary as much.

The mean playing time for males is higher than the mean playing time for females.

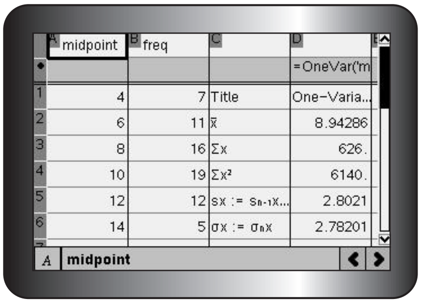
The data for the females is more dispersed than the data for the males.

Danica's Solution: Using technology to determine $\overline{x}$ and σ



	A midpoint	B freq	C	D
◆				
1	4	7		
2	6	11		
3	8	16		
4	10	19		
5	12	12		
6	14	5		

First, I determined the midpoints of the intervals for Janessa's data and entered these values in one list on my calculator. Then I entered the frequency in a second list.



	A midpoint	B freq	C	D
◆				=OneVar('m
1	4	7	Title	One−Varia...
2	6	11	$\overline{x}$	8.94286
3	8	16	Σx	626.
4	10	19	Σx²	6140.
5	12	12	sx := sn-1x...	2.8021
6	14	5	σx := σnx	2.78201

I determined the mean and the standard deviation.

Gaming hours per week for Grade 11 females:

$\overline{x} = 8.942...$ h $\qquad \sigma = 2.782...$ h

Gaming hours per week for Grade 11 males:

$\overline{x} = 12.84$ h $\qquad \sigma = 2.16$ h

I compared the results for the two groups.

The standard deviation for the females is higher than the standard deviation for the males.

Therefore, the females' times vary more from their mean of about 9 h.

The standard deviation for the males is lower. Therefore, their data is more consistent, even though their mean is higher.

The females, on average, spent less time playing video games per week. However, some females played a lot more or a lot less than the mean.

The males played more hours per week, and their times were fairly close to their mean.

Your Turn

Could the mean and standard deviation for the female data differ from those determined by Danica and Cole, if the actual data is used? Explain.

In Summary

Key Ideas

- To determine how scattered or clustered the data in a set is, determine the mean of the data and compare each data value to the mean.
- The standard deviation, σ, is a measure of the dispersion of data about the mean.
- The mean and standard deviation can be determined using technology for any set of numerical data, whether or not the data is grouped.

Need to Know

- When data is concentrated close to the mean, the standard deviation, σ, is low. When data is spread far from the mean, the standard deviation is high. As a result, standard deviation is a useful statistic to compare the dispersion of two or more sets of data.
- When determining the standard deviation, σ, for a set of data using technology, this is the process that is followed:
  1. The square of the deviation of each data value (or the midpoint of the interval) from the mean is determined: $(x - \overline{x})^2$
  2. The mean of the squared deviations of all the data values is determined.
  3. The square root of the mean from step 2 is determined. This value is the standard deviation.
- Standard deviation is often used as a measure of consistency. When data is closely clustered around the mean, the process that was used to generate the data can be interpreted as being more consistent than a process that generated data scattered far from the mean.

CHECK *Your Understanding*

1. a) Determine, by hand, the standard deviation of test marks for the two chemistry classes shown.
   b) Verify your results from part a) using technology.
   c) Which class had the more consistent marks over the first five tests? Explain.

Test	Class A (%)	Class B (%)
1	94	84
2	56	77
3	89	76
4	67	81
5	84	74

Use technology to determine the mean and standard deviation, as needed, in questions 2 to 14.

2. Ali bowls in a peewee league. Determine the mean and standard deviation of Ali's bowling scores, rounded to two decimal places.

135	156	118	133
141	127	124	139
109	131	129	123

3. The bowling scores for the six players on Ali's team are shown at the right.
   a) Determine the mean and standard deviation of the bowling scores for Ali's team, rounded to two decimal places.
   b) Using the mean and standard deviation, compare Ali's data from question 2 to the team's data.

Bowling Scores	Frequency
101–105	1
106–110	3
111–115	4
116–120	7
121–125	9
126–130	14
131–135	11
136–140	8
141–145	6
146–150	5
151–155	3
156–160	1

4. Marie, a Métis beadwork artist, ordered packages of beads from two online companies. She is weighing the packages because the sizes seem inconsistent. The standard deviation of the masses of the packages from company A is 11.7 g. The standard deviation of the masses of the packages from company B is 18.2 g.
   a) What does this information tell you about the dispersion of the masses of the packages from each company?
   b) Marie is working on an important project. She needs to make sure that her next order will contain enough beads to complete the project. Should she order from company A or company B?

The Métis are known for floral beadwork. The symmetrical traditional beadwork is illustrated here on the deerskin coat of Louis Riel. Seeds were used to create the beads.

PRACTISING

5. Four groups of students recorded their pulse rates, as given below.

Group 1	63	78	79	75	73	72	62	75	63	77	77	65	70	69	80
Group 2	72	66	73	80	74	75	64	68	67	70	70	69	69	74	74
Group 3	68	75	78	73	75	68	71	78	65	67	63	69	59	68	79
Group 4	78	75	76	76	79	78	78	76	74	81	78	76	79	74	76

Determine the mean and standard deviation for each group, to one decimal place. Which group has the lowest mean pulse rate? Which group has the most consistent pulse rate?

6. Nazra and Diko are laying patio stones. Their supervisor records how many stones they lay each hour.

Hour	1	2	3	4	5	6
Nazra	34	41	40	38	38	45
Diko	51	28	36	44	41	46

a) Which worker lays more stones during the day?
b) Which worker is more consistent?

Milt Stegall

7. Former Winnipeg Blue Bomber Milt Stegall broke several Canadian Football League (CFL) records, including the most touchdowns (TDs) in a season and the most TDs in a career.

Year	'95	'96	'97	'98	'99	'00	'01	'02	'03	'04	'05	'06	'07	'08
TDs	4	6	14	7	7	15	14	23	15	7	17	7	8	3

a) Determine the mean and standard deviation of the TDs that Milt scored in the years he played, to one decimal place.
b) Why do you think that his first and last years had the lowest number of TDs?
c) Determine the mean and standard deviation, to one decimal place, for the years 1996 to 2007.
d) Compare your results from parts a) and c). What do you notice?

8. Milt Stegall also broke the CFL record for most yards receiving.
a) Determine the mean and standard deviation of his statistics, to one decimal place.

Year	'95	'96	'97	'98	'99	'00	'01	'02	'03	'04	'05	'06	'07	'08
Yards	469	613	1616	403	1193	1499	1214	1862	1144	1121	1184	1252	1108	470

b) Allen Pitts, who played for the Calgary Stampeders, held the CFL record for most yards receiving until Milt Stegall surpassed him. Pitts had the following statistics for yards gained per year: mean 1353.7 and standard deviation 357.1. Which player was more consistent in terms of yards gained per year?

9. Two health clubs monitor the number of hours per month that a random sample of their members spend working out.

Fitness Express	
Hours	**Frequency**
8–10	9
10–12	18
12–14	23
14–16	32
16–18	39
18–20	42
20–22	31
22–24	22
24–26	16
26–28	11
28–30	7

Fit for Life	
Hours	**Frequency**
6–9	8
9–12	13
12–15	32
15–18	47
18–21	52
21–24	42
24–27	27
27–30	19

a) Determine the mean and standard deviation of the hours per month for members of each club, to one decimal place.

b) The health clubs believe that workout consistency is more important than workout length. Which club is more successful at encouraging its members to work out consistently?

10. Jaime has 20 min to get to her after-school job. Despite her best efforts, she is frequently late. Her employer says that unless she arrives to work on time consistently, she will lose her job. She has recorded her travel times (in minutes) for the last two weeks: 18, 20, 22, 27, 16, 23, 25, 26, 19, 28. Over the next two weeks, she continues to record her travel times: 22, 20, 19, 16, 20, 23, 25, 18, 19, 17. Do you think Jaime will lose her job? Use statistics to justify your answer.

11. The manager of a customer support line currently has 200 unionized employees. Their contract states that the mean number of calls that an employee should handle per day is 45, with a maximum standard deviation of 6 calls. The manager tracked the number of calls that each employees handles. Does the manager need to hire more employees if the calls continue in this pattern?

Daily Calls	Frequency
26–30	2
31–35	13
36–40	42
41–45	53
46–50	42
51–55	36
56–60	8
61–65	4

Jordin Tootoo, from Rankin Inlet, Nunavut, played for the Brandon Wheat Kings of the WHL from 1999 to 2003. In 2003, he became the first player of Inuit descent to play in a regular-season NHL game.

12. The following table shows Jordin Tootoo's regular season statistics while playing in the Western Hockey League (WHL) and the NHL from 1999 to 2010.

Season	Games Played	Goals	Assists	Points
1999–2000	45	6	10	16
2000–2001	60	20	28	48
2001–2002	64	32	39	71
2002–2003	51	35	39	74
2003–2004	70	4	4	8
2005–2006	34	4	6	10
2006–2007	65	3	6	9
2007–2008	63	11	7	18
2008–2009	72	4	12	16
2009–2010	51	6	10	16

a) Determine the mean and standard deviation for each column of data.

b) In the 2005–2006 season, Jordin played only 34 games. As a result, he had fewer opportunities to score points. Predict the effect on the standard deviation of each column if this season is omitted.

c) Determine the mean and standard deviation for each column, excluding the 2005–2006 season.

d) How do your results for part c) compare with your results for part a)?

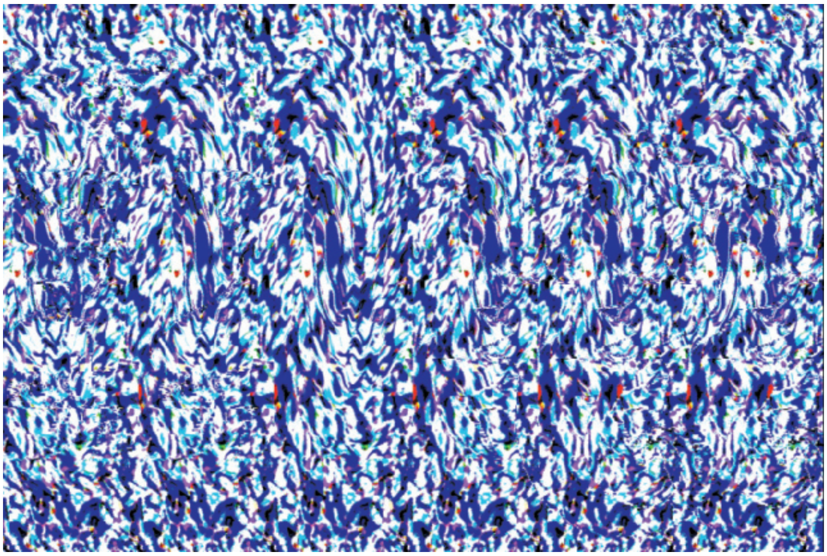
e) Justify the following statement: "Goals + assists = points, whether you are looking at the data for each season or you are looking at the means and standard deviations for many seasons."

Closing

13. Twins Jordana and Jane wrote a total of 10 tests in math class. They have the same mean test score, but different standard deviations. Explain how this is possible.

Extending

14. If you glance at a random dot stereogram (also called a "Magic Eye image"), it looks like a collection of dots or shapes. If you look at it properly, however, it resolves into a 3-D image.



©2010 Magic Eye Inc.

An experiment was done to determine whether people could "fuse" the image faster if they knew what shape they were looking for. The results of the experiment are shown in the table to the right. The people in group A were given no information about the image. The people in group B were given visual information about the image. The number of seconds that the people needed to recognize the 3-D image are listed in the table.

a) Determine the mean and standard deviation for each group, rounded to the nearest hundredth of a second.

b) Were the people who were given visual information able to recognize the image more quickly? Which group was more consistent?

Times for Group A (s)		Times for Group B (s)	
47.2	5.6	19.7	3.6
22.0	4.7	16.2	3.5
20.4	4.7	15.9	3.3
19.7	4.3	15.4	3.3
17.4	4.2	9.7	2.9
14.7	3.9	8.9	2.8
13.4	3.4	8.6	2.7
13.0	3.1	8.6	2.4
12.3	3.1	7.4	2.3
12.2	2.7	6.3	2.0
10.3	2.4	6.1	1.8
9.7	2.3	6.0	1.7
9.7	2.3	6.0	1.7
9.5	2.1	5.9	1.6
9.1	2.1	4.9	1.4
8.9	2.0	4.6	1.2
8.9	1.9	1.0	1.1
8.4	1.7	3.8	
8.1	1.7		
7.9	6.9		
7.8	6.3		
6.1			

5 Mid-Chapter Review

FREQUENTLY ASKED Questions

Study | ***Aid***

- See Lesson 5.1 and Lesson 5.2, Examples 1 and 2.
- Try Mid-Chapter Review Questions 1 to 3.

Q: How can you make a large set of data more manageable to work with and analyze?

A: Create a frequency distribution by organizing the data into equal-sized intervals. This simplifies the data into a manageable number of intervals, which show how the data is distributed. To visualize the data in the frequency table more easily, draw a histogram or a frequency polygon.

For example, consider the weight, in pounds, of the members of the Edmonton Eskimos CFL team.

165	185	188	195	205	210	225	235	250	270	300	320
170	185	189	196	210	215	225	235	255	300	302	335
175	185	191	197	210	215	229	240	255	300	310	
183	188	195	205	210	216	230	248	265	300	315	

First, determine the lowest weight and highest weight, and the range.

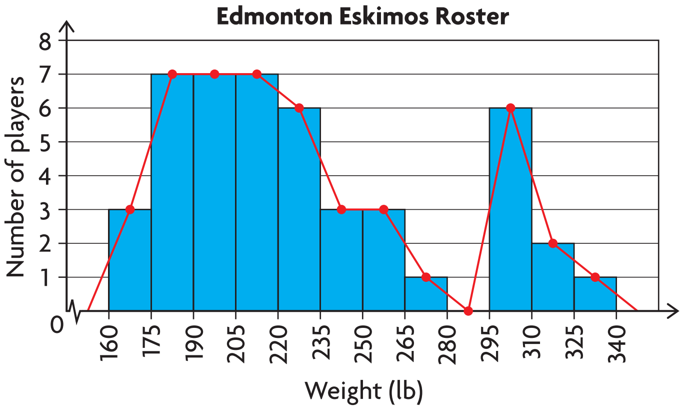
Range = 335 − 165 or 170 lb

Next, decide on the number of intervals and the interval width. Generally, the number of intervals is from 5 to 12. Since the range is 170, an interval width of 15 would require 12 intervals.

The first interval should start slightly below the lowest weight. The last interval should end slightly above the highest weight. Use a tally to count the number of players in each interval.

Weight	Frequency
160–175	3
175–190	7
190–205	7
205–220	7
220–235	6
235–250	3
250–265	3
265–280	1
280–295	0
295–310	6
310–325	2
325–340	1

Draw the graph. To plot each vertex for the frequency polygon, determine the midpoint of the interval. Join the vertices with line segments, and then connect the endpoints to the horizontal axis.



Q: How can you determine standard deviation using technology?

A1: For ungrouped data, enter the data into a list. Use the appropriate statistics formula to determine the standard deviation.

A2: For grouped data, determine the midpoint of each interval using a graphing calculator or spreadsheet. Enter the midpoints into column 1 and the frequencies into column 2. Use the appropriate functions to determine the standard deviation.

Study *Aid*

- See Lesson 5.3, Examples 1 and 2.
- Try Mid-Chapter Review Questions 4 to 6.

Q: What does standard deviation measure, and how do you interpret it?

A: Standard deviation measures the dispersion of a data set. It is an indication of how far away most of the data is from the mean. It is useful for comparing two or more sets of data.

For example, consider two different health clubs.

Health Club	Mean Age of Members (years)	Standard Deviation of Ages (years)
A	37	5.3
B	43	11.4

The standard deviation for Health Club B is more than double the standard deviation for Health Club A. This means that there is a much wider range of ages in Health Club B, and the ages are less clustered around the mean.

Study *Aid*

- See Lesson 5.3.
- Try Mid-Chapter Review Questions 1 to 3.

PRACTISING

Lesson 5.1

1. Compare the mean monthly temperatures, in degrees Celsius, for Paris, France, and Sydney, Australia.

	Jan.	Feb.	Mar.	Apr.	May	Jun.	Jul.	Aug.	Sept.	Oct.	Nov.	Dec.
Paris	6	7	12	16	20	23	25	24	21	16	10	7
Sydney	23	23	21	19	15	13	12	13	15	18	20	22

Lesson 5.2

2. Construct a graph for Wayne Gretzky's goals per NHL season. Describe the data distribution.

Goals	Number of Seasons
0 to 10	2
10 to 20	3
20 to 30	3
30 to 40	2
40 to 50	3
50 to 60	4
60 to 70	1
70 to 80	2
80 to 90	1
90 to 100	1

3. Jackson and Jillian are trying to control the number of text messages they send. They record the number they send every day in April.

Jackson: 2, 7, 20, 4, 11, 25, 6, 27, 3, 6, 18, 5, 13, 4, 10, 16, 23, 22, 5, 8, 3, 12, 6, 13, 12, 7, 8, 26, 9, 17
Jillian: 2, 9, 11, 15, 8, 8, 0, 21, 16, 12, 14, 14, 15, 20, 11, 12, 10, 9, 8, 0, 7, 24, 19, 18, 19, 15, 12, 8, 1, 13

a) Choose an interval width.
b) Create a frequency table for the data.
c) Compare the two sets of data using a frequency polygon.

Lesson 5.3

4. Compare the two sets of data in question 3 by determining the means and standard deviations, to the nearest tenth.

5. Liam keeps track of the amount he spends, in dollars, on weekly lunches during one semester:

19	15	6	24	27	26	48
19	23	18	29	17	14	22
19	26	20	17	28		

a) Determine the range, mean, and standard deviation, correct to two decimal places.
b) Remove the greatest and the least weekly amounts. Then determine the mean, standard deviation, and range for the remaining amounts.
c) What effect does removing the greatest and least amounts have on the standard deviation?

6. Tiffany researched the annual salaries of males versus females for her project on gender issues. She obtained the following data about 441 full-time employees who work in laboratories. Determine the mean and standard deviation for each set of data, and compare the data.

Females	
Salary Range ($)	**Frequency**
20 000–25 000	92
25 000–30 000	52
30 000–35 000	19
35 000–40 000	10
40 000–45 000	4
45 000–50 000	1
50 000–55 000	3
55 000–60 000	3

Males	
Salary Range ($)	**Frequency**
20 000–30 000	86
30 000–40 000	78
40 000–50 000	28
50 000–60 000	20
60 000–70 000	22
70 000–80 000	10
80 000–90 000	4
90 000–100 000	5
100 000–110 000	2
110 000–120 000	1
120 000–130 000	0
130 000–140 000	1

5.4 The Normal Distribution

GOAL

Determine the properties of a normal distribution, and compare normally distributed data.

YOU WILL NEED

- calculator
- grid paper
- dice
- dice simulator

INVESTIGATE the Math

Many games require dice. For example, the game of Yacht requires five dice.

? What shape is the data distribution for the sum of the numbers rolled with dice, using various numbers of dice?

A. If you rolled a single die 50 000 times, what do you think the graph would look like?

B. Predict what the graph would look like if you rolled two dice 50 000 times.

C. With a partner, roll two dice 50 times. Record the sum for each roll in a frequency distribution table. Then draw a graph to represent the distribution of the data. Comment on the distribution of the data.

D. Combine the data for the entire class, and draw a graph. Comment on the distribution of the combined data.

E. Using a dice simulator, roll two dice 50 000 times. Compare the graph for this set of data with the graph you drew in part D.

F. Make a conjecture about what the graph would look like if you rolled three dice 50 000 times.

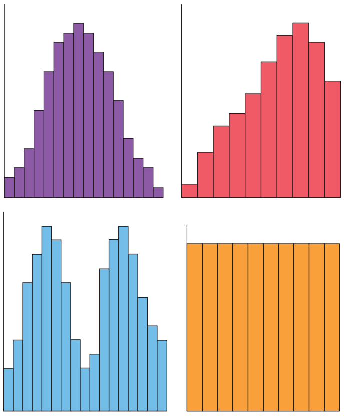
G. Using a dice simulator, roll three dice 50 000 times. What do you notice about the shape of the graph as the number of rolls increases? Draw a frequency polygon to represent the distribution of the data. Was your conjecture correct?

H. Make a conjecture about what the graph would look like if you rolled four dice 50 000 times. What would the graph look like for five dice rolled 50 000 times?

I. Using a dice simulator, roll four dice and then five dice 50 000 times. What do you notice about the shape of the graph as the number of rolls increases? Draw a frequency polygon to represent the distribution of the data for both four dice and five dice. Describe the shape of each polygon. Was your conjecture correct?

EXPLORE...

- Sometimes the distribution of data has a special shape. For example, the first graph below has one peak, so the shape has one mode. Describe the shape of each graph, and suggest a context that the graph could represent.



normal curve

A symmetrical curve that represents the normal distribution; also called a **bell curve**.

normal distribution

Data that, when graphed as a histogram or a frequency polygon, results in a unimodal symmetric distribution about the mean.

Reflecting

J. How does increasing the number of dice rolled each time affect the distribution of the data?

K. How does increasing the sample size affect the distribution of the data for three dice, four dice, and five dice?

L. What do you think a graph that represents 100 000 rolls of 10 dice would look like? Why do you think this shape is called a **normal curve**?

M. As the number of dice increases, the graph approaches a **normal distribution**. What does the line of symmetry in the graph represent?

APPLY the Math

EXAMPLE 1 Examining the properties of a normal distribution

Heidi is opening a new snowboard shop near a local ski resort. She knows that the recommended length of a snowboard is related to a person's height. Her research shows that most of the snowboarders who visit this resort are males, 20 to 39 years old. To ensure that she stocks the most popular snowboard lengths, she collects height data for 1000 Canadian men, 20 to 39 years old. How can she use the data to help her stock her store with boards that are the appropriate lengths?

Height (in.)	Frequency
61 or shorter	3
61–62	4
62–63	10
63–64	18
64–65	30
65–66	52
66–67	64
67–68	116
68–69	128
69–70	147
70–71	129
71–72	115
72–73	63
73–74	53
74–75	29
75–76	20
76–77	12
77–78	5
taller than 78	2

Heidi's Solution

$\bar{x} = 69.521$ in.
$\sigma = 2.987...$ in.

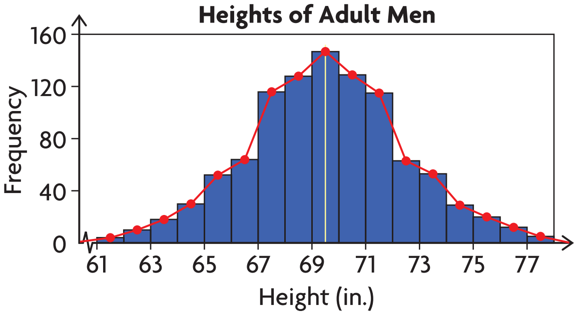
I decided to use 60.5 as the midpoint of the first interval and 78.5 as the midpoint of the last interval. I calculated the mean and standard deviation using a graphing calculator.

There are 147 heights in the 69–70 in. interval. The 73rd person is about midway through this interval, so the median height is approximately 69.5 in.

I examined the table. The median is the average of the 500th and the 501st height.

There are 425 heights less than or equal to 69 in. The median height is the height of the 73rd person in the 69–70 in. interval.

I don't know where the mode is, but it is more likely to be in the interval that contains the greatest number of data values, which is the 69–70 in. interval. I can assume that all three measures of central tendency have about the same value.



I drew a histogram to show the height distribution.

I drew a vertical line on my histogram to represent the mean.

Then I drew a frequency polygon by connecting the midpoints at the top of each bar of the graph.

The data is almost symmetrical about the mean and tapers off in a gradual way on both sides. The frequency polygon resembles a bell shape.

The data has a normal distribution.

Heights within one standard deviation of the mean:
$69.521 - 2.987...$ or 66.5 in.
$69.521 + 2.987...$ or 72.5 in.

Heights within two standard deviations of the mean:
$69.521 - 2(2.987...)$ or 63.5 in.
$69.521 + 2(2.987...)$ or 75.5 in.

I determined the range of heights within one standard deviation of the mean by adding and subtracting the standard deviation from the mean.

I also determined the range of heights within two standard deviations of the mean.

Range	Height Range
$\bar{x} - 1\sigma$ to $\bar{x} + 1\sigma$	about 66.5 in. to 72.5 in.
$\bar{x} - 2\sigma$ to $\bar{x} + 2\sigma$	about 63.5 in. to 75.5 in.

I summarized my results in a table and compared them to my histogram. The range for one standard deviation appears to include most of the data. The range for two standard deviations appears to include almost all of the data.

Number of males within one standard deviation of the mean from 67 in. to 73 in.:
$116 + 128 + 147 + 129 + 115 + 63$, or 698

I estimated the percent of the heights that were within one and two standard deviations of the mean.

To determine the number of heights in each range, I rounded up to find the lower and upper boundaries in each range. Then I summed the number of people from the table that were in these ranges.

Number of males within two standard deviations of the mean from 64 in. to 76 in.:
$30 + 52 + 64 + 116 + 128 + 147 + 129 + 115 + 63 + 53 + 29 + 20$, or 944

Range	Height Range	Percent of Data
$\bar{x} - 1\sigma$ to $\bar{x} + 1\sigma$	66.5 in. to 72.5 in.	$\frac{698}{1000}$ or 69.8%
$\bar{x} - 2\sigma$ to $\bar{x} + 2\sigma$	63.5 in. to 75.5 in.	$\frac{946}{1000}$ or 94.6%

About 70% of the heights are within one standard deviation of the mean.

About 95% of the heights are within two standard deviations of the mean.

Heights within one standard deviation of the mean are most common.

A high percent of heights are within two standard deviations of the mean.

I predict that about 70% of my male customers will need snowboards for heights from 66.5 to 72.5 in.

Your Turn

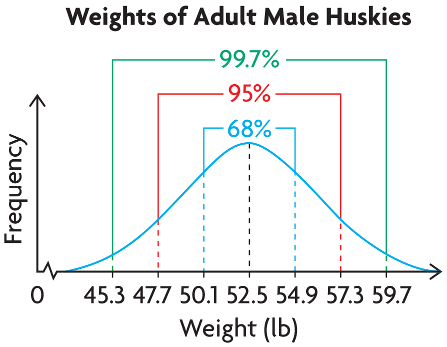
What percent of all the heights is within three standard deviations of the mean?

EXAMPLE 2 Analyzing a normal distribution

Jim raises Siberian husky sled dogs at his kennel. He knows, from the data he has collected over the years, that the weights of adult male dogs are normally distributed, with a mean of 52.5 lb and a standard deviation of 2.4 lb. Jim used this information to sketch a normal curve, with

- 68% of the data within one standard deviation of the mean
- 95% of the data within two standard deviations of the mean
- 99.7% of the data within three standard deviations of the mean

The Canadian Championship Dog Derby, held in Yellowknife, Northwest Territories, is one of the oldest sled-dog races in North America. Top mushers gather to challenge their dogs in the fast-paced, three-day event.

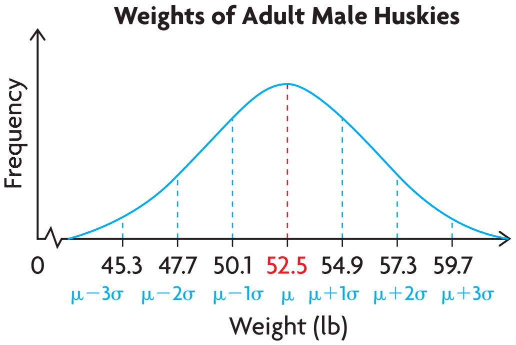


What percent of adult male dogs at Jim's kennel would you expect to have a weight between 47.7 lb and 54.9 lb?

Communication *Tip*

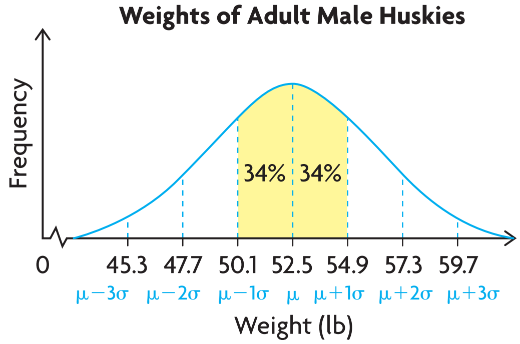
In statistics, when an entire population is involved, use the symbol μ (read as "mu") for the mean of the population.

Ian's Solution

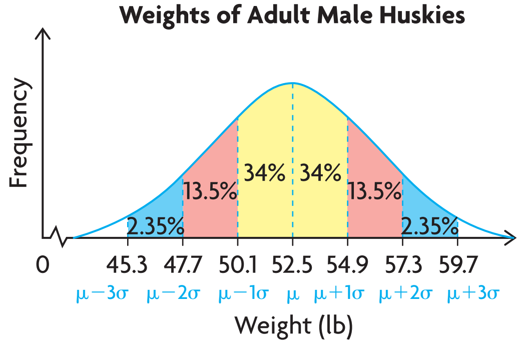


I sketched the graph and labelled the mean, μ, below the horizontal axis. Since the standard deviation is 2.4 lb, I can label the scale to the right of the mean as $\mu + 1\sigma$, $\mu + 2\sigma$, and $\mu + 3\sigma$.

I labelled the scale to the left of the mean as $\mu - 1\sigma$, $\mu - 2\sigma$, and $\mu - 3\sigma$.



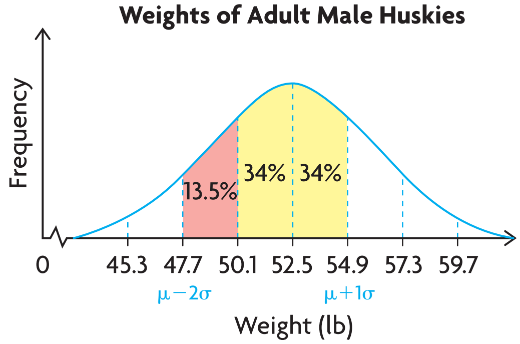
I reasoned that the area under the curve is symmetrical around μ, so if 68% of the dogs have weights within one standard deviation, then 34% must have weights between $\mu - 1\sigma$ and μ, and 34% must have weights between μ and $\mu + 1\sigma$.



I knew that 95% of the weights lie within two standard deviations of the mean. Since 68% of the weights lie within one standard deviation of the mean, 27% of the weights must lie between one and two standard deviations, or 13.5% for each side of the graph.

Using the same reasoning, I figured out the percent of data that would lie between two and three standard deviations from the mean for each side of the graph:

$$\frac{99.7\% - 95\%}{2} = 2.35\%$$



For the percent that fits between 47.7 lb and 54.9 lb, I determined the location of each weight.

$\mu + 1\sigma = 54.9$

$\mu - 2\sigma = 47.7$

I used my diagram to determine the sum of the percent of data between these locations.

The percent of dogs with weights between 47.7 lb and 54.9 lb, x, can be represented as

$$x = 13.5\% + 34\% + 34\%$$
$$x = 81.5\%$$

Approximately 81.5% of adult male dogs should have a weight between 47.7 lb and 54.9 lb.

Your Turn

a) What percent of adult male dogs at Jim's kennel would you expect to have a weight between 50.1 lb and 59.7 lb?

b) What percent of adult male dogs at Jim's kennel would you expect to have a weight less than 45.3 lb?

EXAMPLE 3 Comparing normally distributed data

Team	μ (kg)	σ (kg)
Men	6.35	1.04
Women	6.35	0.59

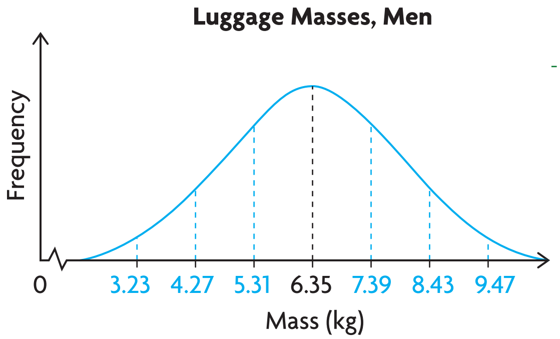
Two baseball teams flew to the North American Indigenous Games. The members of each team had carry-on luggage for their sports equipment. The masses of the carry-on luggage were normally distributed, with the characteristics shown to the right.

a) Sketch a graph to show the distribution of the masses of the luggage for each team.

b) The women's team won the championship. Each member received a medal and a souvenir baseball, with a combined mass of 1.18 kg, which they packed in their carry-on luggage. Sketch a graph that shows how the distribution of the masses of their carry-on luggage changed for the flight home.

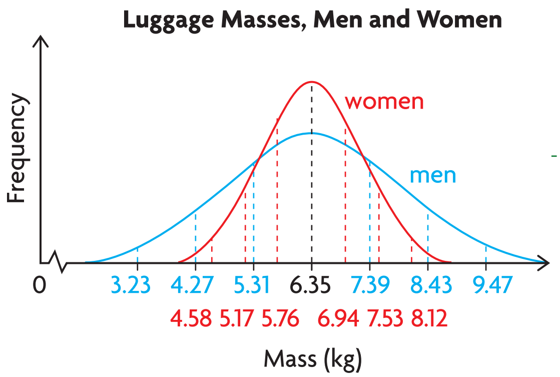
Samara's Solution

a)



I sketched the normal distribution of the masses of the luggage for the men's team. I marked the values for μ, $\mu + 1\sigma$, $\mu + 2\sigma$, $\mu + 3\sigma$, $\mu - 1\sigma$, $\mu - 2\sigma$, and $\mu - 3\sigma$.

I knew that the area under the normal curve represents 100% of the data, so I could think of the area as equal to 1 unit.



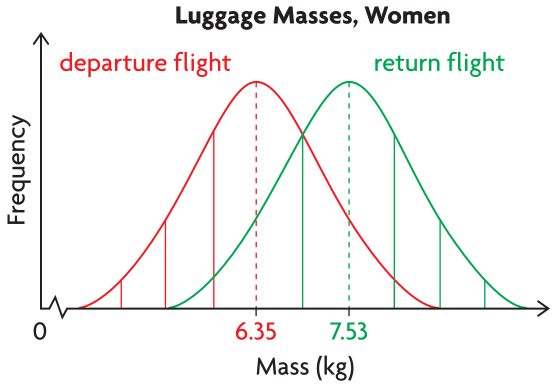
On the same graph, I sketched the normal curve for masses of the luggage for the women's team. I knew that this curve must be narrower than the curve for the men's team, since the standard deviation is lower.

I also knew that the area under this curve represents 100% of the data for the women's luggage, so the area under the red curve is also equal to 1 unit. For the area under both curves to be the same, the normal curve for the women's team must be taller.

b) Data for masses of luggage for women's team on flight home:

$\mu = 6.35 + 1.18$ or 7.53 kg

$\sigma = 0.59$ kg



Since each member of the women's team will add 1.18 kg to the mass of her carry-on luggage, the mean mass will increase by 1.18 kg. Although the mass of each piece of luggage will change, the distribution of the masses will stay the same, and the standard deviation will still be 0.59 kg.

I sketched a graph to show the new masses of the luggage for the women's team. It made sense that the new graph would simply move 1.18 to the right of the old graph. The shape stayed the same because the standard deviations of the two graphs are the same.

Your Turn

Suppose that the women had gone shopping and had also added their purchases to their carry-on luggage. How would you sketch a graph to show the distribution of the masses of their luggage for the trip home? Explain.

EXAMPLE 4 Analyzing data to solve a problem

Shirley wants to buy a new cellphone. She researches the cellphone she is considering and finds the following data on its longevity, in years.

2.0	2.4	3.3	1.7	2.5	3.7	2.0	2.3	2.9	2.2
2.3	2.7	2.5	2.7	1.9	2.4	2.6	2.7	2.8	2.5
1.7	1.1	3.1	3.2	3.1	2.9	2.9	3.0	2.1	2.6
2.6	2.2	2.7	1.8	2.4	2.5	2.4	2.3	2.5	2.6
3.2	2.1	3.4	2.2	2.7	1.9	2.9	2.6	2.7	2.8

a) Does the data approximate a normal distribution?

b) If Shirley purchases this cellphone, what is the likelihood that it will last for more than three years?

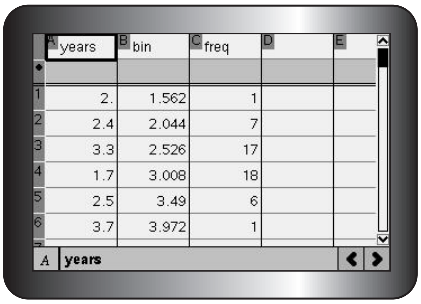
Shirley's Solution

a) $\mu = 2.526$

$\sigma = 0.482$

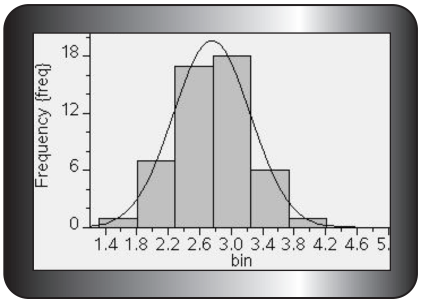
median $= 2.55$

Using my calculator, I determined the mean, the standard deviation, and the median. The median is close to the mean, which indicates that the data may be normally distributed.



	A years	B bin	C freq	D	E
◆					
1	2.	1.562	1		
2	2.4	2.044	7		
3	3.3	2.526	17		
4	1.7	3.008	18		
5	2.5	3.49	6		
6	3.7	3.972	1		

I created a frequency table to generate a histogram.



I created a histogram of the data, with an interval width of σ. My histogram looked almost symmetrical, so I decided to check the data to see how closely it approximates a normal distribution.

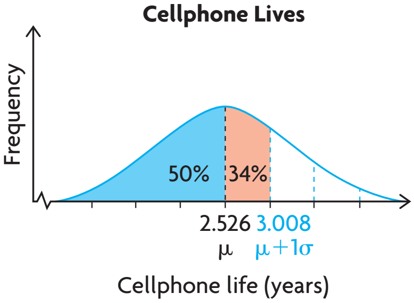
I generated a normal distribution curve on top of the histogram.

-1σ to 1σ	-2σ to 2σ	-3σ to 3σ
$\frac{35}{50} = 70\%$	$\frac{48}{50} = 96\%$	100%

I determined the percent of the data within one, two, and three standard deviations of the mean. I think the percents are reasonably close to those for a normal distribution (68%, 95%, and 99.7%).

The data approximates a normal distribution.

b)



I sketched a normal curve with the mean at 2.526 and the mean + 1 standard deviation at 3.008. I could use this location on the graph to determine the percent of values greater than three years.

I knew that the left-half area of the curve contains 50% of the data, and the area between 2.526 and 3.008 contains approximately 34% of the data.

$\mu + 1\sigma = 2.526 + 0.482$, or 3.008

$100\% - (50\% + 34\%) = 16\%$
About 16% of the cellphones lasted more than three years.

The area under the curve to the right of 3.008 is the white section. I subtracted the area of the coloured sections from 100%.

Your Turn

If Shirley purchases this cellphone, what is the likelihood that it will last at least 18 months?

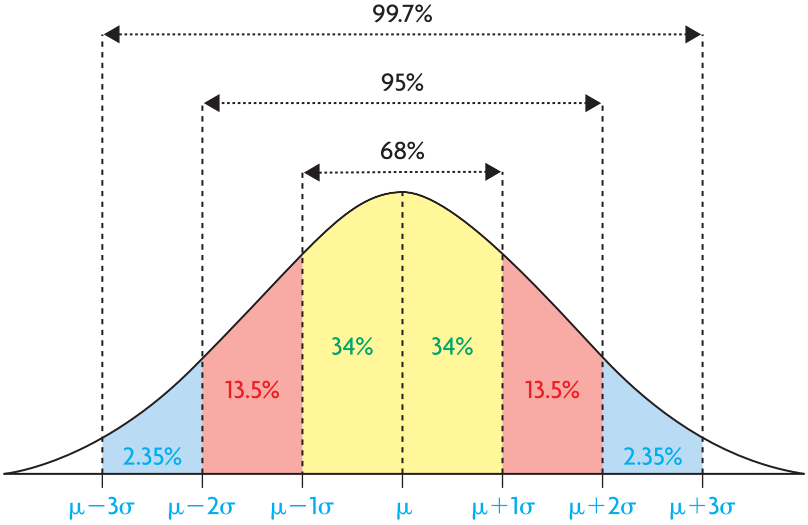
In Summary

Key Ideas

- Graphing a set of grouped data can help you determine whether the shape of the frequency polygon can be approximated by a normal curve.
- You can make reasonable estimates about data that approximates a normal distribution, because data that is normally distributed has special characteristics.
- Normal curves can vary in two main ways: the mean determines the location of the centre of the curve on the horizontal axis, and the standard deviation determines the width and height of the curve.

Need to Know

- The properties of a normal distribution can be summarized as follows:
  - The graph is symmetrical. The mean, median, and mode are equal (or close) and fall at the line of symmetry.
  - The normal curve is shaped like a bell, peaking in the middle, sloping down toward the sides, and approaching zero at the extremes.
  - About 68% of the data is within one standard deviation of the mean.
  - About 95% of the data is within two standard deviations of the mean.
  - About 99.7% of the data is within three standard deviations of the mean.
  - The area under the curve can be considered as 1 unit, since it represents 100% of the data.



- Generally, measurements of living things (such as mass, height, and length) have a normal distribution.

CHECK Your Understanding

1. The ages of members of a seniors curling club are normally distributed, with a mean of 63 years and a standard deviation of 4 years. What percent of the curlers is in each of the following age groups?
   a) between 55 and 63 years old
   b) between 67 and 75 years old
   c) older than 75 years old

Test	Mean (μ)	Standard Deviation (σ)
1	77	3.9
2	83	3.9
3	77	7.4

2. A teacher is analyzing the class results for three biology tests. Each set of marks is normally distributed.
   a) Sketch normal curves for tests 1 and 2 on one graph. Sketch normal curves for tests 1 and 3 on a different graph.
   b) Examine your graphs. How do tests 1 and 3 compare? How do tests 1 and 2 compare?
   c) Determine Oliver's marks on each test, given the information shown at the right.

Test	Oliver's Mark
1	$\mu + 2\sigma$
2	$\mu - 1\sigma$
3	$\mu + 3\sigma$

3. Is the data in each set normally distributed? Explain.

a)

Interval	10–19	20–29	30–39	40–49	50–59	60–69
Frequency	3	5	17	20	11	4

b)

Interval	2–5	6–9	10–13	14–17	18–21	22–25
Frequency	2	8	8	3	4	5

c)

Interval	10–24	25–39	40–54	55–69	70–84	85–99
Frequency	2	7	16	10	4	1

PRACTISING

4. Tiegan is organizing her movie collection. She decides to record the length of each movie, in minutes.

91	129	95	96	96	90	101	87	100	90
86	78	105	99	81	106	101	122	91	102
89	125	162	155	89	89	180	94	84	99
73	100	99	100	117	135	100	89	87	110
125	103	94	99	98	102	96	88	154	144

   a) Determine the mean and standard deviation for the set of data.
   b) Create a frequency table, using σ as the interval width.
   c) Are the lengths of Tiegan's movies normally distributed? Explain.

The Indian monsoon, or rainy season, usually begins in June or July, depending on location, and ends late in September.

5. The data in each of the following sets has been ordered from least to greatest. For each set,

i) calculate the mean, median, and standard deviation;
ii) create a frequency polygon; and
iii) explain why the distribution is or is not approximately normal.

a) daily maximum temperatures (°C) in monsoon season in India:
41.5, 42.4, 42.6, 42.7, 42.9, 43.0, 43.6, 44.0, 44.5, 44.6, 44.6, 44.8, 45.0, 45.3, 45.5, 45.5, 45.6, 45.7, 45.8, 46.1, 46.3, 46.4, 46.5, 46.6, 46.8, 47.0, 47.2, 47.6, 47.6, 47.9

b) class marks on a pop quiz out of 15:
2, 4, 5, 6, 6, 6, 7, 7, 7, 7, 8, 8, 8, 8, 8, 8, 9, 9, 9, 10, 10, 11, 11, 11, 12, 12, 13, 13, 15

6. A manufacturer offers a warranty on its coffee makers. The coffee makers have a mean lifespan of 4.5 years, with a standard deviation of 0.7 years. For how long should the coffee makers be covered by the warranty, if the manufacturer wants to repair no more than 2.5% of the coffee makers sold?

Rolling 3 Dice

Sum	Frequency
3	1
4	3
5	6
6	10
7	15
8	21
9	25
10	27
11	27
12	25
13	21
14	15
15	10
16	6
17	3
18	1

7. Hila found the data at the left that shows the number of ways that each sum can be obtained when rolling three dice.

a) Determine the mean and the standard deviation.
b) Draw a frequency polygon to show the data.
c) Does the data have a normal distribution? Explain.

8. The company payroll of Sweetwater Communications has a mean monthly salary of $5400, with a standard deviation of $800.

a) Sketch a normal curve to represent the salaries for the company.
b) Sketch a curve to show the effects of Proposal 1: Each employee receives a raise of $270 per month.
c) Sketch a curve to show the effects of Proposal 2: Each employee receives a 5% raise on the original salary.

9. The results for the first round of the 2009 Masters golf tournament are given below.

65	68	70	71	72	73	74	76
66	69	70	71	72	73	74	76
66	69	70	72	73	73	75	77
67	69	71	72	73	73	75	77
67	69	71	72	73	73	75	77
68	69	71	72	73	73	75	78
68	69	71	72	73	73	75	78
68	70	71	72	73	73	75	78
68	70	71	72	73	73	75	79
68	70	71	72	73	74	75	79
68	70	71	72	73	74	75	79
68	70	71	72	73	74	76	80

a) Are the golf scores normally distributed?
b) Explain how the measures of central tendency support your decision in part a).

10. A school of 130 bottlenose dolphins is living in a protected environment. The life expectancy of the dolphins is normally distributed, with a mean of 39 years and a standard deviation of 3.5 years. How many of these dolphins can be expected to live more than 46 years?

11. Julie is an engineer who designs roller coasters. She wants to design a roller coaster that 95% of the population can ride. The average adult in North America has a mass of 71.8 kg, with a standard deviation of 13.6 kg.
a) What range of masses should Julie consider in her design?
b) If Julie wanted to design a roller coaster that 99.7% of the population could ride, what range of masses should she consider?
c) What assumption is being made, which could cause problems if it is not valid?

12. A new video game is being tested with a sample of students. The scores on the first attempt for each player are recorded in the table.
a) Graph the data. Does the data appear to have a normal distribution?
b) Determine the mean and standard deviation of the data. Do these values validate your answer to part a)?

Scores	Freq.
less than 18 000	2
18 000–27 000	5
27 000–36 000	14
36 000–45 000	36
45 000–54 000	77
54 000–63 000	128
63 000–72 000	163
72 000–81 000	163
81 000–90 000	127
90 000–99 000	80
99 000–108 000	33
108 000–117 000	14
117 000–126 000	6
greater than 126 000	2

13. In a dog obedience class, the masses of the 60 dogs enrolled were normally distributed, with a mean of 11.2 kg and a standard deviation of 2.8 kg. How many dogs would you expect to fall within each range of masses?
a) between 8.4 kg and 14.0 kg
b) between 5.6 kg and 16.8 kg
c) between 2.8 kg and 19.6 kg
d) less than 11.2 kg

14. The mass of an Appaloosa horse is generally in the range of 431 kg to 533 kg. Assuming that the data is normally distributed, determine the mean and standard deviation for the mass of an Appaloosa. Justify your answers.

The Appaloosa Horse Club of Canada Museum is located in Claresholm, Alberta.

Closing

15. Explain why a selection of 10 students from a class can have marks that are not normally distributed, even when the marks of the whole class are normally distributed.

Extending

16. Newfoundland dogs have masses that are normally distributed. The mean mass of a male dog is 63.5 kg, with a standard deviation of 1.51 kg. The mean mass of a female dog is 49.9 kg, with a standard deviation of 1.51 kg. Esteban claims that he used to have two adult Newfoundland dogs: a male that was 78.9 kg and a female that was 29.9 kg. Using your knowledge of normal distribution, do you think he is being truthful? Explain.

Applying Problem-Solving Strategies

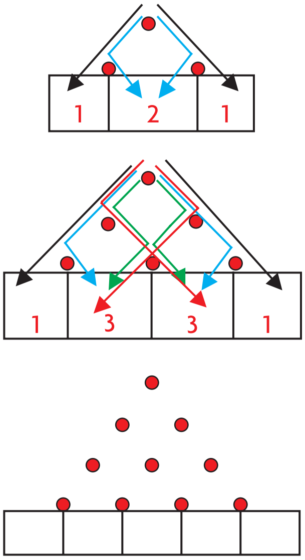
Predicting Possible Pathways

A Galton board is a triangular array of pegs that is used for statistical experiments. Balls are dropped, one at a time, onto the top peg and fall either right or left. Then, as they hit a peg in the next row, they fall either right or left again, until they finally pass through a slot at the bottom, where they can be counted. Each ball is equally likely to fall either way.

The Puzzle

A. For an array with 2 rows and 3 pegs, there is one way for a ball to fall into each end slot, and two ways for a ball to fall into the middle slot. For an array with 3 rows and 6 pegs, there is one way for a ball to fall into each end slot, and three ways for a ball to fall into the two middle slots.

Determine the number of ways for a ball to fall through an array with 4 rows and 10 pegs.



B. Examine each array in step A. Look for a pattern to determine how many pegs there would be in a 5-row array and a 6-row array.

C. Look for a pattern in the number of ways for a ball to fall into each slot in the arrays in part A. Use this pattern to determine the number of ways for a ball to fall into each slot in a 5-row array and a 6-row array.

The Strategy

D. Describe a strategy you could use to determine the number of ways for a ball to fall into each slot in an array of any size.

E. Use your strategy to determine the number of ways for a ball to fall into each slot in an array with 10 rows.

F. Draw a histogram or a frequency polygon to illustrate the results for a 10-row array. Comment on the distribution.

5.5 *Z*-Scores

GOAL

Use *z*-scores to compare data, make predictions, and solve problems.

YOU WILL NEED

- calculator
- grid paper
- *z*-score tables (pages 592 to 593)

LEARN ABOUT the Math

Hailey and Serge belong to a running club in Vancouver. Part of their training involves a 200 m sprint. Below are normally distributed times for the 200 m sprint in Vancouver and on a recent trip to Lake Louise. At higher altitudes, run times improve.

Location	Altitude (m)	Club Mean Time: μ for 200 m (s)	Club Standard Deviation: σ (s)	Hailey's Run Time (s)	Serge's Run Time (s)
Vancouver	4	25.75	0.62	24.95	25.45
Lake Louise	1661	25.57	0.60	24.77	26.24

? At which location was Hailey's run time better, when compared with the club results?

EXPLORE...

- Alexis plays in her school jazz band. Band members practise an average of 16.5 h per week, with a standard deviation of 4.2 h. Alexis practises an average of 22 h per week. As a class, discuss how you might estimate the percent of the band that, on average, practises a greater number of hours than Alexis.

EXAMPLE 1 Comparing *z*-scores

Determine at which location Hailey's run time was better, when compared with the club results.

Marcel's Solution

For any given score, x, from a normal distribution, $x = \mu + z\sigma$, where z represents the number of standard deviations of the score from the mean.

Hailey's run time is less at Lake Louise, but so is the club's mean run time. I can't compare these times directly, because the means and standard deviations are different for the two locations. To make the comparison, I have to standardize Hailey's times to fit a common normal distribution.

Solving for z results in a formula for a **z-score**:

$$z = \frac{x - \mu}{\sigma}$$

A *z*-score indicates the position of a data value on a **standard normal distribution**.

***z*-score**

A standardized value that indicates the number of standard deviations of a data value above or below the mean.

standard normal distribution

A normal distribution that has a mean of zero and a standard deviation of one.

Vancouver:

$$z = \frac{x - \mu}{\sigma}$$

$$z = \frac{24.95 - 25.75}{0.62}$$

$$z = -1.290 \ldots$$

Lake Louise:

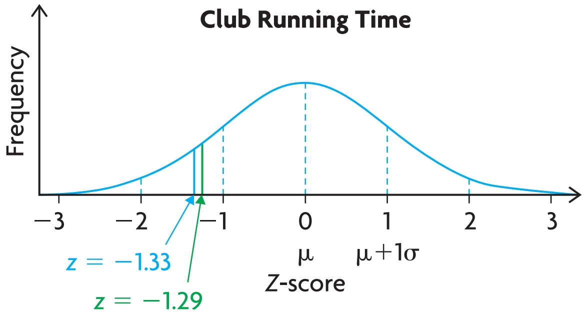
$$z = \frac{x - \mu}{\sigma}$$

$$z = \frac{24.77 - 25.57}{0.60}$$

$$z = -1.333 \ldots$$

I know that *z*-scores can be used to compare data values from different normal distributions. I calculated the *z*-score for Hailey's run times at each location.

Hailey's run time is about 1.29 standard deviations below the mean in Vancouver, and 1.33 standard deviations below the mean in Lake Louise.



I sketched the standard normal curve, which has a mean of zero and a standard deviation of 1. Then I drew a line on the graph for each *z*-score.

The *z*-score for Hailey's Lake Louise run is farther to the left than the *z*-score for her Vancouver run.

Hailey's time for 200 m was better than the club's mean in both locations. However, Hailey's *z*-score for Lake Louise was lower than her *z*-score for Vancouver, so her time was better in Lake Louise.

I can make this comparison because both times have been translated to a normal distribution that has the same mean and standard deviation.

Reflecting

A. Use *z*-scores to determine which of Serge's runs was better.

B. Explain why the lower *z*-score represents a relatively faster run.

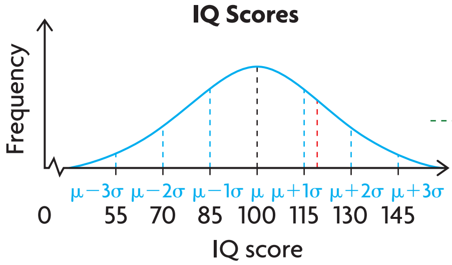
C. What can you say about a data value if you know that its *z*-score is negative? positive? zero?

APPLY the Math

EXAMPLE 2 | Using *z*-scores to determine the percent of data less than a given value

IQ tests are sometimes used to measure a person's intellectual capacity at a particular time. IQ scores are normally distributed, with a mean of 100 and a standard deviation of 15. If a person scores 119 on an IQ test, how does this score compare with the scores of the general population?

Malia's Solution: Using a *z*-score table



First, I sketched a normal curve and determined the IQ scores for one, two, and three standard deviations from the mean.

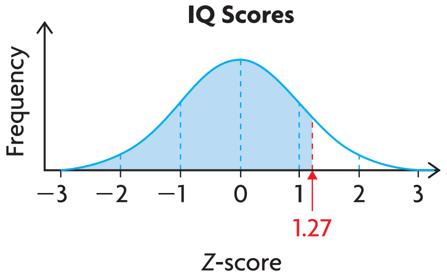
Then I drew a line that represented an IQ score of 119.

I noticed that my line was between one and two standard deviations above the mean.

$$z = \frac{x - \mu}{\sigma}$$

$$z = \frac{119 - 100}{15}$$

$$z = 1.2666...$$



I determined the *z*-score for an IQ of 119.

An IQ score of 119 is about 1.27 standard deviations above the mean. I sketched this on a standard normal curve.

I knew that I needed to determine the percent of people with IQ scores less than 119. This is equivalent to the area under the curve to the left of 1.27 on the standard normal curve.

***z*-score table**

A table that displays the fraction of data with a *z*-score that is less than any given data value in a standard normal distribution.
(There is a *z*-score table on pages 592 to 593.)

z	0.0	0.01	0.06	0.07
0.0	0.5000	0.5040	0.5239	0.5279
0.1	0.5398	0.5438	0.5636	0.5675
1.1	0.8643	0.8665	0.8770	0.8790
1.2	0.8849	0.8869	0.8962	**0.8980**
1.3	0.9032	0.9049	0.9131	0.9147

I used a ***z*-score table**.

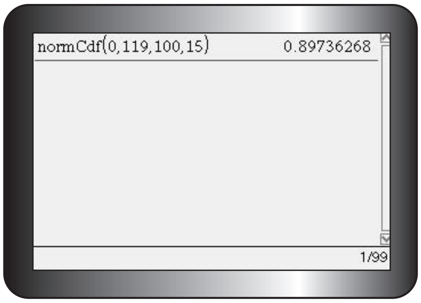
1.27 = 1.2 + 0.07

I used the 1.2 row and the 0.07 column.

The value in the table, 0.8980, is the fraction of the area under the curve to the left of the *z*-score.

The value in the *z*-score table is 0.8980. This means that an IQ score of 119 is greater than 89.80% of IQ scores in the general population.

Desiree's Solution: Using a graphing calculator



```
normCdf(0,119,100,15)                0.89736268
```

I used the statistics function for normal distributions on my calculator to determine the percent of the population that has an IQ score between 0 and 119.

I entered the lower bound of 0, the upper bound of 119, the mean of 100, and the standard deviation of 15.

An IQ score of 119 is greater than 89.74% of all the scores.

My solution is slightly different from Malia's because this method does not use a rounded *z*-score.

Your Turn

Megan determined the area under the normal curve using slightly different reasoning: "I know that the total area under a normal curve is 100%, so the area under the curve to the left of the mean is 50%. I used my graphing calculator to calculate the area between a *z*-score of 0 and a *z*-score of 1.27, by entering these as the lower and upper bounds. My calculator gave a result of 0.397... ."

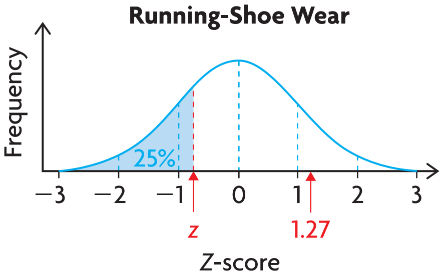
How could Megan use this result to complete her solution?

EXAMPLE 3 Using *z*-scores to determine data values

Athletes should replace their running shoes before the shoes lose their ability to absorb shock.

Running shoes lose their shock-absorption after a mean distance of 640 km, with a standard deviation of 160 km. Zack is an elite runner and wants to replace his shoes at a distance when only 25% of people would replace their shoes. At what distance should he replace his shoes?

Rachelle's Solution: Using a *z*-score table



I sketched the standard normal curve. I needed the *z*-score for 25% of the area under the curve, or 0.25.

z	0.09	0.08	0.07	0.06	0.05
–0.7	0.2148	0.2177	0.2206	0.2236	0.2266
–0.6	0.2451	0.2483	0.2514	0.2546	0.2578
–0.5	0.2776	0.2810	0.2843	0.2877	0.2912

I searched the *z*-score table for a value that is close to 0.25.

The *z*-score that represents an area of 0.25 is about halfway between −0.67 and −0.68, or about −0.675.

$$z = \frac{x - \mu}{\sigma}$$

$$(-0.675) = \frac{x - (640)}{(160)}$$

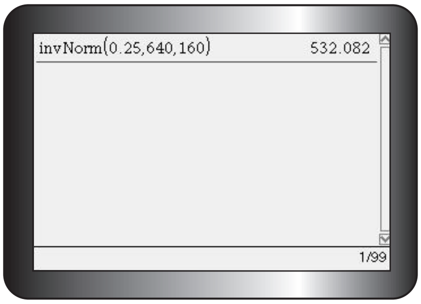
$$-108 = x - 640$$

$$532 = x$$

I substituted the values I knew into the *z*-score formula and solved for *x*.

Zack should replace his running shoes after 532 km.

Renalda's Solution: Using a graphing calculator



```
invNorm(0.25,640,160)                532.082


                                        1/99
```

I used the statistics function on my calculator.

I entered the decimal value for the percent of data to the left of the *z*-score I needed. Then I entered the mean and standard deviation of the data.

Zack should replace his running shoes after 532 km.

Therefore, 25% of people would replace their shoes after 532 km.

Your Turn

Quinn is a recreational runner. He plans to replace his running shoes when 70% of people would replace their shoes. After how many kilometres should he replace his running shoes?

EXAMPLE 4 Solving a quality control problem

The ABC Company produces bungee cords. When the manufacturing process is running well, the lengths of the bungee cords produced are normally distributed, with a mean of 45.2 cm and a standard deviation of 1.3 cm. Bungee cords that are shorter than 42.0 cm or longer than 48.0 cm are rejected by the quality control workers.

a) If 20 000 bungee cords are manufactured each day, how many bungee cords would you expect the quality control workers to reject?

b) What action might the company take as a result of these findings?

Logan's Solution: Using a *z*-score table

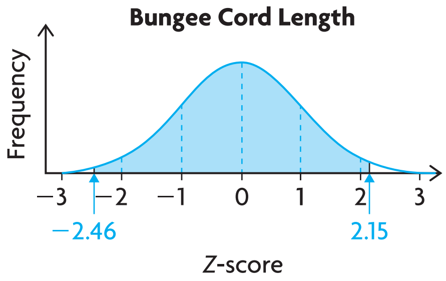
a) Minimum length = 42 cm   Maximum length = 48 cm

$$z_{\text{min}} = \frac{x - \mu}{\sigma} \qquad z_{\text{max}} = \frac{x - \mu}{\sigma}$$

$$z_{\text{min}} = \frac{42.0 - 45.2}{1.3} \qquad z_{\text{max}} = \frac{48.0 - 45.2}{1.3}$$

$$z_{\text{min}} = -2.461... \qquad z_{\text{max}} = 2.153...$$

I determined the *z*-scores for the minimum and maximum acceptable lengths.



I sketched the standard normal curve. The area under the curve to the left of −2.46 represents the percent of rejected bungee cords less than 42 cm. The area under the curve to the right of 2.15 represents the percent of rejected bungee cords greater than 48 cm.

Area to left of −2.46 = 0.0069

Area to right of 2.15 = 1 − 0.9842
Area to right of 2.15 = 0.0158

I looked up each *z*-score in the *z*-score table. The *z*-score table gives the area to the left of the *z*-score, which I want for 42 cm.

Since I wanted the area to the right of the *z*-score for 48 cm, I had to subtract the corresponding area from 1.

Percent rejected = Area to the left of −2.46
+ Area to the right of 2.15
Percent rejected = 0.0069 + 0.0158
Percent rejected = 0.0227 or 2.27%

I added the two areas to determine the percent of bungee cords that are rejected.

Total rejected = (0.0227)(20 000) or 454

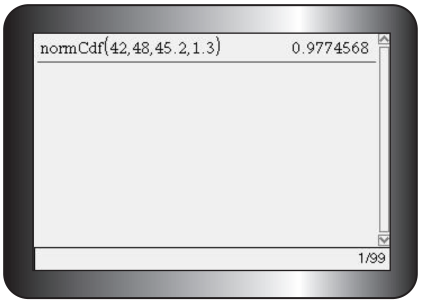
I determined the number of bungee cords that are rejected.

b) ABC needs a more consistent process, because 454 seems like a large number of bungee cords to reject. The company should adjust its equipment so that the standard deviation is lowered.

Lowering the standard deviation will reduce the percent of rejected bungee cords.

Nathan's Solution: Using a graphing calculator

a)



```
normCdf(42,48,45.2,1.3)          0.9774568
                                      1/99
```

I used the statistics function on my calculator to determine the percent of bungee cords that are an acceptable length. I entered the minimum and maximum acceptable lengths and then the mean and standard deviation.

Number accepted = 20 000 × 0.977...
Number accepted = 19 549.135...
About 19 549 bungee cords meet the standard every day, so 451 bungee cords are rejected every day.

I determined the number of bungee cords that meet the standard. Then I subtracted to determine the number rejected.

My solution is slightly different from Logan's solution because this method does not use a rounded *z*-score value.

b) I think the company should adjust its equipment to get a lower standard deviation, so fewer bungee cords are discarded.

Your Turn

a) What percent of all the bungee cords are accepted?

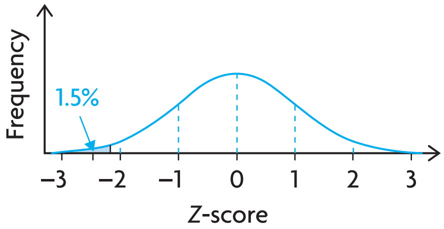
b) A client has placed an order for 12 000 bungee cords, but will only accept bungee cords that are between 44.0 cm and 46.0 cm in length. Can this client's order be filled by one day's production, with the equipment operating as is? Explain.

EXAMPLE 5 Determining warranty periods

A manufacturer of personal music players has determined that the mean life of the players is 32.4 months, with a standard deviation of 6.3 months. What length of warranty should be offered if the manufacturer wants to restrict repairs to less than 1.5% of all the players sold?

Sacha's Solution

$1.5\% = 0.015$



I used my graphing calculator to determine the *z*-score that corresponds to an area under the normal curve of 0.015.

$z = -2.17$

$$z = \frac{x - \mu}{\sigma}$$

$$(-2.17) = \frac{x - (32.4)}{(6.3)}$$

$$-13.671 = x - 32.4$$

$$18.729 = x$$

I substituted the known values into the *z*-score formula and solved for *x*.

The manufacturer should offer an 18-month warranty.

Since the manufacturer wants to repair less than 1.5% of the music players, I rounded down to 18 months.

Your Turn

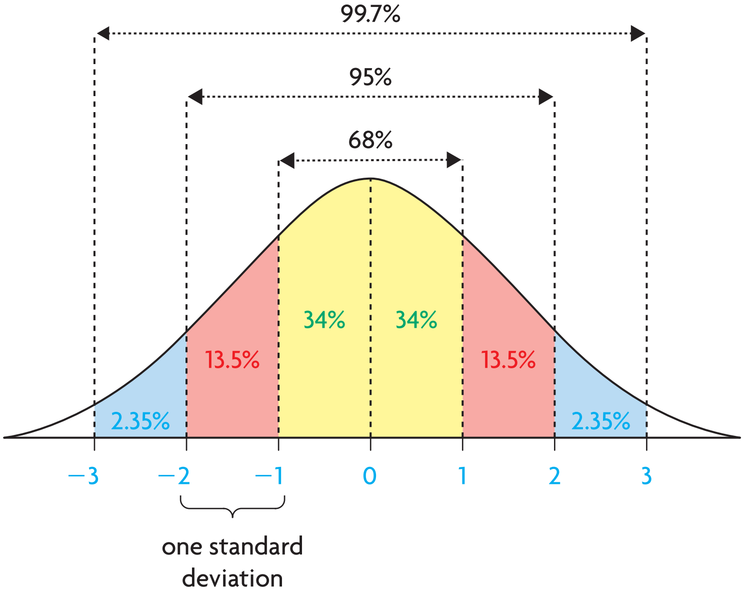
a) If 10 000 personal music players are sold, how many could the manufacturer expect to receive for repairs under warranty?

b) The manufacturer wants to offer the option of purchasing an extended warranty. If the manufacturer wants to repair, at most, 20% of the players under the extended warranty, what length of extended warranty should be offered?

In Summary

Key Ideas

- The standard normal distribution is a normal distribution with mean, μ, of 0 and a standard deviation, σ, of 1. The area under the curve of a normal distribution is 1.
- *Z*-scores can be used to compare data from different normally distributed sets by converting their distributions to the standard normal distribution.



Need to Know

- A *z*-score indicates the number of standard deviations that a data value lies from the mean. It is calculated using this formula:

$$z = \frac{x - \mu}{\sigma}$$

- A positive *z*-score indicates that the data value lies above the mean. A negative *z*-score indicates that the data value lies below the mean.
- The area under the standard normal curve, to the left of a particular *z*-score, can be found in a *z*-score table or determined using a graphing calculator.

CHECK Your Understanding

1. Determine the z-score for each value of x.
   a) $\mu = 112, \sigma = 15.5, x = 174$
   b) $\mu = 53.46, \sigma = 8.24, x = 47.28$
   c) $\mu = 82, \sigma = 12.5, x = 58$
   d) $\mu = 245, \sigma = 22.4, x = 300$

2. Using a z-score table (such as the table on pages 592 to 593), determine the percent of the data to the left of each z-score.
   a) $z = 1.24$ b) $z = -2.35$ c) $z = 2.17$ d) $z = -0.64$

3. Determine the percent of the data between each pair of z-scores.
   a) $z = -2.88$ and $z = -1.47$ b) $z = -0.85$ and $z = 1.64$

4. What z-score is required for each situation?
   a) 10% of the data is to the left of the z-score.
   b) 10% of the data is to the right of the z-score.
   c) 60% of the data is below the z-score.
   d) 60% of the data is above the z-score.

PRACTISING

In the following questions, assume that the data approximates a normal distribution.

5. Calculate the z-score for each value of x.
   a) $\mu = 24, \sigma = 2.8, x = 29.3$
   b) $\mu = 165, \sigma = 48, x = 36$
   c) $\mu = 784, \sigma = 65.3, x = 817$
   d) $\mu = 2.9, \sigma = 0.3, x = 3.4$

6. Determine the percent of the data to the left of each z-score.
   a) $z = 0.56$ b) $z = -1.76$ c) $z = -2.98$ d) $z = 2.39$

7. Determine the percent of the data to the right of each z-score.
   a) $z = -1.35$ b) $z = 2.63$ c) $z = 0.68$ d) $z = -3.14$

8. Determine the percent of the data between each pair of z-scores.
   a) $z = 0.24$ and $z = 2.53$ b) $z = -1.64$ and $z = 1.64$

9. Determine the z-score for each situation.
   a) 33% of the data is to the left of the z-score.
   b) 20% of the data is to the right of the z-score.

10. Meg wonders if she should consider a career in the sciences, because she does well in mathematics. However, she also does well in English and has thought about becoming a journalist.
    a) Determine the z-score for each of Meg's marks.
    b) Which subject is Meg better in, relative to her peers?
    c) What other factors should Meg consider?

Subject	Standard Test Results (%)		Meg's Mark (%)
	μ	σ	
English	77	6.8	93
math	74	5.4	91

11. A hardwood flooring company produces flooring that has an average thickness of 175 mm, with a standard deviation of 0.4 mm. For premium-quality floors, the flooring must have a thickness between 174 mm and 175.6 mm. What percent, to the nearest whole number, of the total production can be sold for premium-quality floors?

12. Violeta took part in a study that compared the heart-rate responses of water walking versus treadmill walking for healthy college females. Violeta's heart rate was 68 on the treadmill for the 2.55 km/h walk and 145 in the water for the 3.02 km/h walk. For which event was her heart rate lower, compared with the others who took part in the study?

	Treadmill (beats/min)		Water (beats/min)	
Speed	μ	σ	μ	σ
Resting	68	8.43	71	6.15
2.55 km/h	76	9.15	130	13.50
2.77 km/h	79	11.66	146	11.96
3.02 km/h	81	11.33	160	13.50
3.31 km/h	81	10.27	167	12.58

13. In 2006, the ages of mothers who had children aged 4 and under were approximately normally distributed, with a mean age of 32 years and a standard deviation of 5.9 years. The data is shown in the table at the right.

a) Determine the percent of mothers who were less than 40 years old.

b) Determine the percent of mothers who were less than 21 years old.

c) Determine the percent of mothers who were 18 years old or less. Why might someone want to know this?

Age of Mother (years)	2006 Census (%)
15–19	1.1
20–24	8.8
25–29	23.2
30–34	33.7
35–39	23.8
40–44	8.2
45–49	1.2
Total	100

Statistics Canada

14. In a population, 50% of the adults are taller than 180 cm and 10% are taller than 200 cm. Determine the mean height and standard deviation for this population.

15. A medical diagnostic test counts the number of blood cells in a sample. The red blood cell count (in millions per cubic microlitre) is normally distributed, with a mean of 4.8 and a standard deviation of 0.3.

a) What percent of people have a red blood cell count that is less than 4?

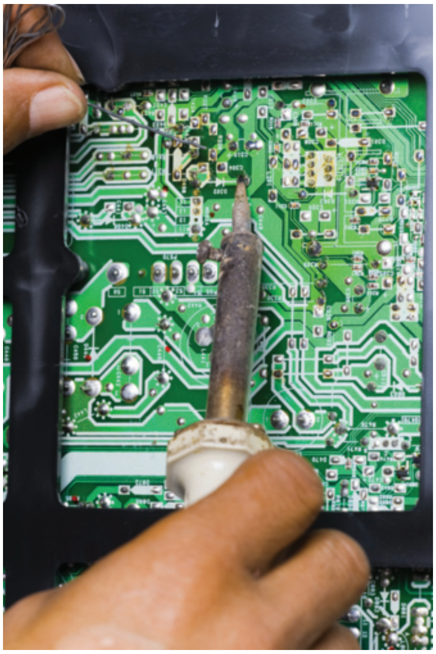
b) What percent of people have a count between 4.7 and 5.0?

c) What red blood cell count would someone have if 95% of people have a lower count?

16. An MP3 player has a one-year warranty. The mean lifespan of the player is 2.6 years, with a standard deviation of 0.48 years.

a) A store sells 4000 players. How many of these players will fail before the warranty expires?

b) Tyler is offered an extended warranty, for one extra year, when he buys a player. What is the likelihood that he will make a claim on this warranty if he takes it?



17. A manufacturer of plasma televisions has determined that the televisions require servicing after a mean of 67 months, with a standard deviation of 7.2 months. What length of warranty should be offered, if the manufacturer wants to repair less than 1% of the televisions under the warranty?

18. A tutor guarantees that 10% of her students will obtain an A on every test they write. For the last test, the mean mark is 68 and the standard deviation is 6. What mark is required to receive an A on the test?

19. In the insurance industry, standard deviation is used to quantify risk—the greater the risk, the higher the standard deviation. For example, consider the cost of a car accident for two different cars: a high-priced luxury car and a mid-priced car. The expected cost of repairs for both cars is \$2500. However, the standard deviation for the high-priced car is \$1000, and the standard deviation for the mid-priced car is \$400. Explain why the probability that the repairs will cost more than \$3000 is 31% for the high-priced car but only 11% for the mid-priced car.

20. **a)** A club accepts members only if they have an IQ score that is greater than the scores for 98% of the population. What IQ score would you need to be accepted into this club? (Recall that $\mu = 100$ and $\sigma = 15$ for the general population.)

b) Only 0.38% of the population are considered to be geniuses, as measured by IQ scores. What is the minimum IQ score that is required to be considered a genius?

c) Jarrod was told that his IQ score is in the top 30% of the population. What is his IQ score?

Closing

21. What is a *z*-score, how do you determine it, and what is it used for?

Extending

22. A company packages sugar into 5 kg bags. The filling machine can be calibrated to fill to any specified mean, with a standard deviation of 0.065 kg. Any bags with masses that are less than 4.9 kg cannot be sold and must be repackaged.

a) If the company wants to repackage no more than 3% of the bags, at what mean should they set the machine?

b) Assuming that the company sets the machine at the mean you determined in part a), what percent of the bags will have more than 5 kg of sugar? Do you think the company will be satisfied with this percent?

The LSAT must be taken by people who want to gain admission to a law school. The test focuses on logical and verbal reasoning skills.

23. Approximately 40% of those who take the LSAT, or Law School Admission Test, score from 145 to 155. About 70% score from 140 to 160.

a) Determine the mean score and standard deviation for the LSAT.

b) Harvard University also uses other methods to choose students for its law school, but the minimum LSAT score that is required is about 172. What percent of people who take the LSAT would be considered by Harvard for admission?

24. Create your own problem involving *z*-score analysis, using any of the normally distributed data from Lesson 5.4. Exchange problems with classmates, and solve the problems. Provide suggestions for improving the problems.

5.6 Confidence Intervals

GOAL

Use the normal distribution to solve problems that involve confidence intervals.

LEARN ABOUT the Math

A telephone survey of 600 randomly selected people was conducted in an urban area. The survey determined that 76% of people, from 18 to 34 years of age, have a social networking account.

The results are accurate within plus or minus 4 percent points, 19 times out of 20.

? How can this result be interpreted, if the total population of 18- to 34-year-olds is 92 500?

EXAMPLE 1 Analyzing and applying survey results

Calculate the range of people that have a social networking account, and determine the certainty of the results.

Danica's Solution

The **margin of error** for the data is ± 4%, so the **confidence interval** is 76% ± 4%, which is from (76 − 4)% or 72% to (76 + 4)% or 80%.

I interpreted the survey statement.

The **confidence level** of the survey is 95%. The probability of error for this result is 5%. If the survey were conducted 100 times, then 95 times out of 100, the percent of people in the population with a social networking account would be from 72% to 80%.

The results are accurate 19 times out of 20, which is 95% of the time.

YOU WILL NEED

- calculator
- *z*-score tables (pages 592 to 593)

EXPLORE...

- Often it is impractical to survey an entire population. For example, if a light bulb company wants to test the number of hours that a light bulb will burn before failing, it cannot test every bulb. Propose a method that the company could use to determine the longevity of its light bulbs.

margin of error

The possible difference between the estimate of the value you're trying to determine, as determined from a random sample, and the true value for the population; the margin of error is generally expressed as a plus or minus percent, such as ±5%.

confidence interval

The interval in which the true value you're trying to determine is estimated to lie, with a stated degree of probability; the confidence interval may be expressed using ± notation, such as 54.0% ± 3.5%, or ranging from 50.5% to 57.5%.

confidence level

The likelihood that the result for the "true" population lies within the range of the confidence interval; surveys and other studies usually use a confidence level of 95%, although 90% or 99% is sometimes used.

$92\,500 \times 76\% = 70\,300$

$92\,500 \times 4\% = 3700$

The confidence interval for the population is 70 300 ± 3700.

I used the confidence interval, 76% ± 4%, and the population, 92 500, to calculate the range of the number of people in the population who have a social networking account.

$70\,300 - 3700 = 66\,600$

$70\,300 + 3700 = 74\,000$

It can be said, with 95% confidence, that 66 600 to 74 000 people, in a population of 92 500 people from ages 18 to 34, have a social networking account.

Reflecting

A. Based on this survey, what is the range for 18- to 34-year-olds who do not have a social networking account?

B. The same telephone survey was conducted by a different company, using a sample of 600 randomly selected people in both urban and rural areas. According to this survey, 76% of people, from 18 to 34 years of age, have a social networking account. (The results are accurate within plus or minus 5.3 percent points, 99 times out of 100.) How are the results of this survey different from those of the first survey? How are they the same?

APPLY the Math

EXAMPLE 2 Analyzing the effect of sample size on margin of error and confidence intervals

Polling organizations in Canada frequently survey samples of the population to gauge voter preference prior to elections. People are asked:

1. "If an election were held today, which party would you vote for?"

If they say they don't know, then they are asked:

2. "Which party are you leaning toward voting for?"

The results of three different polls taken during the first week of November, 2010, are shown on the next page. The results of each poll are considered accurate 19 times out of 20.

Polling Organization & Data	Conservative (%)	Liberal (%)	NDP (%)	Bloc Quebecois (%)	Green Party (%)	Other (%)
Ekos	29	29	19	9	11	3
sample size, 1815	margin of error, ±2.3%					
Nanos	37	32	15	11	5	n.a.
sample size, 844	margin of error, ±3.4%					
Ipsos	35	29	12	11	12	n.a.
sample size, 1000	margin of error, ±3.1%					

source: http://www.sfu.ca/~aheard/elections/polls.html

How does the sample size used in a poll affect:

a) the margin of error in the reported results?

b) the confidence interval in the reported results?

Martin's Solution

a) $\frac{19}{20} = 95\%$

In this case, the confidence level used by each polling organization is the same. This enables me to compare the effect that sample size has on the margin of error.

The confidence level of each poll is 95%.

	Nanos	Ipsos	Ekos
sample size	844	1000	1815
margin of error	±3.4%	±3.1%	±2.3%

I created a table to compare the polls. I arranged the polls in increasing order of sample size, then looked for a trend in the margin of error.

If polls are assessed using the same confidence level, when the sample size increases, the margin of error decreases.

A larger sample size results in the possibility of a poll that more accurately represents the population.

My observation makes sense because a larger sample should be a better indicator of how the population might vote.

b) Let n represent the number of people polled.

Nanos

$n = 844$

37% ± 3.4% or 33.6% to 40.4%

Ipsos

$n = 1000$

35% ± 3.1% or 31.9% to 38.1%

Ekos

$n = 1815$

29% ± 2.3% or 26.7% to 31.3%

I decided to compare the confidence interval for the Conservative Party for each of the 3 different polls. I wrote the confidence interval for each poll in increasing order of sample size.

The Nanos poll predicts that 33.6% to 40.4% of the population will vote for a Conservative. That is a range of 6.8%.

The Ipsos poll predicts that 31.9 % to 38.1% of the population will vote for a Conservative. That is a range of 6.2%.

The Ekos poll predicts that 26.7% to 31.3% of the population will vote for a Conservative. That is a range of 4.6%.

I interpreted each of the confidence intervals for these polls.

If polls are conducted using the same confidence level, when the sample size increases, the range in the confidence interval decreases.

My observation makes sense because the confidence interval is determined by the margin of error. So, as the sample size increases, the margin of error decreases and in turn the range of the confidence interval decreases.

Your Turn

Compare the confidence intervals for the Liberal Party for each of the three polls. Do your results reflect Martin's results above? Explain.

EXAMPLE 3 Analyzing the effect of confidence levels on sample size

To meet regulation standards, baseballs must have a mass from 142.0 g to 149.0 g. A manufacturing company has set its production equipment to create baseballs that have a mean mass of 145.0 g.

To ensure that the production equipment continues to operate as expected, the quality control engineer takes a random sample of baseballs each day and measures their mass to determine the mean mass. If the mean mass of the random sample is 144.7 g to 145.3 g, then the production equipment is running correctly. If the mean mass of the sample is outside the acceptable level, the production equipment is shut down and adjusted. The quality control engineer refers to the chart shown on the next page when conducting random sampling.

Confidence Level	Sample Size Needed
99%	110
95%	65
90%	45

a) What is the confidence interval and margin of error the engineer is using for quality control tests?

b) Interpret the table.

c) What is the relationship between confidence level and sample size?

Geoffrey's Solution

a) The confidence interval is 144.7 g to 145.3 g.

The confidence interval is the range that the mean mass of the random sample can fall in and be acceptable.

Margin of error:
145.3 − 145.0 or 0.3
144.7 − 145.0 or −0.3
The margin of error is ±0.3 g.

I subtracted the mean from the upper and lower limits of the confidence interval to determine the margin of error.

b)
- In order to be confident that, 99 out of 100 times, the mean mass of the sample measures from 144.7 g to 145.3 g, the engineer needs to take a random sample of 110 baseballs from the production line.
- In order to be confident that, 95 out of 100 times, the mean mass of the sample measures from 144.7 g to 145.3 g, the engineer needs to take a random sample of 65 baseballs from the production line.
- In order to be confident that, 90 out of 100 times, the mean mass of the sample measures from 144.7 g to 145.3 g, the engineer needs to take a random sample of 55 baseballs from the production line.

I interpreted each entry in the table.

c) For a constant margin of error, as the confidence level increases, the size of the sample needed to attain that confidence level increases. To have greater confidence that the baseballs meet quality standards, the engineer must use a larger sample.

I observed the trend in the table.

Your Turn

After making adjustments in equipment, the quality control engineer decided that the mean mass of baseballs must lie in the range 144.2 g to 146.4 g.

a) What margin of error is being used in the new sampling process?

b) What is the mean mass of a baseball that the engineer is trying to achieve?

c) Will the new baseballs meet regulation standards?

EXAMPLE 4 Analyzing statistical data to support a position

A poll was conducted to ask voters the following question: If an election were held today, whom would you vote for? The results indicated that 53% would vote for Smith and 47% would vote for Jones. The results were stated as being accurate within 3.8 percent points, 19 times out of 20. Who will win the election?

Kylie's Solution

Smith would have 53% of the votes, and Jones would have only 47% of the votes. Based on these numbers, Smith should win.

I examined the mean percent of votes that each candidate would receive, based on the poll.

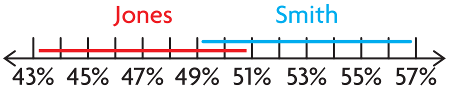
Percent of votes for Jones in the population:
47% ± 3.8%

Confidence interval:
43.2% to 50.8%

Percent of votes for Smith in the population:
53% ± 3.8%

Confidence interval:
49.2% to 56.8%

The margin of error is 3.8%. I used this value to determine the confidence interval for both candidates.



The two confidence intervals overlap from 49.2% to 50.8%.

I graphed the confidence intervals on a number line.

If the poll is accurate, Smith is more likely to win. However, there also is a chance that Jones will win, since the confidence intervals overlap by 1.6% of the votes.

If voters' opinions are the same on election day, Smith may receive only 49.5% of the votes and Jones could receive 50.5% of the votes.

Your Turn

Is it possible that Smith could receive more than 56.8% of the votes, according to this survey? Explain why or why not.

In Summary

Key Ideas

- It is often impractical, if not impossible, to obtain data for a complete population. Instead, random samples of the population are taken, and the mean and standard deviation of the data are determined. This information is then used to make predictions about the population.
- When data approximates a normal distribution, a confidence interval indicates the range in which the mean of any sample of data of a given size would be expected to lie, with a stated level of confidence. This confidence interval can then be used to estimate the range of the mean for the population.
- Sample size, confidence level, and population size determine the size of the confidence interval for a given confidence level.

Need to Know

- A confidence interval is expressed as the survey or poll result, plus or minus the margin of error.
- The margin of error increases as the confidence level increases (with a constant sample size). The sample size that is needed also increases as the confidence level increases (with a constant margin of error).
- The sample size affects the margin of error. A larger sample results in a smaller margin of error, assuming that the same confidence level is required.

  For example:

  - A sample of 1000 is considered to be accurate to within ±3.1%, 19 times out of 20.
  - A sample of 2000 is considered to be accurate to within ±2.2%, 19 times out of 20.
  - A sample of 3000 is considered to be accurate to within ±1.8%, 19 times out of 20.

CHECK Your Understanding

In the following questions, assume that the data approximates a normal distribution.

1. A poll determined that 81% of people who live in Canada know that climate change is affecting Inuit more than the rest of Canadians. The results of the survey are considered accurate within ±3.1 percent points, 19 times out of 20.

a) State the confidence level.
b) Determine the confidence interval.
c) The population of Canada was 33.5 million at the time of the survey. State the range of the number of people who knew that climate change is affecting Inuit more than the rest of Canadians.

2. A cereal company takes a random sample to check the masses of boxes of cereal. For a sample of 200 boxes, the mean mass is 542 g, with a margin of error of ±1.9 g. The result is considered accurate 95% of the time.

a) State the confidence interval for the mean mass of the cereal boxes.
b) Three other samples of different sizes were taken using the same confidence level, as shown at left, but the margin of error for each sample was mixed up. Match the correct margin of error with each sample size.

Sample Size	Margin of Error (g)
50	1.2
100	3.9
500	2.7

PRACTISING

3. An advertisement for a new toothpaste states that 64% of users reported better dental checkups. The results of the poll are accurate within 3.4 percent points, 9 times out of 10.

a) State the confidence level.
b) Determine the confidence interval.
c) If all 32 students in a mathematics class used this toothpaste, determine the range of the mean number of classmates who could expect better dental checkups.

4. In a 2006 Centre de recherche sur l'opinion publique (CROP) poll, 81% of Canadians indicated that they support bilingualism in Canada and that they want Canada to remain a bilingual country. This poll was reported accurate ±2.2%, 19 times out of 20.

a) Interpret the poll.
b) Mark claims that this poll must be flawed because if the majority of Canadians felt this way, then most people would speak both French and English, but they don't. Do you agree with Mark? Justify your decision.

5. The responses to another question in the poll from question 1 were summarized as follows: 58% of people living in Canada know that the cost of living for the average Inuit is 50% higher than the cost of living for other Canadians.

a) Determine the confidence interval for this question.

b) Predict the range of the mean number of people in your city or town who could have answered this question correctly.

6. Toxic materials, such as arsenic, lead, and mercury, can be released into the air if a discarded cellphone is incinerated. Toxins can be released into groundwater if a discarded cellphone ends up in a landfill. In a recent survey, 89% of those surveyed answered yes to the following question: Would you recycle your cellphone if it were convenient? The survey is considered accurate to within 4.3 percent points, 99 times out of 100.

a) Determine the confidence level and the confidence interval.

b) If 23 500 000 people in Canada own cellphones, state the range of the number of people who would indicate that they would recycle their cellphone if it were convenient.

7. a) Look in print or electronic media to find an example of a poll or survey that used a confidence level to report the results.

b) Determine the confidence interval.

c) Do you agree or disagree with any concluding statements that were made about the data from the poll or survey? Explain.

8. A company produces regulation ultimate discs. The discs have a mean mass of 175.0 g, with a standard deviation of 0.9 g. To ensure that few discs are rejected, the quality control manager must ensure that the mean mass of the discs lies in the acceptable range of 174.8 g to 175.2 g. During each shift, a random sample of discs is selected and the mass of each disc in the sample is measured. The table shown helps guide the sampling process.

Confidence Level	Sample Size Needed
90%	55
95%	78
99%	135

a) What is the confidence interval and margin of error this company is using for its quality control tests?

b) Approximately how many discs should be measured to ensure the mean mass is within ±0.2 g, 99% of the time?

c) The manager wants to save on labour costs by using a smaller sample. She knows that any discs that do not meet the regulation standards can be sold as recreational discs. Approximately how many discs should be measured to ensure that the mean mass is within ±0.2 g, 90% of the time?

d) Estimate the number of discs the company should measure to be confident that the mean mass of the ultimate discs lies in the acceptable range 19 times out of 20.

Math in Action

Finding and Interpreting Data in the Media

Research and polling firms regularly conduct surveys of samples of specific populations in areas such as agribusiness, food and animal health, energy, financial services, health care, higher education, lotteries and gaming, media and entertainment, retail, technology and communications, travel and tourism, and public opinion on political issues. Many companies and organizations rely on the data that is collected to make business and marketing decisions.

- With a partner or in a small group, look in print or electronic media for examples that cite confidence intervals or confidence levels. List the important statistical information that is given.
- Do you agree or disagree with the statements about the survey results in your examples? Use your knowledge of statistics to support your reasoning.
- What could be done to change the outcome of a future survey about one of these subjects? (For example, many different approaches have been taken to curb smoking addictions in Canada.)
- How much of a change would there have to be in the survey responses to make the outcome noticeably different?

9. Use confidence intervals to interpret each of the following statements.

a) In a recent survey, 54% of post-secondary graduates indicated that they expected to earn at least $100 000/year by the time they retire. The survey is considered accurate within ±4.5%, 9 times in 10.

b) A market research firm found that among online shoppers, 63% search for online coupons or deals when they purchase something on the Internet. The survey is considered accurate within ±2.1 percent points, 99% of the time.

c) A recent report indicated that Canadians spend an average of 18.1 h per week online, compared with 16.9 h per week watching television. The results are considered accurate with a margin of error of ±3.38%, 19 times out of 20.

d) A survey conducted at the expense of the political party that holds office indicated that 39% of decided voters said they would not vote for candidates of that party in the next election. The result is considered accurate within ±3%, 95% of the time.

Closing

10. Explain why, for a given confidence level,

a) the margin of error decreases as the sample size increases

b) the margin of error increases as the confidence level increases

Extending

11. As sample size increases, the margin of error, expressed as a percent, decreases. Consider the table below.

Sample Size	Margin of Error (%)
100	9.80
400	4.90
900	3.27
1600	2.45
2500	1.96
3600	1.63

a) What mathematical relationship exists between increased sample size and the margin of error?

b) What would be the margin of error for a sample size of

i) 4900

ii) 2000

c) Use your results from parts a) and b) to explain why a relatively small sample will give a fairly accurate indication of the trend for an entire population.

5 Chapter Self-Test

1. Students recorded their heights, in inches, when they graduated from kindergarten in 1999 and again when they graduated from high school in 2011.

1999:	39	41	41	43	45	46	47	46	48	47	44	38	41	39	43	46	44
2011:	60	74	76	62	64	61	66	68	71	76	74	73	72	69	64	63	60

a) Determine the mean and standard deviation for each year.

b) Compare the heights for the two years. Which set of heights has a greater standard deviation? Describe some of the possible reasons for this greater deviation.

2. The chest circumferences of Scottish militiamen in the 19th century are given in the frequency table to the right.

a) Are the chest circumferences normally distributed? Explain.

b) Determine the *z*-score for a Scottish militiaman with a 42 in. chest circumference.

Chest Circumference (in.)	Frequency
33	3
34	18
35	81
36	185
37	420
38	749
39	1073
40	1079
41	934
42	658
43	370
44	92
45	50
46	21
47	4
48	1

3. Brenda searched the Environment Canada website and found that the mean daily temperature in Edmonton in March is −2.6 °C, with a standard deviation of 3.2 °C. The mean daily temperature in Calgary in March is −1.9 °C, with a standard deviation of 2.8 °C. Compare the temperatures at these two locations in March.

4. In a Canada Day poll, 1009 Canadians were asked "What are things about Canada that make you proud?" 88% of respondents said the flag, 80% said hockey, and 44% said the Canadian justice system. The poll was reported accurate to within ±3.1%, 19 times out of 20.

a) Use confidence intervals to interpret the poll results.

b) At the time of the poll, Statistics Canada estimated Canada's population at 34 019 000. Determine the range of people in Canada, based on this poll, who would answer hockey makes them proud of Canada.

c) If the polling company conducted this same survey using the same sample size, but used a confidence level of 99%, what would happen to the margin of error? Explain.

WHAT DO You Think Now? Revisit **What Do You Think?** on page 209. How have your answers and explanations changed?

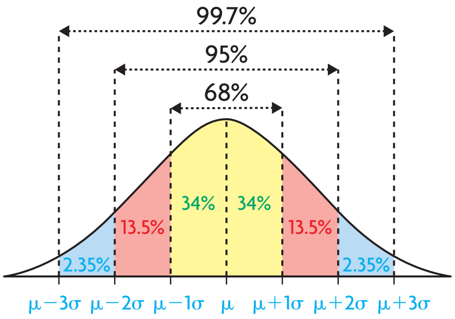
5 Chapter Review

FREQUENTLY ASKED Questions

Study **Aid**

- See Lesson 5.4, Examples 1 and 2.
- Try Chapter Review Questions 7 and 8.

Q: What is a normal distribution, and what are its properties?



A: When data is normally distributed, 50% of the data is above the mean and 50% is below the mean. This makes the distribution symmetrical. The measures of central tendency are equal or close to each other. The graph of a normal distribution (data values versus frequency) is a bell curve. For a normal distribution, approximately 68% of the data is within one standard deviation of the mean, 95% is within two standard deviations of the mean, and 99.7% is within three standard deviations of the mean.

Study **Aid**

- See Lesson 5.5, Examples 1, 2, and 3.
- Try Chapter Review Questions 9 and 10.

Q: What is a *z*-score, and how do I calculate it?

A: A *z*-score indicates the distance of a data value from the mean of the set, measured in standard deviations. If the *z*-score is positive, the data value is greater than the mean. If the *z*-score is negative, the data value is less than the mean.

For example, a *z*-score of 2.00 means that the data value is 2.00 standard deviations above the mean.

To calculate a *z*-score, use the following formula:

$$z = \frac{x - \mu}{\sigma}$$

To determine the *z*-score, given $x = 23.5$, $\mu = 18.6$, and $\sigma = 3.2$, substitute each value into the *z*-score formula.

$$z = \frac{23.5 - 18.6}{3.2}$$

$$z = 1.531...$$

The value is about 1.53 standard deviations above the mean.

Q: How do I compare two values from two normally distributed sets of data?

A: Determine the *z*-score of each piece of data. The value with the higher *z*-score is the greater relative value.

City	μ (\$)	σ (\$)
Edmonton	375 000	75 000
Calgary	415 000	80 000

For example, Max sells his house in Edmonton for \$392 000 and purchases a house in Calgary for \$417 000. The mean and standard deviations for houses in each city are shown in the table above.

Edmonton:	Calgary:
$z = \frac{392\,000 - 375\,000}{75\,000}$	$z = \frac{417\,000 - 415\,000}{80\,000}$
$z = 0.226...$	$z = 0.025$

The house in Edmonton has the greater relative value because the *z*-score is higher.

Study **Aid**

- See Lesson 5.5, Example 1.
- Try Chapter Review Question 10.

Q: What is the difference between margin of error, confidence interval, and confidence level?

A: The purpose of a poll or survey is to gather information that can be used to make predictions about a population.

For example, in a recent telephone poll, 33% of Canadians, 18 years of age and older, thought that Olympic athletes who were caught using performance-enhancing drugs should be banned from competition for life (Nanos National Poll, Dec. 2009). The results were accurate to within 3.1 percent points, 19 times out of 20.

The margin of error is ± 3.1%, which indicates the sampling error in the poll. The margin of error can be combined with the result of the poll to generate a confidence interval. For this poll, we expect that if the entire population of Canadians, 18 years of age and older, were asked the same question, between 29.9% and 36.1% would indicate that they want drug-using athletes banned.

The confidence level of the poll is stated as 19 times out of 20, which is equivalent to 95%. If this poll were conducted over and over again, 95% of the time the result would fall within the confidence interval, 29.9% to 36.1%.

Study **Aid**

- See Lesson 5.6, Examples 1 and 4.
- Try Chapter Review Questions 11 and 12.

PRACTISING

Lesson 5.1

1. Twila and Amber keep a log of the amount of time, in minutes, they spend on homework each school day for two weeks. Determine the mean and range for each girl's data, and compare the two sets of data.

Twila:	45	55	50	40	55	40	60	45	40	35
Amber:	80	10	65	15	75	30	40	85	20	35

Lesson 5.2

2. Melody is comparing education levels of her generation with education levels of her parents' generation. She obtained the data in the table. Draw two frequency polygons on the same graph to compare the education levels. Comment on the results.

Level of Education	People 25 to 34 Years Old (%)	People 55 to 64 Years Old (%)
less than high school	11	23
high school diploma	23	24
trades certificate	10	13
college diploma	23	16
university certificate or diploma	5	6
university degree	29	18

Lesson 5.3

3. a) Predict which girl's data in question 1 will have the lowest standard deviation. Justify your answer.

b) Determine the standard deviation for each girl's data. Was your prediction correct?

4. The following data was taken from a 2000 federal government survey on the mean salary in each province for three categories.

Education	**Salary (thousands of dollars)**												
Province	**NL**	**PE**	**NS**	**NB**	**QC**	**ON**	**MB**	**SK**	**AB**	**BC**	**YT**	**NT**	**NU**
no diploma	16	15	18	17	21	23	19	18	22	22	19	20	15
high school	17	18	21	20	24	28	23	22	26	26	26	32	27
post-secondary	31	29	33	32	35	44	34	33	41	38	38	48	43

a) Determine the mean and standard deviation for each level of education.

b) Which level of education yields the highest mean salary?

c) Which level of education has the greatest variability in salary?

5. Marc usually puts a bag of either sunflower seeds or raisins in his lunch. The first table shows the number of sunflower seeds in the last 30 bags of sunflower seeds that Marc has had in his lunch. The second table shows the number of raisins in the last 30 bags of raisins that Marc has had in his lunch. Is Marc more likely to get the mean number of items in a bag of sunflower seeds or a bag of raisins? Justify your thinking.

Sunflower Seeds per Bag	28	29	30	31	32	33
Frequency	2	4	11	9	3	1

Raisins per Bag	27	28	29	30	31
Frequency	1	12	7	3	7

6. Scientists monitor the masses of polar bears. In 2010, the following data was obtained:

Adult Female	$\bar{x} = 247$ kg	$\sigma = 33$ kg
Adult Male	$\bar{x} = 461$ kg	$\sigma = 51$ kg

The masses of two polar bears were measured. The female had a mass of 277 kg, and the male had a mass of 499 kg. Use *z*-scores to determine which bear had the greater mass compared with other bears of the same sex.

Polar bears go ashore when the sea ice melts. If the sea ice melts too early in the year, the polar bears must go ashore before they are ready. This forces them into a prolonged fast. Global warming may result in polar bears becoming too thin to reproduce.

Lesson 5.4

7. Judy always waits until her gas tank is nearly empty before refuelling. She keeps track of the distance she drives on each tank of gas. The distance varies depending on the weather and the amount she drives on the highway. The distance has a mean of 824 km and a standard deviation of 28 km.

   a) Sketch a normal curve to show the distribution of the driving distances for a tank of gas. Mark the kilometres driven for values that are 1, 2, and 3 standard deviations from the mean.

   b) What percent of the time does Judy drive between 796 km and 852 km on a tank of gas?

   c) What percent of the time does she drive between 740 km and 796 km on a tank of gas?

   d) Between what two values will she drive 95% of the time?

8. The body temperatures of 130 adults are recorded in the frequency table to the right.

   a) Determine the mean and standard deviation of the data.

   b) Are the temperatures normally distributed? Explain.

Temperature (°C)	Frequency
35.8	2
36.0	3
36.2	5
36.4	11
36.6	14
36.8	29
37.0	27
37.2	20
37.4	13
37.6	3
37.8	2
38.0	0
38.2	0
38.4	1

Lesson 5.5

9. TJ is a Congo African Grey parrot. This species of parrot has a life expectancy of 50 years, with a standard deviation of 8 years. What is the likelihood that TJ will live over 60 years?

10. *Computers For All* offers an extended 3-year replacement warranty on its computers. The mean lifespan of its computers is 3.8 years, with a standard deviation of 0.45. *Everything Electronic* offers a 2-year replacement warranty on its computers. The mean lifespan of an *Everything Electronic* computer is 2.6 years, with a standard deviation of 0.31. Which computer is more likely to fail before its warranty period is over?

Lesson 5.6

11. A poll was conducted to determine where Canadians obtain health-related information. 61.9% said they research information on the Internet, 68.9% said they ask friends or family, and 17.9% said they call a health line. The results of this survey are considered accurate within ±1.4 percent points, 99 times out of 100.

a) Determine the confidence interval for each information source.

b) In a city with an adult population of 345 000, predict the range of the number of adults who will say they use each source.

12. Two different market research companies conducted a survey on the same issue. Company A used a 99% confidence level and company B used a 95% confidence level.

a) If both companies used a sample size of 1000, what does this imply about the margin of error for each survey?

b) If both companies used the same margin of error of ±2%, what does this imply about the sample size for each survey?

True-False Tests

Sometimes you may have to make a decision based on limited or no knowledge.

Have you ever wondered what would happen if you guessed all the answers on a true-false test? Would it matter how many questions were on the test?

? How can you determine the likelihood of passing a true-false test if you guessed all the answers?

Part 1: Generate the Data

A. Write a short true-false question about yourself. Make sure that none of your classmates would know the answer to your question. Example: "The name of my first pet was Fido" or "I had spaghetti for dinner last Sunday." Give your question, along with the correct answer, to your teacher.

B. As a class, take the "test." Your teacher will read each question, and you will write "true" or "false" on a sheet of paper. Leave the answer to the question that you created blank. Your teacher will add one question at the end of the test.

C. Mark the tests as a class, and convert each test score to a percent. Record each person's score on the board, both with and without your teacher's question included.

Part 2: Analyze the Data

D. Determine the mean and standard deviation of the data that does not include your teacher's question. Is the data normally distributed? Explain.

E. Repeat step D for the data that includes your teacher's question.

F. Compare the results for the two sets of data. What do you notice?

G. Is it likely that you would pass a true-false test if you guessed all the answers?

Task ***Checklist***

✔ Did you present the data effectively?

✔ Did you use appropriate mathematical language?

✔ Were your conclusions presented clearly and concisely?

5 Project Connection

Analyzing Your Data

Statistical tools can help you analyze and interpret the data you collect. You need to think carefully about which statistical tool to use and when, because other people will be scrutinizing your data. A summary of relevant tools is given below.

Measures of Central Tendency

Selecting which measure of central tendency (mean, median, or mode) to use depends on the distribution of your data. As the researcher, you must decide which measure most accurately describes the tendencies of the population. Consider the following criteria when you are deciding which measure of central tendency best describes your set of data.

- Outliers affect the mean the most. If the data includes outliers, use the median to avoid misrepresenting the data. If you want to use the mean, remove the outliers before calculating the mean.
- If the distribution of the data is not symmetrical, but instead strongly skewed, the median may best represent the set of data.
- If the distribution of the data is roughly symmetrical, the mean and median will be close, so either may be appropriate to use.
- If the data is not numeric (for example, colour), or if the frequency of the data is more important than the values, use the mode.

Measures of Dispersion

Both the range and the standard deviation give you information about the distribution of the data in a set.

The range of a set of data changes considerably because of outliers. The disadvantage of using range is that it does not show where most of the data in a set lies—it only shows the spread between the highest and lowest values. The range is an informative tool that can be used to supplement other measures, such as standard deviation, but it is rarely used as the only measure of dispersion.

Standard deviation is the measure of dispersion that is most commonly used in statistical analysis when the mean is used to calculate central tendency. It measures the spread relative to the mean for most of the data in the set.

Outliers can affect standard deviation significantly. Standard deviation is a very useful measure of spread for symmetrical distributions with no outliers.

Standard deviation helps with comparing the spread of two sets of data that have approximately the same mean. The set of data with the smaller standard deviation has a narrower spread of measurements around the mean, and therefore usually has comparatively fewer high or low values.

Normal Distribution and *Z*-Scores

When working with several sets of data that approximate normal distributions, you can use *z*-scores to compare the data values. A *z*-score table enables you to find the area under a normal distribution curve with a mean of zero and a standard deviation of one. To determine the *z*-score for any data value in a set that is normally distributed, you can use the formula

$$z = \frac{x - \mu}{\sigma}$$

where x is any observed data value, μ is the mean of the set, and σ is the standard deviation of the set.

Margin of Error and Confidence Level

When analyzing the results of a survey, you may need to interpret and explain the significance of some additional statistics. Most surveys and polls draw their conclusions from a sample of a larger group. The margin of error and the confidence level indicate how well the sample represents the larger group. For example, a survey may have a margin of error of plus or minus 3% at a 95% level of confidence. This means that if the survey were conducted 100 times, the data would be within three percent points above or below the reported results in 95 of the 100 surveys.

The size of the sample that is used for a poll affects the margin of error. If you are collecting data, consider the size of the sample you need for a desired margin of error.

Project Example — Choosing tools to analyze your data

Sarah chose the changes in population of the Western provinces and the territories over the last century as her topic. Below, she describes how she determined which statistical tools to use.

Sarah's Analysis

I obtained my data from a government census. Since a census surveys the entire population and not a sample, I do not need to consider margin of error or confidence level. I am using time series data, which shows trends from 1900 to 2000. The data is not normally distributed, so I do not need to use *z*-scores.

I could use a measure of central tendency to represent the "average" population of each province or territory over this period. I am not interested in frequency, so the mode is not appropriate. I think the mean would be the best measure to use in this situation. I could also look at the spread in population for each province or territory over this period using range and standard deviation. These values may allow me to compare the populations of the provinces and territories.

Your Turn

A. Which statistical tools are appropriate for your data? Explain why.

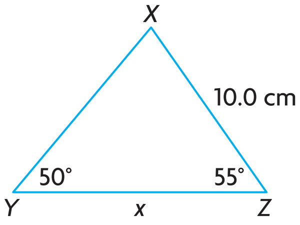
B. Use the tools you selected, and calculate the statistics.

C. Use these statistics to analyze your data.

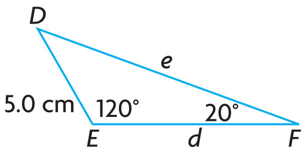
3-5 Cumulative Review

1. Determine the measure of all indicated sides to the nearest tenth of a unit.

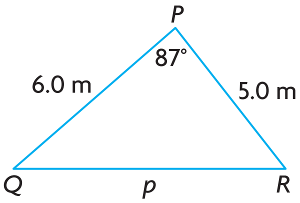
a)



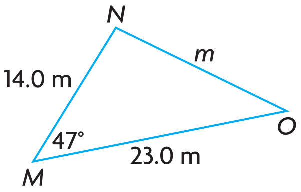
c)



b)

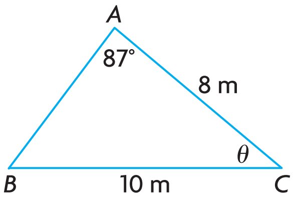


d)

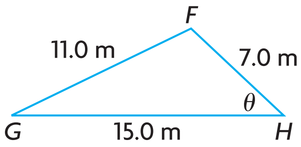


2. Determine the measure of each indicated angle to the nearest degree.

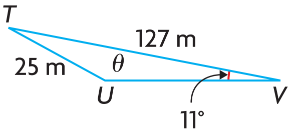
a)



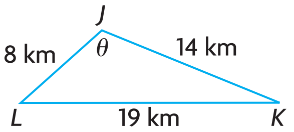
c)



b)



d)



3. In △ *HIJ*, ∠ *I* = 48°, *i* = 9 cm, and *j* = 11 cm. Solve △ *HIJ*. Round your answers to the nearest tenth of a unit.

4. In △ *DEF*, ∠ *D* is 58°, *e* is 10.0 cm, and *f* is 14.0 cm. Solve △ *DEF*. Round your answers to the nearest tenth of a unit.

5. Mohammed has been driving his ATV on the Vedder Mountain Trail System, near Chilliwack, British Columbia, for 3.2 km. He has been travelling in a compass direction of N54°E. He uses his compass to change direction to a new course of S5°W and continues for 4.6 km. If Mohammed wants to return directly to his starting point, how far must he travel, to the nearest tenth of a kilometre? In which direction should he travel, to the nearest tenth of a degree?

6. On a 520 m hole, a golfer's tee shot travels 175 m, 17° to the right of the direct path to the flag. The golfer's second shot flies 15° farther to the right, but only travels 150 m. How far, to the nearest metre, is the golfball from the flag?

7. In $\triangle QRS$, $\angle Q$ is acute. Explain, with the help of a diagram, the relationship between $\angle Q$, sides q and r, and the height of the triangle, for each of the following situations to occur.

a) No triangle is possible.
b) Only one type of triangle is possible.
c) Two types of triangles are possible.

8. Caitlin wants to determine the height of a tree on the opposite bank of a river. She starts by laying out a baseline that is 100 m long. Then she estimates the angles from the ends of the baseline to the base of the tree as 80° and 30°. From the end of the baseline with the 80° angle, she estimates the angle of elevation to the top of the tree as 20°.

a) Sketch a diagram to model this situation.
b) Determine the height of the tree, to the nearest tenth of a metre.

9. In a study of the longevity of a particular breed of dog, veterinarians recorded the lifespans of 30 dogs.

Lifespans of Dogs (years)				
12.9	13.2	14.1	13.9	12.8
13.1	13.2	13.6	13.0	13.4
12.9	13.3	11.8	12.8	14.6
10.4	14.8	11.5	13.5	13.6
9.6	14.5	13.5	13.8	14.4
13.1	13.6	12.8	12.9	13.3

a) Create a frequency table and histogram for the data.
b) Does the data approximate a normal distribution? Explain.
c) Determine the range and standard deviation of the data. Describe what these measures tell you about the data.

10. The average daily temperature in Winnipeg, Manitoba, during the month of January is −17.8 °C, with a standard deviation of 3.9 °C. The average daily temperature in Whitehorse, Yukon, during the month of January is −17.7 °C, with a standard deviation of 7.3 °C. Compare the temperatures at these two locations in January.

11. Zac is 195 cm tall. In a recent survey of students at his school, it was determined that the heights of the students are normally distributed, with a mean of 170 cm and a standard deviation of 12.5 cm.

a) What percent of the students at Zac's school are shorter than Zac?
b) What percent of the students are taller than Zac?

12. A manufacturer of smart phones has created a new phone model. The mean life of this new model is 48 months, with a standard deviation of 12 months. The manufacturer has offered a 24-month warranty on this model.

a) Determine the percent of phones that are expected to malfunction during the warranty period.
b) What percent of phones is expected to malfunction during the second and third year of use?

13. From May 29 to June 3, 2010, Nanos Research conducted a random telephone survey of 1008 Canadians, 18 years of age and older, to ask the following question: What are the most important issues facing Canadians today? The responses are shown in the table.

a) What is the margin of error for this survey?
b) Determine the confidence level for this survey.
c) State the confidence interval for each of the following responses.
i) health care
ii) environment

Responses	(%)*
health care	23.1
jobs/economy	19.2
environment	12.6
high taxes	5.3
education	2.5
unsure	13.3

*Percent values may not add to 100 due to rounding.

This survey is accurate, plus or minus 3.1 percent points, 19 times out of 20.

14. Explain why the confidence level for a poll or survey is decreased when

a) the margin of error decreases for a specific sample size
b) the sample size that is needed for a specific margin of error decreases