

# 3

## Two-Variable Data



### What You'll Learn

To distinguish types of data, and to analyse and represent two-variable data from primary and secondary sources

### And Why

Analysing data to look for relationships is part of many college courses and professions. Fishery and forestry managers, sports trainers, medical researchers, and lab technicians all work with two-variable data.

### Key Words

- variable
- one-variable data
- two-variable data
- scatter plot
- dependent variable
- independent variable
- correlation
- cause-and-effect relationship
- line of best fit
- outlier
- interpolation
- extrapolation
- non-linear data
- linear correlation
- correlation coefficient
- conjecture

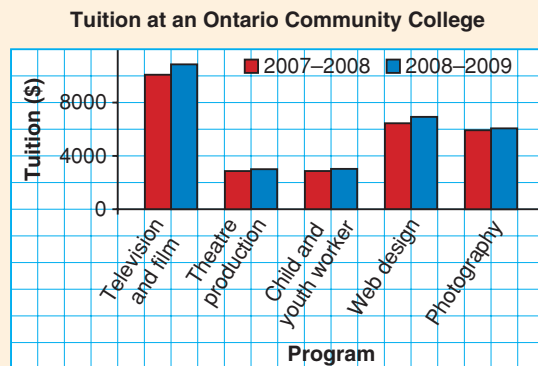
## Interpreting Data Graphs

Prior Knowledge for 3.1

Data can be presented in a variety of ways. Graphical representations can include bar graphs, histograms, scatter plots, and circle graphs.

**Example**

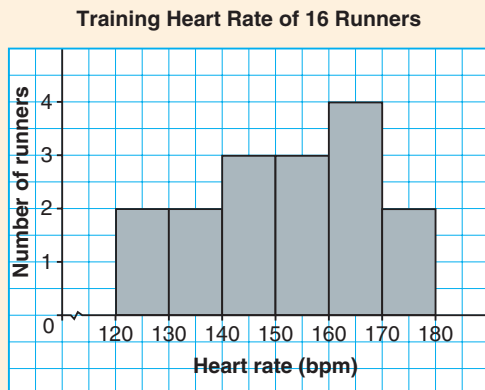
- What do the red bars in this graph represent? The blue bars?
- For which program did the tuition change the most?
- Estimate the tuition for photography in 2007–2008; in 2008–2009.

**Solution**

- The red bars represent the tuition for programs at an Ontario community college in 2007–2008. The blue bars represent the tuition for the same programs in 2008–2009.
- Look at the graph for the pair of bars with the greatest difference in height: television and film is the program with the greatest increase in tuition.
- The tuition for photography was about \$5900 in 2007–2008; about \$6100 in 2008–2009.

**CHECK** ✓

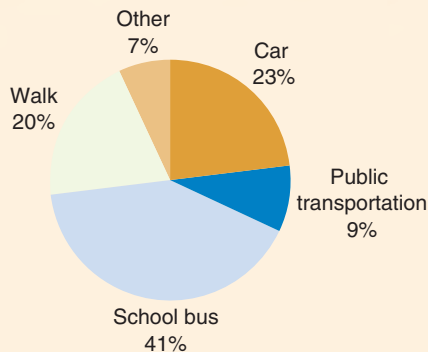
- The histogram shows the training heart rates, in beats per minute (bpm), of 16 runners.
  - How many runners had heart rates between 140 and 149 bpm?
  - What interval was the most common?
  - What were the minimum and maximum heart rates any of these runners could have had?



2. Miguel surveyed 250 high school students about how they usually get to school. He displayed the data in this circle graph.

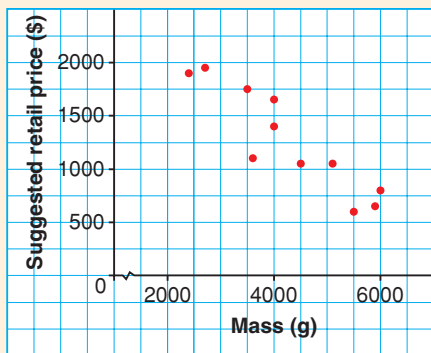
- What does the light blue area represent?  
Is it more or less than half the graph?
- What conclusions might you make from this graph?

Usual Transportation for High School Students



3. This scatter plot compares the mass and price of a selection of laptops.

Price and Mass of Laptop Computers



- What does each point show?
  - How many of the laptops have mass less than 4000 g?  
How many of these cost less than \$1250?
  - How many of the laptops have mass greater than 4000 g?  
How many of these cost less than \$1250?
4. One hundred students were asked to identify their leisure activities.

Activity	Number of students
Reading	12
Playing sports	32
Watching TV	83
Visiting friends or family	55
Hobbies	41

- What is the sum of the numbers in the second column of the table?
- Why would a circle graph not be a good tool for displaying these data?
- What type of graph would be good for these data? Why do you think so?

**Slope** is a measure of the steepness of a line.

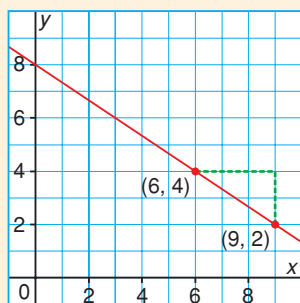
- A positive slope means the line goes up to the right.
- A negative slope means the line goes down to the right.

To determine the slope of a line, we need the coordinates of two points on the line.

A line with slope  $m$  and  $y$ -intercept  $b$  can be represented by the equation  $y = mx + b$ .

### Example

- a) Determine the slope of the line shown.
- b) Write the equation of the line in the form  $y = mx + b$ .

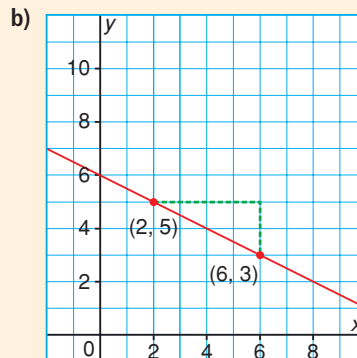
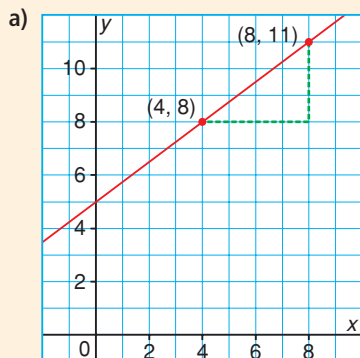


### Solution

- a) Choose two points on the line:  $(6, 4)$  and  $(9, 2)$   
 Slope:  $\frac{\text{rise}}{\text{run}} = \frac{2 - 4}{9 - 6}$ , or  $-\frac{2}{3}$
- b) From the graph, the  $y$ -intercept is 8. The slope is  $-\frac{2}{3}$ .  
 So, the equation of the line is  $y = -\frac{2}{3}x + 8$ .

### CHECK ✓

1. Determine the slope of each line. Write the equation of each line in the form  $y = mx + b$ .





## Getting Extra Practice

# Transitions

People in diverse fields have shown that regular practice is a key part of their success. Consider the practice associated with:

- sports
- music
- acting
- firefighting
- computer applications
- aviation

Practice helps you learn to carry out procedures efficiently and accurately.

### Strategies for Success

- Practise often, every day if possible.
- Push yourself beyond your current level of competence.
- Focus as you practise.
- Reflect on what you are doing well, and where you want to improve.

To practise the concepts and skills presented in your math class:

- Try a variety of questions to deepen your understanding.
- Once you feel confident doing questions at one level, try some at the next level.
- Create new problems, then solve them or ask classmates to solve them.
- Spend some time practising with other people. Compare answers to discover new approaches.

To get ready for a test or for the next chapter:

- Read the *Study Guide* at the end of the chapter.
- Try the *Practice Test*.
- Choose questions from each lesson, mixing up the order.
- Use the Internet to find some interactive mathematics questions. Try search words related to the math and technology in the chapter.
- Find data and graphs in newspapers, magazines, and Internet news sites. Create and solve your own questions about them. Trade questions and compare solutions.

#### Search words

- Virtual manipulative
- Java applet
- Fathom tutorial

# 3.1

## One- and Two-Variable Data

One application may have a wide variety of data that are helpful to gather, record, and analyse. What sorts of data do you think about when following a sport like women's hockey?



### Investigate

### Interpreting and Comparing Data

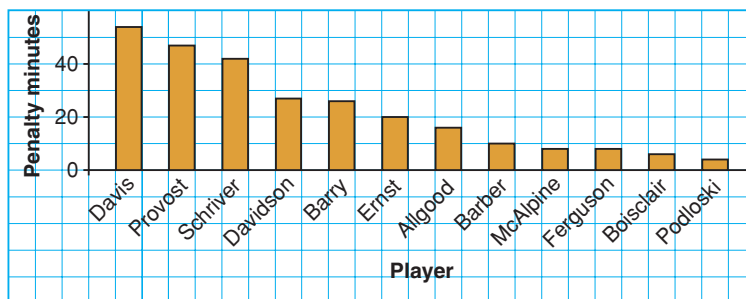
Work with a partner.

The graphs and table provide information about Canadian women's university hockey points leaders in the 2006–2007 season.

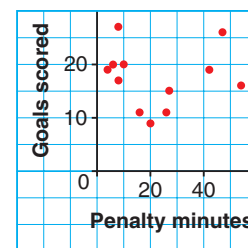
- Create a question that can be answered:
  - from the bar graph
  - from the scatter plot
  - from the table

Record each question and your answer.

Penalty Minutes for Top Scorers in Women's University Hockey



Penalty Minutes and Goals Scored



Player	Games played	Goals	Assists	Points	Penalty minutes
Lindsay McAlpine	24	27	30	57	8
Tarin Podloski	22	19	31	50	4
Mariève Provost	21	26	21	47	47
Valerie Boisclair	21	20	21	41	6
Jenna Barber	24	20	20	40	10
Courtney Schriver	21	19	16	35	42
Christina Davis	20	16	17	33	54
Candice Ernst	18	9	24	33	20
Kate Allgood	24	11	20	31	16
Brayden Ferguson	19	17	14	31	27
Vanessa Davidson	17	15	16	31	27
Taryn Barry	24	11	19	30	26

### Reflect

- Exchange questions with another pair of students. Answer each others' questions.
- Only one graph displays two-variable data. Which one do you think it is? Why?

## Connect the Ideas

### One-variable and two-variable data

In statistics, a **variable** is an attribute that can be measured.

**One-variable data** sets give measures of one attribute. You can recognize one-variable situations when you see:

- Tally charts
- Frequency tables
- Bar graphs
- Histograms
- Pictographs
- Circle graphs

One-variable data can be analysed using mean, median, or mode.

**Two-variable data** sets give measures of two attributes for each item in a sample. You can recognize two-variable situations when you see:

- Ordered pairs
- Scatter plots
- Two-column tables of values

### Example 1

### Identifying Situations Involving One- and Two-Variable Data

State whether each situation involves one-variable or two-variable data.

Justify your answers.

- a) Noah researches annual hours of sunshine in Canadian cities.
- b) A study compared the length of time children spend playing video games and the time they spend reading.

#### **Solution**

- a) Noah's research could be analysed using mean, median, or mode. So, the situation involves one-variable data.
- b) The study involves two pieces of information for each child, time spent on video games and time spent reading. These data could be represented in a scatter plot. So, the study involves two-variable data.

### Example 2

### Deciding Which Type of Graph to Draw

For a class project, Dylan surveyed students about their part-time jobs.

Student	Hours Spent at Part-Time Job	
	During the week (h)	On the weekend (h)
Adil	0.0	18.0
Anya	5.0	12.5
Ellen	8.0	12.0
Fiona	17.0	8.0
Aaron	0.0	16.5
Leila	10.0	16.0
Mason	9.5	8.0
Petra	15.0	6.0

- a) What type of graph would be best to show how many hours each student worked on the weekend? Justify your choice. Does the graph display one-variable or two-variable data?
- b) What type of graph would best show a possible relationship between weekday and weekend hours? Justify your choice.



### Solution

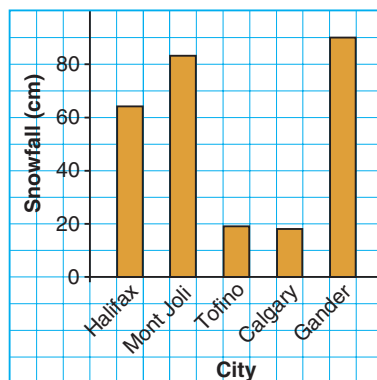
- a) A bar graph would display all the information together so that the reader can easily compare the number of hours for each student. One piece of data would be displayed for each student. So, the graph would display one-variable data.
- b) A scatter plot could show a possible relationship between weekday hours and weekend hours. Since each point on the scatter plot would display two pieces of information about a student, the graph would reflect two-variable data.

## Practice

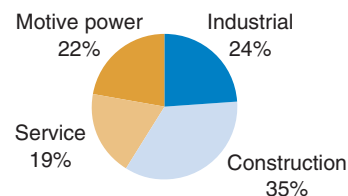
A

1. a) Does each graph illustrate one-variable or two-variable data?

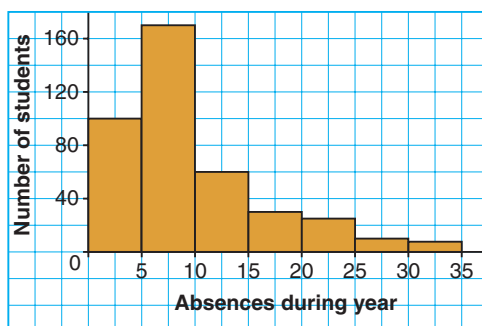
i) Snowfall in January



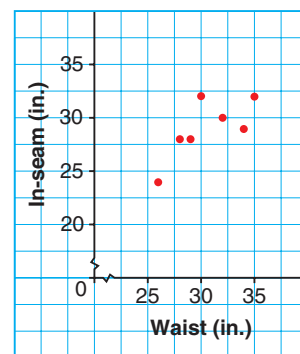
ii) Destination of Apprenticing Students



iii) Frequency of Student Absences



iv) Pant Measurements



- b) Choose one graph from part a. How did you decide whether the graph showed one- or two-variable data?

2. a) Does each table illustrate one-variable or two-variable data?

i)

Household size	Number of TVs in household
1	2
2	3
3	3
4	3
5	4

ii)

Student	Dollars
Anne	15
Lars	25
Mason	5
Thom	20
Riaz	25
Loni	10

iii)

Candle burn time (h)	10	15	25	40
Cost (\$)	7.49	10.99	15.49	19.99

b) Choose one table from part a. How did you decide whether the table showed one- or two-variable data?

3. Identify the two variables in each situation.

- The more purchases made with a credit card, the more reward points earned.
- Anthropology students read an article that claimed that people with greater brain mass have higher IQs.
- Across Ontario, the mosquito population remained low due to the lower than average rainfall.

4. a) State whether each situation involves one-variable or two-variable data.

- Marcus calculates the median mark of the class on an exam.
  - An article discusses the possible link between prolonged cell phone use and the increased probability of brain cancer.
  - A classroom survey shows that 70% of the students plan to attend university, 15% plan to attend college, 10% are going directly into the workplace, and 5% are uncertain.
- b) Choose one situation from part a. Explain how you decided whether the situation involved one- or two-variable data.

The median is the middle value when data are arranged in numerical order. If there are two middle values, their mean is the median.

■ For help with question 3, see Example 1.

**B**

For help with question 5, see Example 2.

5. What type of graph would you use to display the data in each table? Justify your choices.

a)

Number of Sit-Ups Students Can Do in 1 min					
Number of sit-ups	0–9	10–19	20–29	30–39	40–49
Frequency	0	3	5	8	4

b)

Land Area of Selected Provinces/Territories							
Province or territory	Alberta	Manitoba	Ontario	Quebec	Nova Scotia	PEI	Nunavut
Land area (1000 km <sup>2</sup> )	642	554	918	1365	53	6	1936

c)

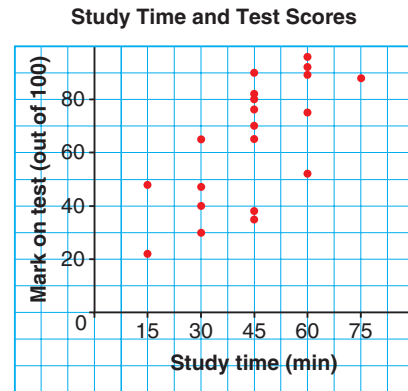
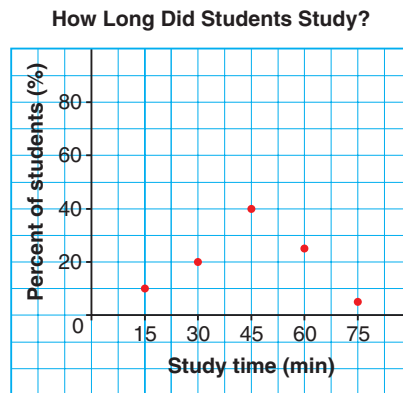
Ages of Selected Students by Grade							
Grade	9	11	12	9	10	10	11
Age	15	18	18	14	15	16	16

6. A company manufactures fuses. The quality control department frequently tests sample lots to determine how many fuses are defective. This chart shows the results of testing over several days.

Quality Control Results								
Sample size	50	50	100	150	150	200	200	200
Defective fuses	2	1	3	3	4	3	5	4

- a) Suppose the supervisor wants a graph showing the frequency of each sample size. What type of graph would be best? Why?
- b) What kind of graph would you use to display all the data in the table? Justify your choice.
- c) If you were in charge of quality control for this company, which of the samples would concern you most? Least? Explain.
7. Cheyenne recorded the foot lengths of 10 students in her math class. Ayub recorded the heights of 10 students in his English class. They created a scatter plot with the variables foot length and height to determine if there is a relationship between them.
- a) Why are the data they collected not two-variable data?
- b) What kind of analysis could they do with their data?

8. A teacher surveyed her students about how long they had studied for a test.
- a) Which graph displays two-variable data? Justify your answer.



- b) Which graph provides information about a possible relationship? What variables does the relationship involve?
- c) For each graph, write a question someone could answer using the data in the graph. Answer your questions.
9. **Literacy in Math** Select a graph from this lesson. Describe one piece of information you can learn from the graph.
10. **Assessment Focus** A school basketball team had these data for a game.

Player	Minutes played	Points scored				Total
		1st quarter	2nd quarter	3rd quarter	4th quarter	
Willans	23	2	6	4	4	16
Ohira	11	3	1	3	0	7
Jasquith	19	5	2	2	6	15
Meyers	14	0	0	1	2	3
Salinski	28	5	5	4	4	18
Tobin	26	6	4	6	4	20
Ramanathan	8	0	2	2	0	4
Olander-Hinns	10	2	0	0	2	4
Leenders	5	2	0	0	0	2
Wardhaugh	16	4	2	2	2	10
<b>Total</b>		29	22	24	24	

- a) Suppose you calculated the mean number of minutes played by a team member in the game. Is this one-variable or two-variable statistics? Explain.
- b) Suppose you want to create a graph to display the number of points scored in each quarter. Would the graph display one- or two-variable data? What type of graph would you create? Explain your thinking.
- c) Suppose you want to create a graph to display the number of minutes a player played in the game and the number of points the player scored. Would the graph display one- or two-variable data? What type of graph would you create? Explain your thinking.



**C**

11. a) Construct the graphs you described in parts b and c of question 10.
- b) Write a question about the graphs that would involve one-variable analysis.
- c) Write a question about the graphs that would involve two-variable analysis.
- d) Trade graphs and questions with a classmate. Answer your classmate's questions. Check each other's work.

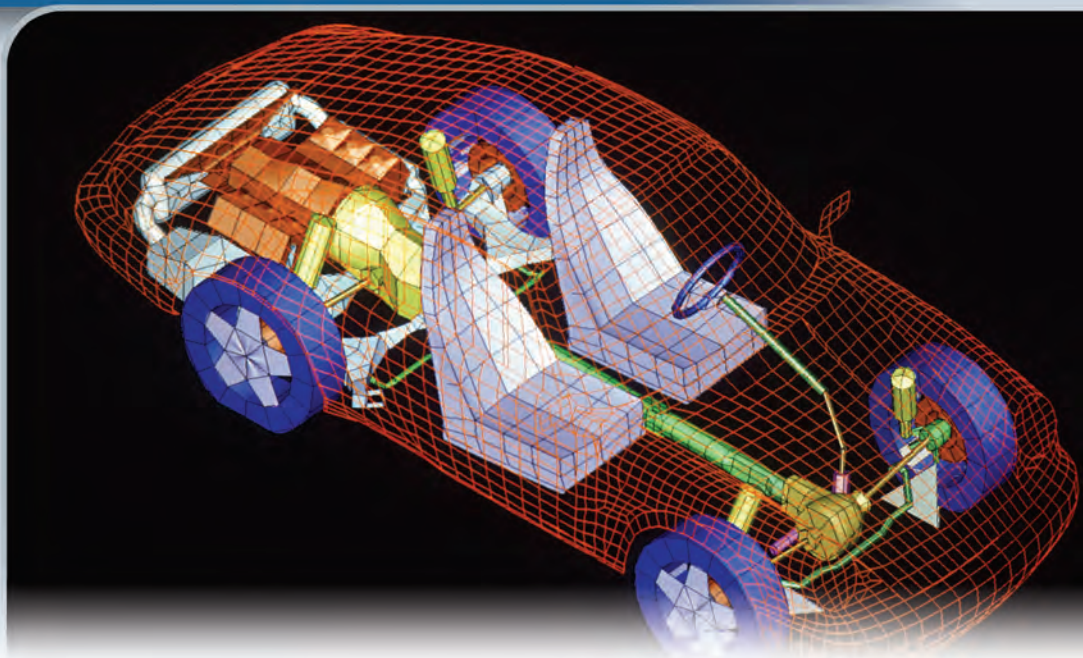
### *In Your Own Words*

Describe a situation where you would need to collect, analyse, and graph two-variable data.

# 3.2

## Using Scatter Plots to Identify Relationships

*Ergonomic designs suit the way humans think, see, and move. They are based on the measurement and study of human body dimensions. Ergonomics is studied in college programs such as industrial design, fitness, and lifestyle management.*



### Investigate

#### Materials

- grid paper
- measuring tape at least 3 ft. or 1 m long

### Creating a Scatter Plot

Work with a partner to collect data from at least 10 students.

- For each student, measure and record the distance:
  - From the elbow to the outstretched tip of the middle finger
  - From the knee to the ankle
- Graph your data set, and describe the results.
- Do you think the two body measurements are related? Why or why not?

Elbow-to-fingertip length	Knee-to-ankle length

### Reflect

- Compare graphs with other pairs. Do your graphs appear to show the same relationship between the measurements?
- Identify at least one reason the relationship you identified may not be true in general. What could you do to find out?

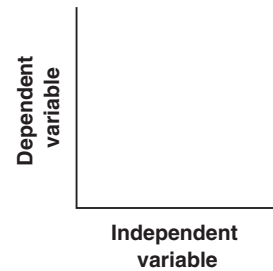
## Connect the Ideas

### Independent variable and dependent variable

Scatter plots represent two-variable data as points. Scatter plots may reveal a relationship between the two variables.

In two-variable situations, one variable may be **dependent** on another: its value changes according to the value of the **independent** variable. For example, the value of a car depends on its age.

Typically, we plot the independent variable on the horizontal axis, and the dependent variable on the vertical axis.



### Example 1

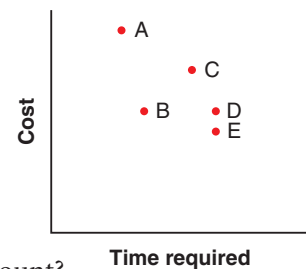
#### Interpreting a Scatter Plot

Jay researched estimates for a job painting his house.

The scatter plot shows Jay's results.

- Which is the dependent variable? Justify your choice.
- Which two companies will take the longest? Which of these is cheaper?
- Which two companies charge the same amount?
- Why might you pick company E? Company B?

Hiring a Painting Company



#### Solution

- Employees are paid for the time they work, so the cost depends on the time required for the job. The dependent variable is cost.
- Company D and Company E take the longest time for the job. Of Companies D and E, Company E is cheaper.
- Company B and Company D charge the same amount.
- You might pick Company E if you are not in a hurry. They take a long time, but are the cheapest.

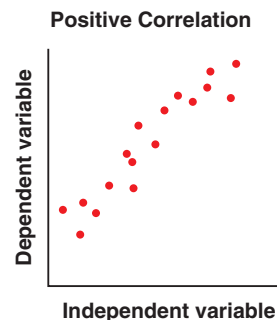
Company B is the second fastest and charges a low price. So, you might pick Company B if you want the job done fairly quickly, but economically.

## Types of correlations

A **correlation** is a relationship between two variables. Graphing two-variable data on a scatter plot may show a correlation between the variables.

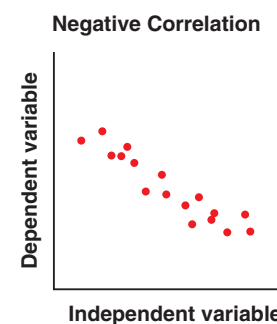
A *positive correlation* describes a situation in which both variables increase together.

On a scatter plot, the points go up to the right.



A *negative correlation* describes a situation in which one variable decreases as the other variable increases.

On a scatter plot, the points go down to the right.



### Example 2

#### Analysing Data Using a Scatter Plot

Davis conducted an experiment comparing a person's leg length and how long it takes the same person to walk 100 m. He gathered these data showing (leg length in centimetres, time taken in seconds). (80, 66), (73, 74), (60, 83), (64, 62), (63, 75), (78, 76), (83, 64), (54, 81), (73, 70), (78, 76)



- Graph the data.
- Does the graph suggest a relationship between leg length and time taken to walk 100 m? If so, describe the relationship.
- Use the scatter plot to estimate the time it would take a person with leg length 85 cm to walk 100 m. Explain.
- How might Davis make the results of his experiment more reliable?



### Solution

a) Time taken may depend on leg length, so use the vertical axis to show time taken, and the horizontal axis to show leg length.

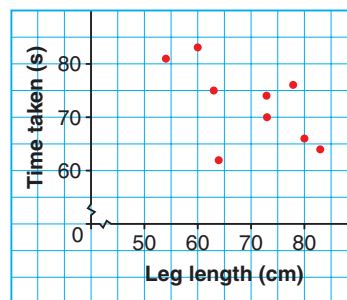
b) The points go down to the right.

This suggests a negative correlation: the time taken to walk 100 m tends to decrease as leg length increases.

c) A point with a first coordinate of 85 might have a second coordinate of about 63. So, someone with legs 85 cm long might take about 63 s to walk 100 m.

d) More data would make Davis' results more reliable. He could recruit people with a wider range of leg lengths while keeping other factors, such as age, gender, and body type, very similar.

Leg Length and Time Taken to Walk 100 m



### Cause-and-effect relationships

Observing a relationship between two variables does not mean that one variable causes a change in the other. Other factors could be involved, or the correlation could be a coincidence.

Demonstrating cause and effect conclusively is a challenging task that requires careful analysis and specialized statistical tools.

### Example 3

### Considering Possible Cause and Effect

State whether the claim in each situation is reasonable.

- A scientific study showed a negative correlation between aerobic exercise and blood pressure. It claimed that the increase in aerobic activity was the cause of the decrease in blood pressure.
- Mila discovered a positive correlation between gasoline price and average monthly temperature. She concluded that temperature determines the price of gasoline.
- Since the 1950s, the concentration of carbon dioxide in the atmosphere has been increasing. Crime rates in many countries have also increased over this time period. Does more carbon dioxide in the atmosphere cause people to commit crimes?

### Solution

- a) It is reasonable to think there may be a cause-and-effect relationship. There are many factors that affect blood pressure, however. Since this is a scientific study, we might reasonably expect that the researchers made efforts to neutralize the other factors, for example by studying subjects in a very close age and fitness range.
- b) It is not reasonable to say there is a cause-and-effect relationship between temperature and gasoline prices. A more likely explanation for the correlation is that higher temperatures occur in the summer, when more people are travelling out of town for weekends and vacations. This increased demand could cause the price increase.
- c) It is not reasonable to say there is a cause-and-effect relationship. It is much more likely that carbon dioxide levels and crime rates are each determined by many other factors, such as increasing populations.

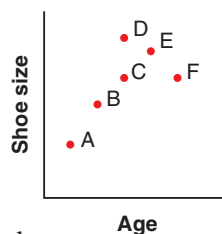
## Practice

**A**

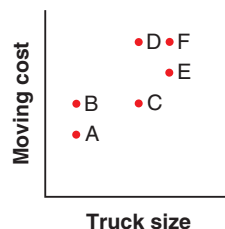
■ For help with questions 1 or 3, see Example 1.

1.
  - a) What does each point show?
  - b) Which child wears the smallest shoe?  
The biggest shoe?
  - c) Which two children are the same age?
  - d) Which two children wear the same shoe size?
2. Identify the dependent variable in the scatter plot in question 1. If you do not think there is a dependent variable, tell why.
3.
  - a) Which company has the lowest moving cost?
  - b) Which two companies use the largest trucks?
  - c) Which two companies use the smallest trucks?
  - d) Based on the scatter plot, how many different truck sizes are there?
  - e) Which company would you use for a small load? For a very large load?
4. Identify the dependent variable in the scatter plot in question 3. If you do not think there is a dependent variable, explain why.

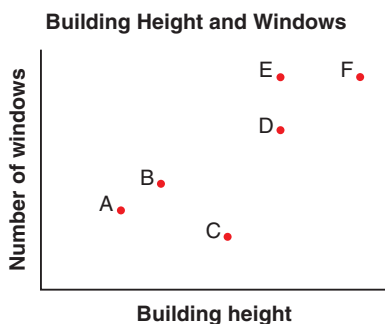
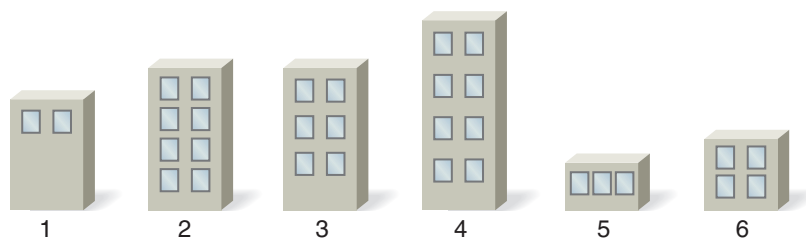
Age and Shoe Size of Six Children



Moving Company Estimates



5. Which building is represented by each point in the scatter plot?



6. For each situation, state whether you think the two variables would have a positive correlation, a negative correlation, or no correlation.

- Cost of a restaurant bill and the amount left as a tip
- Blood pressure reading and IQ
- Number of applicants for a job and probability that you will get the job
- Speed of current and time taken to travel upstream
- Number of kilometres driven and price of gas per litre

**B**

7. Create a scatter plot of the data in each table.

a)

<b>Number of people in household</b>	5	4	5	6	3
<b>Average electricity consumption per month (kWh)</b>	1152	928	953	1067	893

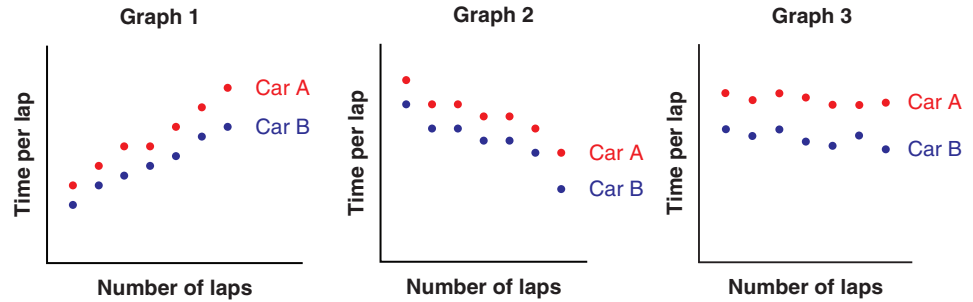
b)

<b>Number of daylight hours</b>	10.54	10.57	10.60	10.63	10.66
<b>Time spent brushing teeth (s)</b>	47	52	38	40	42

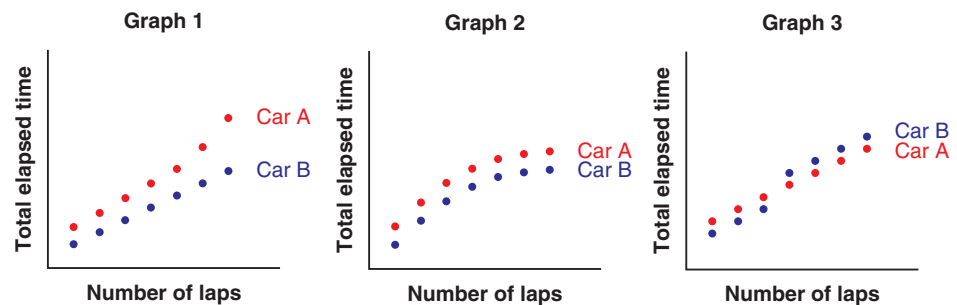
8. For each part, state whether you think the two variables will have any correlation. If you think a correlation exists, describe it briefly.

- Summer temperatures and sales of bottled water
- Price of gasoline and number of people who go to movies
- Price of gasoline and number of day trips people take
- Cost of tuition and number of students who apply to college

9. For each variable, describe a variable that could be correlated with it.
- The amount of time you sleep and ...
  - The size of an in-ground pool and ...
  - The amount of traffic on the road and ...
10. Each graph shows the time taken for two race cars to travel around an oval track for a few practice laps.



- How do these graphs show that there are two cars?
  - How many laps are shown in each graph? What assumption are you making?
  - Which graph shows both cars maintaining a fairly constant lap time?
  - In each graph, which car has the greater average speed?
  - Suppose it starts to rain, making the track slippery and forcing the cars to slow down. Which graph would best show this? Explain.
11. Each graph shows the time taken for two cars to travel around an oval track for the first few laps of a race.



- In Graph 1, which car completes each lap faster?
- Which graph shows both cars increasing their speed during the first few laps?
- Which graph shows Car B making a brief pit stop to repair a loose tire and being passed by Car A?
- Which graph shows Car A having engine trouble and therefore taking longer and longer to complete each lap? Justify your choice.

■ For help with question 12, see Example 2.

**12.** This table shows the hourly cost of heating a pool with a given surface area.

Surface area (sq. ft.)	Hourly cost (cents)
100	24
200	50
400	102
500	127
800	206
1200	310
1500	390

- Using the table, describe what happens to the hourly cost as the surface area increases. What does this suggest about the trends you might see in a scatter plot of these data?
- Create a scatter plot for these data. How does it compare to your prediction in part a?
- Does the scatter plot reveal a correlation between the two variables? If so, describe it.
- About how much would it cost to heat a 20 feet  $\times$  40 feet pool for 24 h?

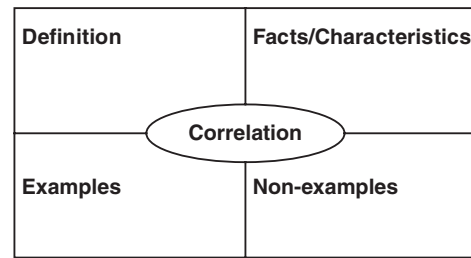
**13. Assessment Focus** In Canada, fuel consumption ratings for vehicles are expressed in L/100 km, or the number of litres of fuel used to travel 100 km.

Fuel Consumption for a Vehicle Driven at Different Speeds on a Test Track								
Speed (km/h)	60	70	80	90	100	110	120	130
Fuel consumption (L/100 km)	5.6	5.9	6.2	6.7	7.4	8.1	8.7	9.4

- Which is the independent variable? Justify your choice.
- What happens to the fuel consumption as the speed increases? Predict the general shape of a scatter plot of the data. Justify your prediction.
- Create a scatter plot of the data. How does the shape of the scatter plot compare to your prediction from part b?
- Describe the relationship between the variables.
- Use the plotted data to estimate the fuel consumption for this car at a speed of 85 km/h. Explain your thinking.



- 14. Literacy in Math** Create a Frayer model about correlations. Include descriptions of positive and negative correlations under *Facts/Characteristics*.



■ For help with question 15, see Example 2.

- 15.** For each situation, decide whether you think it's reasonable to conclude a cause-and-effect relationship.
- Scientific studies found that as exposure to second-hand smoke increased, so did the risk of lung cancer. Based on these studies, a panel concluded that exposure to second-hand smoke increases the risk of lung cancer.
  - Jovanna found a strong correlation between students' science and English marks. She concluded that a student's success in science causes increased marks in English.
  - In Ontario, there seems to be a negative correlation between the consumption of hot chocolate and the number of motorcycle accidents. So, drinking more hot chocolate causes a decrease in the number of motorcycle accidents.
  - Zach researched a negative correlation between the price of a concert ticket and the distance of the seat from the stage. He concluded that the distance from the stage determines the price of a ticket.



**C**

- 16.** Vivek is an amateur meteorologist. He collected data on temperature and relative humidity, each hour for 12 h.

Time	Temperature (°C)	Relative humidity (%)
1:00 p.m.	30	51
2:00 p.m.	31	48
3:00 p.m.	32	46
4:00 p.m.	31	46
5:00 p.m.	28	62
6:00 p.m.	30	49
7:00 p.m.	30	47
8:00 p.m.	29	52
9:00 p.m.	27	58
10:00 p.m.	24	69
11:00 p.m.	23	73
12:00 a.m.	21	82

- a) Which is the independent variable? Why do you think so?
- b) Create a scatter plot of Vivek's data.
- c) Does there appear to be a correlation? If so, describe it. Otherwise, explain why you think there is no trend.
- 17.** Use the scatter plot you drew in question 16.
- a) Vivek heard that the temperature will reach a low of 18°C overnight. Estimate the relative humidity overnight.
- b) Vivek has found it is very likely to rain when the relative humidity reaches 90%. Is it very likely to rain overnight? Explain.

### *In Your Own Words*

Describe a situation where a set of data can be displayed in a scatter plot. How could the scatter plot help you interpret the data?

# 3.3

## Line of Best Fit

*In recent years, natural disasters seem to be more intense and occur more frequently. Will this trend continue? Climatologists try to answer questions like this by identifying trends in historical data and using them to make predictions about future events.*



### Investigate

#### Materials

- grid paper
- coloured pencils

### Sketching a Line of Best Fit

Work with a partner.

Sabah researches the cost of houses in a new area. She is looking for

- A two-storey, detached house
  - Asking prices of \$300 000 or less
  - 2000 square feet of living space
- Graph the data. Use axes that show house sizes up to 2500 square feet and prices of up to \$400 000.
  - Describe any trend you see in the graph. With a coloured pencil, draw the line that you think comes closest to matching the trend.
  - Use the line to estimate how much a 2000 square foot house might cost in Sabah's area.

Size (sq. ft.)	Price (\$)
1700	271 900
1850	289 900
1600	277 900
1650	289 900
1700	279 000
1800	294 900
1550	269 900

You've drawn the **line of best fit**.



Sabah notices a new *For Sale* sign on a house in the area. At 2300 square feet and \$384 500, it does not match her criteria. However, she decides to add the data to her collection anyway.

- Add the point (2300, 384 500) to your scatter plot.
- With a different colour, draw the new line of best fit.
- Use this line to estimate the price for a 2000 square foot house in Sabah's neighbourhood. Compare this estimate to your first estimate.

### Reflect

- How does the arrangement of plotted data affect the way you draw a line of best fit? How does it affect your confidence in using the line to make predictions?
- A real estate agent tells Sabah that 2000 square foot houses often come up for sale in this area, usually at prices from \$295 000 to \$305 000. How close were your estimates to these prices? Which line of best fit helped you make a closer prediction?

## Connect the Ideas

### Line of best fit

A **line of best fit** is a line drawn through data points to best represent a linear relationship between two variables. Other names for a line of best fit are *regression line* or *trend line*.

Drawing the line of best fit involves more than just finding a pathway through the middle of the data. The line of best fit is the line that is closest to each point. The more varied the position of points, the more difficult it is to draw the line of best fit.

### Outliers

In a scatter plot, a point that lies far away from the main cluster of points is an **outlier**. An outlier may be caused by inaccurate measurements, or it may be an unusual, but still valid, result.

### Outliers and the line of best fit

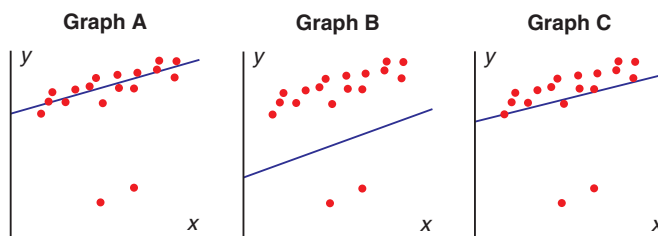
The line of best fit should reflect all valid points from a data set. This includes outliers.

If most of the plotted points are clustered along a linear path, even a single outlier can affect the path of the line of best fit.

### Example 1

### Exploring the Effect of Outliers on a Line of Best Fit

Which line is the line of best fit? Justify your choice.



#### Solution

The line in Graph C seems most likely to be the line of best fit.

- The line in Graph A passes through the middle of the main cluster of data. It suggests the two outliers are not present or are invalid.
- The line in Graph B follows a path exactly halfway between the main cluster of data and the outliers. It suggests that the outliers and the main cluster of data are equally important.
- The line in Graph C passes just below the middle of the main cluster of data. Its path is affected by the outliers, but it is affected more by the main cluster of data. It is the most reasonable choice.

### Interpolating and extrapolating from a line of best fit

A line of best fit can be used to estimate or predict values.

- Estimating values that lie among the known values on the graph is **interpolation**.
- Predicting values that lie beyond the known values is **extrapolation**.

### Example 2

### Using a Line of Best Fit to Make Predictions

These are the pre-exam term marks and exam marks for some students in a Grade 12 math course.

Term mark (%)	84	76	70	95	92	61	25	55	51	73	62
Exam mark (%)	80	72	68	96	90	58	29	60	53	77	67

- Graph the data and draw the line of best fit.
- Determine the equation of the line of best fit.
- Use the data to predict the exam mark of a student with a pre-exam term mark of 98%.
- Use the data to predict the exam mark of a student with a pre-exam term mark of 10%.

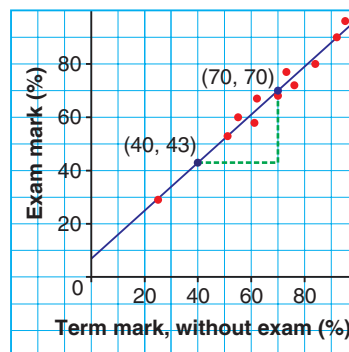
## Solution

- a) Plot the points. Place a ruler on the graph so that it comes as close as possible to all of the points. Draw a line along the ruler's edge.
- b) The equation of a line is  $y = mx + b$ , where  $m$  is the slope and  $b$  is the vertical intercept. Choose two points on the line to find the slope:

(40, 43) and (70, 70)

$$\begin{aligned}\text{Slope: } \frac{\text{rise}}{\text{run}} &= \frac{70 - 43}{70 - 40} \\ &= \frac{27}{30}, \text{ or } 0.9\end{aligned}$$

Term Marks and Exam Marks



The point where a line crosses the vertical axis is called the *vertical intercept*.

$x$  and  $y$  represent the marks as percents, so do not convert them to decimals.

From the graph, the vertical intercept of the line is about 7.

The equation of the line of best fit is approximately  $y = 0.9x + 7$ .

- c) Extrapolating from the graph, a student with a pre-exam term mark of 98% would get an exam mark of about 95%.
- d) Substitute  $x = 10$  into the equation of the line of best fit.

$$\begin{aligned}y &= 0.9x + 7 \\ &= 0.9 \times 10 + 7 \\ &= 16\end{aligned}$$

The predicted exam mark is 16%.

How confident can we be of predictions made from scatter plots?

## Data spread and reliability

A model with data spread over a small interval is less reliable than a model based on data spread over a larger interval.

The farther we get from the main cluster of data, the less confidence we should have on predictions made from that model.

How does this relate to the predictions in Example 2?

## Sample size and reliability

The more data we use, the more reliable the prediction should be.

## Non-linear data

Not all relationships between variables are linear. Over a small interval, a linear model may provide a reasonable fit for non-linear data, but it will not be reliable at the extremes.

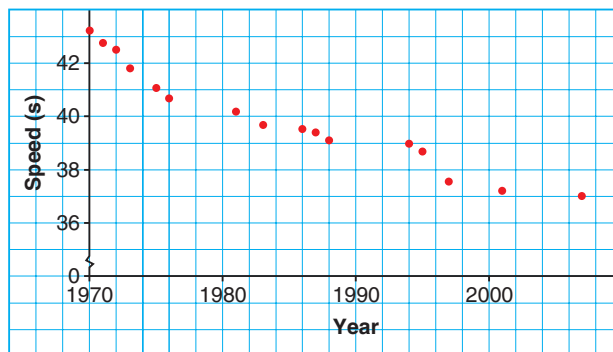
You will learn more about other types of models in Chapter 5.

### Example 3

### Deciding if a Correlation Is Linear

Reanna created a scatter plot using data about the world record times for women's 500-m speed skating. If the world record changed more than once in a year, she used the best time for that year.

World Record for Women's 500-m Speed Skating



Describe the relationship between the two variables. Does it appear to be linear? If not, describe how Reanna could better model the data.

#### **Solution**

As the years increase, the time decreases; there is a strong negative correlation between the variables.

Although the relationship appears to be linear in some places, the overall relationship is non-linear. The world record times decreased most rapidly in the four-year period from 1970 to 1976. A model that decreases steeply at first and then more slowly would better represent the data.

### Assessing a linear model

A linear model may be unreliable in these situations:

- The model is based on too few pieces of data.
- The model is based on data that are clustered together.
- There does not appear to be any correlation.
- There are outliers.
- The data do not appear to be linear.

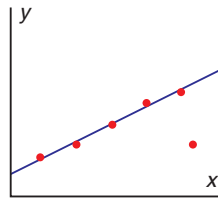
# Practice

**A**

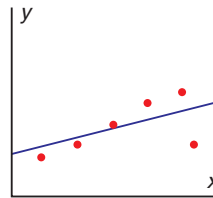
■ For help with questions 1 to 4, see Example 1.

1. For each scatter plot, select the line of best fit. Justify each choice.

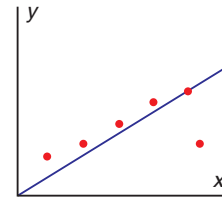
a) **Graph A**



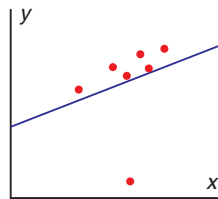
**Graph B**



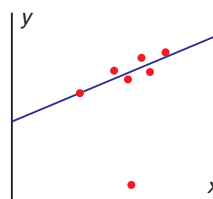
**Graph C**



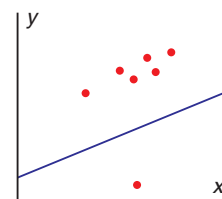
b) **Graph D**



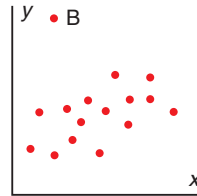
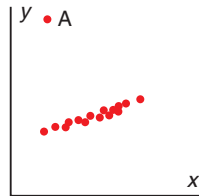
**Graph E**



**Graph F**



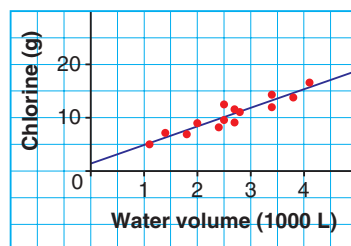
2. Points A and B are both outliers. Which outlier would have the greater effect on the path of the line of best fit? Justify your choice.



3. Use this graph to identify whether you would use interpolation or extrapolation to predict each value.

- Volume of water in a hot tub that requires 10 g chlorine.
- Mass of chlorine needed for a 500 L hot tub.
- Mass of chlorine needed for a 35 000 L hot tub.
- Volume of water in a hot tub that requires 18 g chlorine.

**Hot Tub Maintenance**





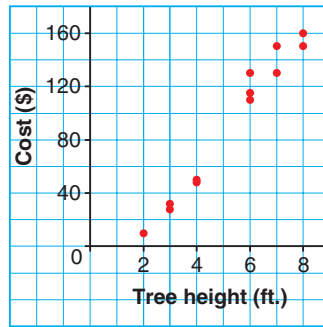
**B**

For help with question 4, see Example 2.

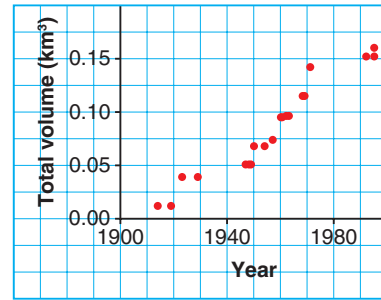
Use the *Course Study Guide* at the end of the book to recall how to calculate slope.

4. a) On a copy of these graphs, sketch the line of best fit for each scatter plot.  
 b) Determine the equation of each line of best fit.

i) Retail Cost of Evergreens

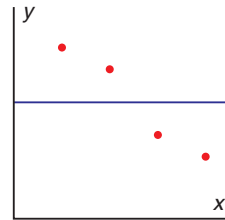


ii) Eruptions of Cerro Negro Volcano since 1900

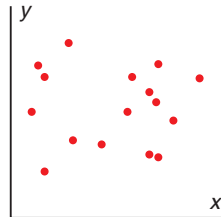


5. The line drawn in this graph passes through the points so that half the points lie above the line and half the points lie below the line.

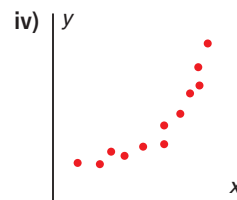
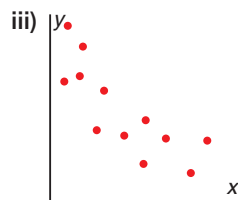
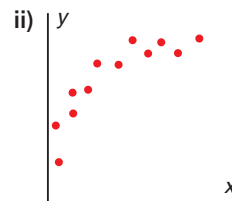
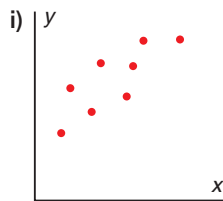
- a) Why is this line not the line of best fit?  
 b) Describe how the line of best fit would look for these data.



6. Describe the problems with drawing the line of best fit for these data.



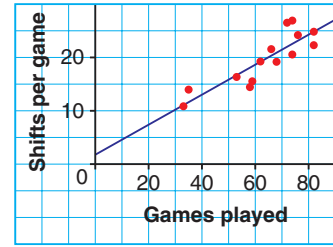
7. a) For each scatter plot, describe the relationship between  $x$  and  $y$ .  
 b) Would you model each relationship with a linear or non-linear model? Justify your answers.



■ For help with question 8, see Example 3.

8. a) Use the line of best fit for this scatter plot to make each prediction.
- Number of games played by a player who averages 10 shifts per game
  - Number of games played by a player who averages 5 shifts per game
- b) Which of your predictions in part a do you think is more reliable?

Play Time for Some Toronto Maple Leaf Forwards



9. A power utility company warns that if the demand for electricity is greater than 27 000 MW, there will be some power outage. The table shows daily high temperature and peak electricity demand for a 10-day period one summer.
- Create a graph and draw the line of best fit for these data.
  - Determine an equation of the line of best fit.
  - Estimate the peak electricity demand for a daily high temperature of  $28^{\circ}\text{C}$ .
  - Predict the likelihood of a power outage when the daily high temperature is  $37^{\circ}\text{C}$ .
  - The power company can take one of their generators offline for maintenance when the demand is below 17 000 MW. On what days would the peak electricity demand fall below 17 000 MW?

Daily high temperature ( $^{\circ}\text{C}$ )	Peak electricity demand (MW)
24	19 503
23	18 832
24	19 150
27	20 613
29	21 544
30	22 237
26	20 082
29	21 819
32	23 488
34	24 950



- 10. Assessment Focus** Use the table of data about life expectancy of Canadian males.

Life Expectancy at Birth of a Canadian Male								
Birth year	1920	1930	1940	1950	1960	1970	1980	1990
Life expectancy at birth (years)	59	60	63	66	68	69	72	75

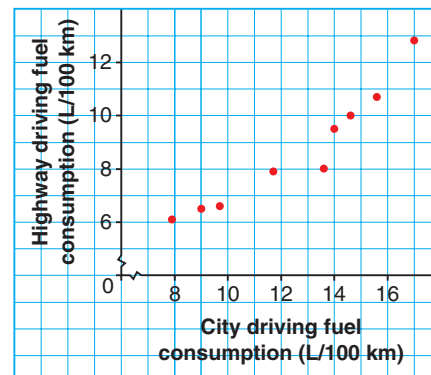
- Create a graph and draw the line of best fit for these data.
  - Find an equation of the line of best fit.
  - Estimate the life expectancy of a male born in 1975.
  - Predict the life expectancy of a male born in 2000.
  - Predict the birth year of males with a life expectancy of 80 years.
  - Write a question someone could answer using your graph. Prepare an answer for the question.
- 11.** Use the table of data about life expectancy of Canadian females.

Life Expectancy at Birth of a Canadian Female								
Birth year	1920	1930	1940	1950	1960	1970	1980	1990
Life expectancy at birth (years)	61	62	66	71	74	76	79	81

- How do the life expectancies for male and female Canadians compare?
  - Repeat question 10 using these data.
- 12. Literacy in Math** Another name for a line of best fit is a *trend line*. Explain why this is an appropriate name.

- 13.** How would you describe the relationship between fuel consumption for city and highway driving? Does it appear to be linear? If not, describe a better model for the data.

Fuel Consumption of Vehicle Models in 2007

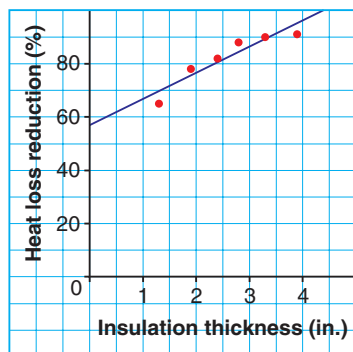




- 14.** For each situation, explain why the data analysis is not reasonable. There may be more than one reason for each graph.

- a) A technician graphed data and concluded he would get a 100% reduction in heat loss by using  $4\frac{1}{2}$  inches of insulation.

**Heat Loss Reduction with Insulation**



- b) A carnival weight-guesser created this scatter plot using data from his last eight customers. He plans to ask customers how tall they are and use the linear model to estimate their weights.

**Carnival Weight-Guesser**



**C**

- 15.** You have seen that outliers affect the placement of the line of best fit.
- a) Do you think that the degree to which an outlier influences the line of best fit depends on the size of the data sample? Explain.
- b) What other factors can you think of that might affect the degree to which an outlier affects the line of best fit?

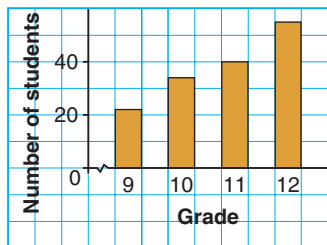
### *In Your Own Words*

Suppose a classmate missed the lesson on line of best fit. Write instructions for her or him about how to draw a line of best fit and why it is useful.

## Mid-Chapter Review

- 3.1** 1. Does the situation illustrate one-variable or two-variable data? Explain.

a) Students Competing in a Math Contest

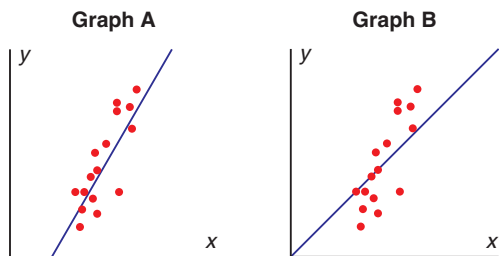


b) Students in a science class measured the height of their plants each week for 12 weeks.

- 3.2** 2. State whether you think the variables in each situation would have a negative correlation, a positive correlation, or no correlation.

- Driving speed and time to travel 100 km
- Size of a house and its interior temperature
- One's age and the number of colds one's had
- Cost of gasoline and fuel efficiency of a vehicle

- 3.3** 3. Which linear model better represents the data? Explain your choice.



- 3.2** 4. This table compares the parking facilities of several large companies.

**3.3**

Available Land and Parking Capacity for Various Companies	
Acres of land	Parking spaces
2.0	145
1.5	160
4.0	500
1.0	95
5.0	600
4.0	425
2.0	550
3.0	280

- Create a scatter plot of the data.
  - Describe any trends you see.
5. Use the table of data on tire pressure.
- Graph the data; draw a line of best fit.
  - Describe the correlation.
  - Predict the pressure at:
    - 70°F
    - 40°F

Tire Pressure and Temperature	
Outside temperature (°F)	Tire pressure (psi)
58	35
79	38
63	36
61	36
85	39
55	34
74	37
88	40

# 3.4

## Analysing Data Using a Graphing Calculator

*Smog is created when air pollutants react with sunshine and heat. The Ontario Ministry of the Environment measures the amount of smog in the air and issues a smog advisory when high smog levels are likely.*



### Inquire

### Analysing Scatter Plots with the TI-83/TI-84

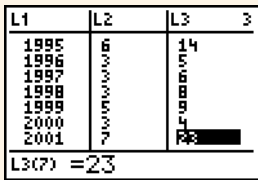
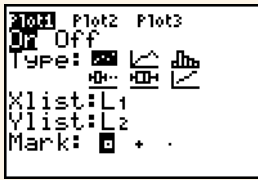
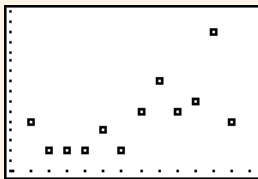
#### Materials

- TI-83 or TI-84 graphing calculator

This table shows the number of smog advisories issued by the Ontario Ministry of the Environment from 1995 to 2006.

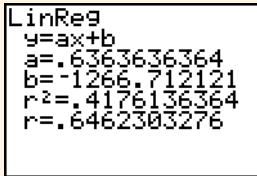
Number of Smog Advisories and Smog Days in Ontario												
Year	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
Number of advisories	6	3	3	3	5	3	7	10	7	8	15	6
Total number of days	14	5	6	8	9	4	23	27	19	20	53	17

## ■ Drawing a scatter plot

Steps	Display	Notes
<p>Press <b>[STAT]</b> 1 to access the Stat List Editor. To clear any data from <b>L1</b>, use the arrow keys to move the cursor to the <b>L1</b> heading. Press <b>[CLEAR]</b> <b>[ENTER]</b>. Repeat for <b>L2</b> and <b>L3</b>.</p> <p>Enter the data from the smog advisory table.</p>		<p>Enter the years in <b>L1</b>, the number of advisories in <b>L2</b>, and the number of days in <b>L3</b>. You can press either <b>[ENTER]</b> or <b>[↓]</b> after each number.</p>
<p>Press <b>[2nd]</b> <b>[Y=]</b> 1 to access the Stat Plots menu for Plot 1. With the cursor over <b>ON</b>, press <b>[ENTER]</b>. Press <b>[↓]</b> <b>[ENTER]</b> to select <b>1: (scatter plot)</b>. Press <b>[↓]</b> <b>[2nd]</b> 1 <b>[ENTER]</b> to set <b>Xlist</b> to <b>L1</b>, <b>[2nd]</b> 2 <b>[ENTER]</b> to set <b>Ylist</b> to <b>L2</b>, and <b>[ENTER]</b> to set <b>Mark</b> to <b>[□]</b> (box).</p>		<p>Make sure <b>Plot2</b> and <b>Plot3</b> are turned off. Use the arrow keys to place the cursor over <b>Plot2</b>. Press <b>[ENTER]</b> <b>[↓]</b> <b>[ENTER]</b>. Repeat for <b>Plot3</b>.</p>
<p>Press <b>[Y=]</b> to see the list of equations stored in the calculator. Clear them or turn them off. Then press <b>[ZOOM]</b> 9 to graph the data from <b>L1</b> and <b>L2</b> on an appropriate scale.</p>		<p>When an equation is turned on, the equals sign is highlighted. To turn off an equation, place the cursor over <b>=</b> and press <b>[ENTER]</b>.</p>

1. a) What does the scatter plot suggest about how the number of smog advisories is changing over time?
- b) Describe any correlation between the variables.
- c) Do you think it would be appropriate to model the relationship with a line of best fit? Justify your answer.

## ■ Creating a line of best fit

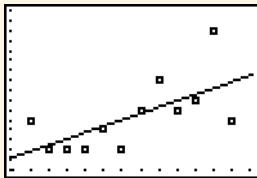
Steps	Display	Notes
<p>Press <b>[STAT]</b> <b>[▶]</b> 4 to select linear regression. Then press <b>[2nd]</b> 1 <b>[,]</b> <b>[2nd]</b> 2 <b>[,]</b> to have the regression performed on the data in lists <b>L1</b> and <b>L2</b>. Press <b>[VARS]</b> <b>[▶]</b> 1 1 to have the equation of the line of best fit stored in <b>Y1</b>. Press <b>[ENTER]</b> to perform the regression.</p>		<p>If the values of <math>r^2</math> and <math>r</math> do not appear on your screen, turn diagnostic mode on. Press <b>[2nd]</b> 0, scroll down to <b>DiagnosticOn</b> and press <b>[ENTER]</b> <b>[ENTER]</b>.</p>

### Correlation coefficient

The correlation coefficient,  $r$ , is a number between  $-1$  and  $1$ . It describes the strength of the linear correlation.

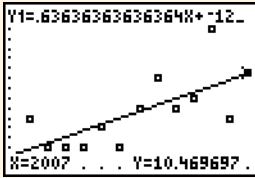
Generally, the closer  $r$  is to  $1$  or  $-1$ , the stronger the linear correlation and the more closely the data approximate a line.

2. a) The screen shows that the equation of the line is  $y = ax + b$  and gives the values of  $a$  and  $b$  to many decimal places. What are these values? Round  $a$  to one decimal place and  $b$  to the nearest whole number. What does  $a$  tell you about the line of best fit?
- b) What is the correlation coefficient,  $r$ ?

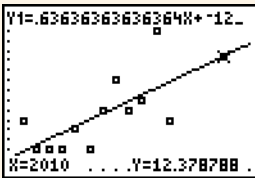
Steps	Display	Notes
<p>Press <b>[GRAPH]</b> to display the scatter plot and line of best fit.</p>		<p>To view the equation of the line of best fit, press <b>[Y=]</b>. Recall that the equation of the line of best fit is stored in <b>Y1</b>.</p>

3. Describe the line of best fit. Is it what you expected? Explain why or why not.

■ Extrapolating with the line of best fit

Steps	Display	Notes
<p>Use the TRACE feature to determine the value of <math>Y</math> when <math>X</math> is 2007. Press <b>TRACE</b>. Then press <b>↓</b> to place the cursor on the line of best fit.</p> <p>Press: 2007 <b>ENTER</b></p> <p>The corresponding <math>Y</math>-value is displayed at the bottom right of the screen.</p>		<p>The TRACE feature will not work for <math>X</math>-values that are beyond the window settings. You could extrapolate algebraically by substituting 2007 for <math>x</math> in the equation of the line of best fit and determining the value of <math>y</math>.</p> <p>Press: <b>CLEAR</b> <b>VARS</b> <b>▶</b> 1 1 <b>(</b> 2007 <b>)</b> <b>ENTER</b></p>

4. a) What was the predicted number of smog advisories for 2007?  
 b) As of October 22, 2007, there were 13 smog advisories in Ontario. Is the total number for the year likely to be much greater than this? Explain.  
 c) Was the prediction from the line of best fit close to the actual number?
  
5. Follow these steps to predict the number of smog advisory days in 2010.

Steps	Display	Notes
<p>Press <b>WINDOW</b> to access the <b>WINDOW</b> menu. The cursor should be on the <b>Xmin</b> setting. Press <b>↓</b> and input 2012 for the <b>Xmax</b> value. Press <b>TRACE</b>. Then press <b>↓</b> to place the cursor on the line of best fit.</p> <p>Press: 2010 <b>ENTER</b>. What is the corresponding <math>Y</math>-value?</p>		<p>You could extrapolate algebraically by substituting 2010 for <math>x</math> in the equation of the line of best fit and determining the value of <math>y</math>.</p> <p>Press: <b>CLEAR</b> <b>VARS</b> <b>▶</b> 1 1 <b>(</b> 2010 <b>)</b> <b>ENTER</b></p>

6. a) Which point appears to be an outlier?
- b) The summer of 2005 was one of the hottest and most humid summers in Ontario. Toronto recorded 41 days with temperatures greater than 30°C. How might this have resulted in the unusual data for 2005?

■ **Removing an outlier**

Steps	Display	Notes
<p>Press <b>[STAT]</b> 1 to access the lists you have stored.</p> <p>In <b>L1</b>, use the <b>[↓]</b> key to move down to the entry for 2005.</p> <p>Then press <b>[DEL]</b> <b>[▶]</b> to delete it and move to <b>L2</b>.</p> <p>Press <b>[DEL]</b> <b>[▶]</b> <b>[DEL]</b> to delete the corresponding entries in <b>L2</b> and <b>L3</b>.</p>		<p>To re-enter 2005 in list <b>L1</b>, place the cursor over 2006 and press <b>[2nd]</b> <b>[DEL]</b> 2005 <b>[ENTER]</b>. Then press <b>[↑]</b> <b>[▶]</b> to move to list <b>L2</b>. Repeat the process to enter the data for lists <b>L2</b> and <b>L3</b>.</p>

7. Repeat the process from *Creating a line of best fit*.
  - a) What are these values of  $a$  and  $b$ ?  
Round  $a$  to one decimal place and  $b$  to the nearest whole number. What does  $a$  tell you about the line of best fit?
  - b) What is the correlation coefficient,  $r$ ?
  - c) How does the correlation coefficient compare to the coefficient from question 2? What does this suggest about the new model?
8. Repeat the process from *Extrapolating with the line of best fit*.
  - a) What are the new predicted number of smog advisories for 2007 and 2010?
  - b) Which line of best fit gave a more accurate prediction for 2007?
  - c) Would you say that including outliers always weakens a model? Justify your answer.



9. Suppose an environmental group wants to show Ontarians that air quality in Ontario is getting worse. Use the data about the number of days with a smog advisory you stored in L3. Be sure to re-enter the data for 2005.

Analyse the data.

- Create a scatter plot relating the year and the number of days with a smog advisory.
  - Describe the relationship as it appears on your scatter plot.
  - Record the  $r$ -value for the correlation and compare it to the previous correlation coefficients you recorded in questions 2 and 7.
  - Create a line of best fit and describe whether it provides a good model for the data.
  - Write a conclusion stating whether the graph supports the environmental group's case.
10. The Ministry of the Environment says air quality in the province has been improving since 1988, but it expects the number of smog advisories to increase because it is doing a better job of monitoring air quality. What does this show about drawing conclusions about cause and effect based only on numerical data?

### Reflect

- Suppose you used linear models to represent two different sets of data. Describe how you could decide which set of data is better represented by a linear model.
- What are advantages of using a graphing calculator to analyse data?



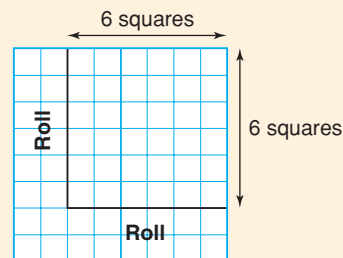
# GAME

## Give It to Me Straight

### Materials

- TI-83 or TI-84 graphing calculator
- dice
- grid paper

Play the game in pairs. Each player will need a calculator.



- On grid paper, draw and label axes as shown.

On the graphing calculator, press **[STAT]** **[ENTER]** to access the Stat List Editor. Clear any data from L1 and L2.

- One player rolls the dice.
- The other player uses the numbers that are rolled as the coordinates of a point. For example, if 3 and 5 are rolled, the player can create (3, 5) or (5, 3). The player enters the coordinates in lists L1 and L2 of the graphing calculator and plots the point on the grid.
- Players take turns rolling the dice and choosing coordinates until each player has created 6 points. The goal is to get a higher correlation coefficient.
- Each player predicts the correlation coefficient for her or his data, and records the prediction as a decimal to the nearest hundredth.
- Players use the graphing calculators to determine the  $r$ -value for their data.  
Press **[STAT]** **[▶]** **4** **[ENTER]** to perform linear regression.
- The player whose data have an  $r$ -value closer to 1 or  $-1$  gets 1 point.
- The player whose prediction of the  $r$ -value is closer to the actual value also gets 1 point.
- After five rounds, the player with more points is the winner.  
If there is a tie, play an extra round. The player with the higher  $r$ -value is the winner.

If the values of  $r^2$  and  $r$  are not showing on the screen, the diagnostic mode is turned off. To turn it on, press **[2nd]** **0**, scroll down the list to **DiagnosticOn** and press **[ENTER]** **[ENTER]**.

### Reflect

- Explain how you decided which number to use as the  $x$ -value and which to use as the  $y$ -value in a coordinate pair.
- What strategy did you use to predict the correlation coefficient?

## 3.5

## Analysing Data Using a Spreadsheet

Do you like to sleep in? Maybe you should move to Resolute Bay, Nunavut where the sun doesn't rise for three months. But you will have to leave by April, when the town begins a three-month stretch of daylight!



## Inquire

## Analysing Scatter Plots Using Software

## Materials

- Microsoft Excel
- daylighthours.xls
- snowrain.xls
- access to the Internet and E-STAT



Work with a partner.

## Part A: Using Microsoft Excel

- Open the file *daylighthours.xls*.

1. The spreadsheet shows the latitudes of different locations and the number of daylight hours on August 15. Describe the relationship between latitude and hours of daylight.

	A	B
1	Daylight Hours on August 15	
2	Latitude	Length of daylight time
3	(°N)	(h)
4	0	12.1
5	10	12.5
6	20	12.8
7	30	13.3
8	40	13.8
9	50	14.5
10	60	15.8
11	70	18.5
12	80	24.0
	90	24.0

- Highlight cells A3 to B12.

Click the **Chart Wizard** icon. 

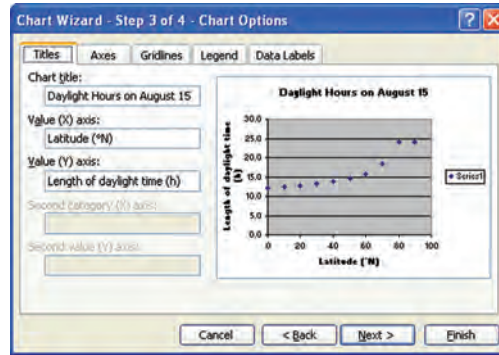
Under *Chart type*: select **XY (Scatter)**.

Click **Next**. Accept the data range by clicking **Next** again.

The title screen should appear. If it does not, click on the **Titles** tab.

Enter the data headings as the chart titles.

You may find it easier to type "degrees North" than "°N".



The latitude of the equator is 0°N.  
The latitude of the North Pole is 90°N.

There are 24 h in 1 day.

- Click on the **Legend** tab. Deselect **Show Legend**.

Click **Finish** to embed the graph into your spreadsheet.

- Right click the horizontal axis of the graph and select **Format Axis....**

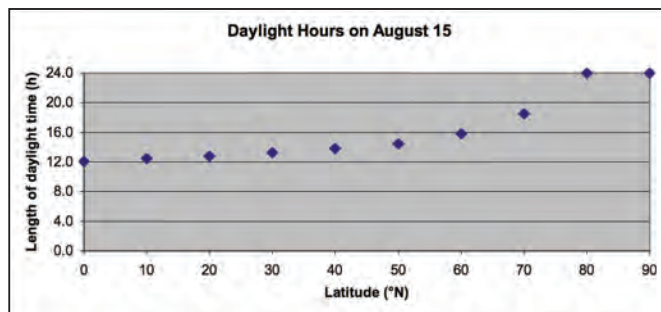
Click on the **Scale** tab.

Set **Minimum** to 0 and **Maximum** to 90.

- Right click the vertical axis and select **Format Axis....**

Click on the **Scale** tab.

Set **Minimum** to 0 and **Maximum** to 24.

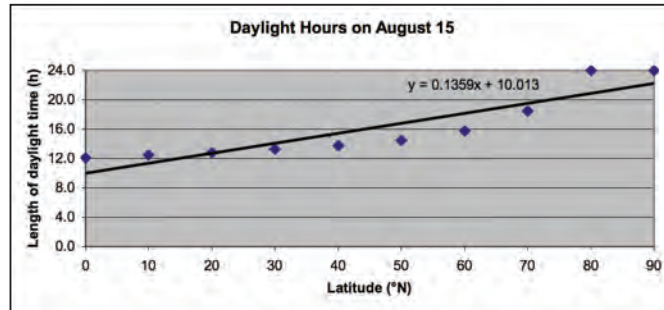


**2.** Describe the correlation between the variables. Compare it with your description of the relationship from question 1.

**3.** Do you think a linear model would represent the data well? Explain your thinking.

The **Chart** menu will not appear unless a graph is selected.

- From the **Chart** menu, select **Add Trendline**.  
Select **Linear** from the list of Trend/regression types.  
Click on the **Options** tab, then select **Display equation on chart**.  
Click **OK**. Your graph should look similar to the one shown here.



4. a) What is the equation of the line of best fit?  
b) Do you think the line does a good job of representing these data? Justify your answer.  
c) Would the linear model provide reliable estimates of daylight hours? Justify your answer.



5. Environment Canada calculates “weather normals” that represent typical weather data for different locations. The current normals are based on data collected from 1971 to 2000.

	A	B	C
1	<b>City or town</b>	<b>Average annual snowfall (cm)</b>	<b>Average annual rainfall (mm)</b>
2	Belleville	155.7	735.9
3	Cameron Falls	237.5	576.6
4	Chalk River	195.4	669.2
5	Cobourg	106.0	765.8
6	Dresden	84.6	759.5
7	Hamilton	161.8	764.8
8	Kapuskasing	313.0	544.6
9	Renfrew	195.5	616.0
10	Samia	125.0	732.6
11	Sault Ste. Marie	302.9	634.3
12	Timmins	313.4	558.1
13	Toronto	133.1	709.8

Open the file *snowrain.xls*.

- a) Do the two variables appear to be related? If so, describe the relationship. If not, explain why not.
- b) Create a scatter plot for the data. Describe any correlation you see. Does the graph support your answer to part a?
- c) Add a line of best fit to the graph. How well do you think it represents the data? Justify your answer.
- d) Petawawa receives an average of 228.5 cm of snow each year. Based on the line of best fit, what would you expect the average annual rainfall to be in Petawawa? How close was the prediction to the actual average of 615.9 mm?

## Part B: Using E-STAT

- Go to the Statistics Canada Web site.

Click **English**.

Select **Learning resources** from the menu on the left.

Click on **E-STAT** in the yellow box on the right.

Click on **Accept and enter**.

If you are working from home, you will need to enter the user name and password assigned to your school.

*CANSIM* provides data taken over time.  
*Census databases* offer information about entire populations taken once every 5 years.

The screenshot shows the Statistics Canada E-STAT website. At the top, there is a navigation bar with links for Français, Contact us, Help, Search, and Canada site. Below this is a secondary navigation bar with Site map, About us, Privacy, Accessibility, and My account. The main header features the Statistics Canada logo and the text 'CANADA'S NATIONAL STATISTICAL AGENCY'. The central content area is titled 'E-STAT: Table of contents' and lists various categories with links to their respective data tables. The categories include Economy, Land and Resources, People, Nation, Historical Censuses of Canada, and Elections Canada. A sidebar on the left contains links for HOME, E-STAT, About E-STAT, What's new in E-STAT, Table of contents, User guides, Search CANSIM, Search Censuses, Search map 2001, Help/Frequently asked questions, Contact E-STAT, and Learning resources.

- The E-STAT table of contents will be displayed.

In the *Land and Resources* section, click on **Agriculture**.

From the list of CANSIM data, click on **Food and nutrition**.

Click on table **002-0011**. This table contains data about the per capita consumption of various food and beverage items throughout Canada.

- Under *Food Categories*, select **Food available**.

Under *Commodity* you will be selecting two items, but there are too many choices to see all at once. Click on **View checklist and footnotes** to display all the items more conveniently.

Scroll down and select **Ice cream (litres per year)** and **Yogurt (litres per year)**.

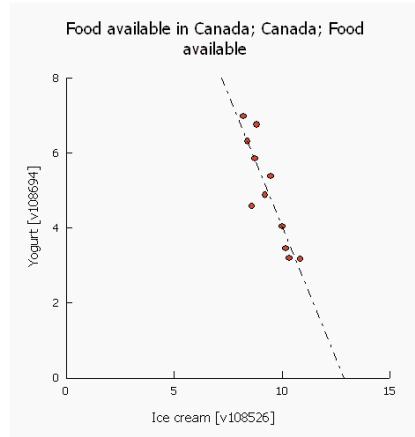
Scroll up to the top of the page and click **Return to picklist**.

Set the Reference period from **1996 to 2006**.

Click on **Retrieve as individual Time Series**.

*Per capita* means per person.

- Create a scatter plot. Select **Scatter graph with line of best fit (linear regression)** and click **Retrieve Now**. Click on **Modify Graphic**. After **Origin**: select **Start axis at 0**. Click **Replot**.



- a) What two variables are being compared in this graph?  
 b) What data are not included in the graph?  
 c) Describe the correlation displayed by the scatter plot.  
 d) What does the direction of the line of best fit tell us about the relationship between ice cream consumption and yogurt consumption? Why do you think this relationship occurs?

- Click on the back button three times to return to the *Output specification screen*.

In the *Screen output – table* box, under *HTML, Table*, select **Time as rows** and click **Retrieve now**.

Click and drag to highlight entire table, including the headings.

With the table highlighted and the cursor on the highlighted table, right click and **Copy**.

- Open a new *Microsoft Excel* spreadsheet.

Right click on cell A1 and select **Paste**.

The data from E-STAT will appear.

Widen the columns to reveal the table headings.

	A	B	C
1	<b>Annual</b>	<b>v108526 - Canada; Food available; Ice cream (litres per year)</b>	<b>v108694 - Canada; Food available; Yogurt (litres per year)</b>
2	1996	10.87	3.17
3	1997	10.35	3.19
4	1998	10.18	3.46
5	1999	10.02	4.05
6	2000	8.62	4.59
7	2001	9.22	4.88
8	2002	9.49	5.39
9	2003	8.76	5.85
10	2004	8.4	6.31
11	2005	8.84	6.76
12	2006	8.22	6.98

- Highlight the columns containing the ice cream and yogurt data. Click on the **Chart Wizard** icon in the toolbar. Follow the steps you learned in the first part of this *Inquire* to create a scatter plot. Enter appropriate titles and embed the graph into your spreadsheet. Add a trend line using the **Chart** menu.
- 7. Compare this scatter plot to the one you created in E-STAT. Which one do you prefer? Justify your answer.



- 8. Return to E-STAT table 002-0011 on food consumption in Canada.
  - a) Retrieve data about the consumption, in litres, of standard (3.25%) milk and partly skimmed (1%) milk between 1996 and 2006.
  - b) Would you expect the consumption of these two products to be related? Justify your answer.
  - c) Graph the data as a scatter plot with a line of best fit. Describe the shape and direction of the correlation. What does the scatter plot tell you about the consumption of these two kinds of milk?

### Reflect

- How does creating a scatter plot help you identify and describe trends in data that might not be obvious from looking at a table?
- Choose a data set from Lesson 3.2 or 3.3. Describe how to use a spreadsheet to graph the data set. What would be an advantage of using spreadsheet software to construct the scatter plot and draw the line of best fit?

## 3.6

Analysing Data Using *Fathom*

In Ontario, every driver is required to have liability insurance on his or her vehicle. Liability insurance covers the potentially enormous costs if a person's actions cause damage to property or injury to a third party.



## Inquire

Creating and Analysing Scatter Plots Using *Fathom*

## Materials

- *Fathom*
- access to the Internet and E-STAT

Work with a partner.

## Part A: Working with Given Data

Lena is researching the cost of the liability portion of car insurance to determine whether it is related to the car's value.

She has collected this data from various insurance brokers.

Value of car (\$)	Liability insurance cost (\$/year)	Collision insurance cost (\$/year)
5 000	247	138
10 000	233	163
10 000	275	154
15 000	291	192
15 000	277	185
20 000	243	257
20 000	315	201
25 000	254	233
30 000	288	252
35 000	302	261




*Fathom* allows only letters, digits, and underscores in attribute names.

Do not enter spaces between digits or the numbers will be treated as text.

To easily see all the information in a table, click and drag the edges of the table object and of the attribute cells.


### ■ Entering data in a table

- Start *Fathom* and begin a new blank document.
- Click on the **Table** icon, then click in the document to start a table. 
- In the table, click **<new>** and enter the attribute **Value\_of\_car**. Press **Enter**.
- A new cell should appear. Click **<new>** and enter the attribute **Liability\_ins\_cost**. Press **Enter**.
- Repeat the process to enter the attribute **Collision\_ins\_cost**.
- Enter the data into the table.
- Double click on **Collection 1** in the upper left corner. Rename it **Car Insurance Costs**. Your table should look something like this.

	Value_of_car	Liability_ins_cost	Collision_ins_cost	<new>
1	5000		247	138
2	10000		233	163
3	10000		275	154
4	15000		291	192
5	15000		277	185
6	20000		243	257
7	20000		315	201
8	25000		254	233
9	30000		288	252
10	35000		302	261

### ■ Creating a scatter plot

You will begin by investigating whether there is a relationship between the liability insurance cost and the value of the car.

- Click on the **Graph** icon, then click in the document to start a graph. 
- Drag the **Value\_of\_Car** table attribute to the horizontal axis of the graph. You should see some data plotted. Ignore them for now.
- Drag the **Liability\_ins\_cost** table attribute to the vertical axis of the graph. The data should be plotted accurately now.

1. Does there appear to be a relationship between the two variables? If so, explain why and describe it. If not, explain why not.

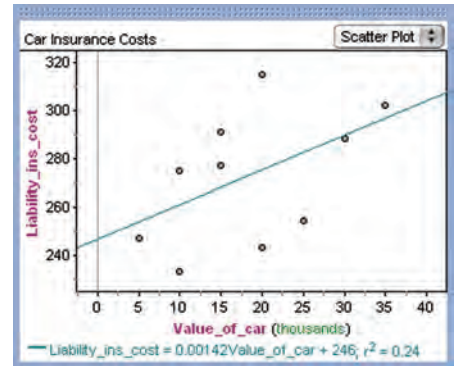
The **Graph** menu is in the toolbar above the **Graph** icon. It will not appear unless a graph is selected.

*Least-squares line* is another term for the *line of best fit*.

### ■ Drawing a line of best fit

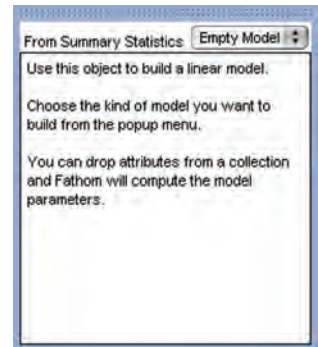
- Click on the graph, then click on the **Graph** menu.
- Select **Least-Squares Line**.  
The line of best fit should appear on the graph.  
The equation of the line appears below the graph.

2. a) Describe the line of best fit.
- b) Do you predict the correlation coefficient will be positive or negative? Will it be closer to 1 or closer to 0?  
Justify your predictions.

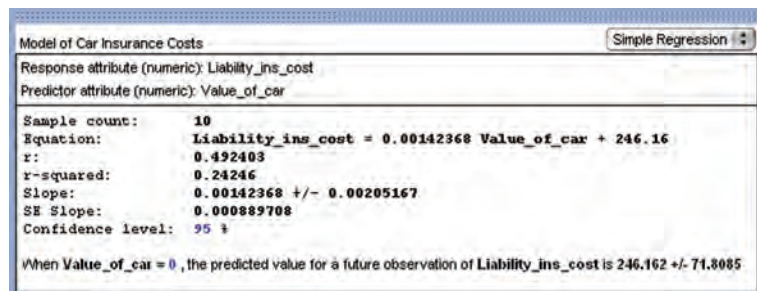


### ■ Determining the correlation coefficient

- Click the **Model** icon, then click in the document to start a statistical model of the data.
- Click on **Empty model**. A pop-up menu should display statistical calculations that can be performed.  
Select **Simple Regression**.
- Drag the **Value\_of\_car** table attribute onto **Predictor attribute** in the model object.
- Drag the **Liability\_ins\_cost** table attribute onto **Response attribute** in the model object.  
Simplify the information that appears by right-clicking in the model object, then deselecting **Verbose**.



The *predictor attribute* is the independent variable. The *response attribute* is the dependent variable.



3. The correlation coefficient,  $r$ , is displayed in the model object.
  - a) Record the value of  $r$  to 2 decimal places.
  - b) Is the correlation is strong or weak? Justify your answer.
  - c) How does it compare to your prediction in question 2 part b?



4. The equation of the line of best fit also appears in the model object.
  - a) Write the equation. Use  $x$  to represent **Value\_of\_car**, and  $y$  to represent **Liability\_ins\_cost**.
  - b) How confident would you be in interpolating or extrapolating liability insurance costs using this equation? Explain your thinking.

■ **Analysing another set of data**

Collision insurance only pays for damages to your vehicle in an accident. Collision insurance is not required by law in Ontario. You will investigate whether collision insurance cost is related to the value of the car.

5. Without deleting your first graph, drop a new graph in the document. You will graph car values and collision insurance costs.
  - a) Which data should you place on the horizontal axis? Which data should you place on the vertical axis? Justify your answers.
  - b) Plot the data as you described in part a.
  - c) Does there appear to be a relationship between the variables? If so, tell why and describe it. If not, tell why not.
6. Display the line of best fit on the graph.
  - a) Describe the line.
  - b) Predict whether the correlation coefficient will be closest to 0, 0.25, 0.5, 0.75, or 1. Explain your thinking.
7. Drop the **Collision\_ins\_cost** table attribute onto **Response attribute** in the model object.
  - a) What is the value of the correlation coefficient? Record your answer to 2 decimal places.
  - b) How does it compare to your prediction in question 6 part b?
8.
  - a) Write the equation of the line of best fit. Use  $x$  to represent **Value\_of\_car**, and  $y$  to represent **Collision\_ins\_cost**.
  - b) How confident would you be in interpolating or extrapolating collision insurance costs using this equation? Explain your thinking.
  - c) Use the equation in part a. Predict the costs of collision insurance for cars with each value.

**Car P:** \$12 000

**Car Q:** \$40 000

9. When Lena was getting a collision insurance quote from a broker, she mistakenly stated that the vehicle would be used for business rather than for pleasure. As a result, one piece of data is invalid.
  - a) Which point do you think it is? What is the name for this kind of point? Delete the point by right-clicking on it and selecting **Delete Case**.
  - b) Describe how the line of best fit changes.
  - c) What is the new  $r$ -value?
  - d) What is the new equation of the line of best fit?
  - e) What change occurred in the table when you deleted the point?
  
10. What assumptions do you need to make about Lena's data collection method in order to consider her data reliable?

### Part B: Working with Data Imported from E-STAT

#### ■ Developing a conjecture

In today's busy world, it can be difficult to juggle work and family. For example, parents may need to take time off work to care for sick children.



11. How might the number of absences due to family responsibilities for working mothers with children under the age of 5 be related to the number of children under 5 in Canada? Describe the correlation, if any, you might expect between these variables. This is your **conjecture** about the relationship.

#### ■ Finding E-STAT data

- Go to the Statistics Canada Web Site. Click **English**. Select **Learning resources** from the menu on the left. Click on **E-STAT** in the yellow box on the right. Click on **Accept and enter**. If you are working from home, you will need to enter the user name and password assigned to your school.

[Français](#) | [Contact us](#) | [Help](#) | [Search](#) | [Canada site](#)  
[Site map](#) | [About us](#) | [Privacy](#) | [Accessibility](#) | [My account](#)

**STATISTICS CANADA**  
 CANADA'S NATIONAL STATISTICAL AGENCY

**E-STAT: Table of contents**

**Economy**  
[Business performance and ownership](#) | [Manufacturing](#)  
[Business, consumer and property services](#) | [Prices and price indexes](#)  
[Construction](#) | [Retail and wholesale](#)  
[Economic accounts](#) | [Science and technology](#)  
[Information and communications technology](#) | [Transportation](#)  
[International trade](#)

**Land and Resources**  
[Agriculture](#) | [Environment](#)  
[Energy](#)

**People**  
[Aboriginal peoples](#) | [Income, pensions, spending and wealth](#)  
[Children and youth](#) | [Labour](#)  
[Culture and leisure](#) | [Languages](#)  
[Education, training and learning](#) | [Population and demography](#)  
[Ethnic diversity and immigration](#) | [Seniors](#)  
[Families, households and housing](#) | [Society and community](#)  
[Health](#) | [Travel and tourism](#)

**Nation**  
[Crime and justice](#) | [Government](#)

**Historical Censuses of Canada**  
 1665-1871

**Elections Canada**  
 2000: Provinces and Territories | [2000: Federal electoral districts](#)  
 1997: Provinces and Territories | [1997: Federal electoral districts](#)

- The table of contents window will open.
- In the *People* section, click on **Labour**. From the list of CANSIM data, click on **Labour mobility, turnover and work absences**. Click on table **279-0033**.
- Under *Sex*, select **Females**.
- Under *Presence of children*, select **Preschoolers, under 5 years**.
- Under *Absence rates*, select **Days lost per worker in a year, personal or family responsibility**. Set the *Reference Period* from **1997 to 2006**.
- Click on **Retrieve as individual Time Series**.

You have selected data about the number of work days lost due to personal or family responsibilities for women with children under 5 from 1997 to 2006. Now you need to find the population of children under 5 over the same period of time.

- Scroll to the bottom of the screen and click on **Add more series**. Click on **Browse by subject**. The CANSIM table of contents appears. Click on **Population and demography**.

[Français](#) | [Contact us](#) | [Help](#) | [Search](#) | [Canada site](#)  
[Site map](#) | [About us](#) | [Privacy](#) | [Accessibility](#) | [My account](#)

**STATISTICS CANADA**  
 CANADA'S NATIONAL STATISTICAL AGENCY

**CANSIM by subject**

- [Aboriginal peoples](#)
- [Agriculture](#)
- [Business performance and ownership](#)
- [Business, consumer and property services](#)
- [Children and youth](#)
- [Construction](#)
- [Crime and justice](#)
- [Culture and leisure](#)
- [Economic accounts](#)
- [Education, training and learning](#)
- [Energy](#)
- [Environment](#)
- [Ethnic diversity and immigration](#)
- [Families, households and housing](#)
- [Government](#)
- [Health](#)
- [Income, pensions, spending and wealth](#)
- [Information and communications technology](#)
- [International trade](#)
- [Labour](#)
- [Languages](#)
- [Manufacturing](#)
- [Population and demography](#)
- [Prices and price indexes](#)
- [Retail and wholesale](#)
- [Science and technology](#)
- [Seniors](#)
- [Society and community](#)
- [Transportation](#)
- [Travel and tourism](#)

- Click on **Population estimates and projections**.  
Then click on table **051-0001**.  
Under *Geography*, select **Canada**.  
Under *Sex*, select **Both sexes**.  
Under *Age group*, select **0 to 4 years**.  
Set the *Reference Period* from **1997 to 2006**.  
Click on **Retrieve as individual Time Series**.  
The titles of both data series you requested should be displayed.

Select an output format for the data.

- In the *Screen output – table* section, under *Plain text*: select **Table, time as rows**.  
Click on **Retrieve now**.  
The table should appear in text format.

### ■ Importing and using E-STAT data in *Fathom*

- Highlight the data in the table, but do not include the column headings or any other information. Right click and **Copy** the highlighted material.
- Open a new *Fathom* document.  
Click the **Collection** icon, then click in the document to start a collection of data. 📦  
Right click and **Paste Cases** into the collection.
- Click on the collection object, then drop a table into the document.  
The E-STAT data should immediately fill the table.  
Change the table attributes to **Year, Population\_under\_5, and Family\_absences**.

To rename a table attribute, double click the name.

Collection 1				
	Year	Population_under_5	Family_absences	<new>
1	1997	1917294	4.1	
2	1998	1872747	3.5	
3	1999	1828982	4.1	
4	2000	1791178	4.1	
5	2001	1759196	4.5	
6	2002	1730473	5.2	
7	2003	1710647	4.8	
8	2004	1705488	4.5	
9	2005	1702406	5.1	
10	2006	1712848	6.2	

- 12.** Drop a graph into the document.
- Make a scatter plot with population as the independent variable and absences as the dependent variable.
  - Describe the relationship and whether your conjecture from question 11 was correct.
- 13.** In the last 10 to 15 years, many companies have improved their policies about personal days for employees. Drop a new graph into the document.
- Make a scatter plot with year as the independent variable and absences as the dependent variable.
  - Describe the relationship.
- 14.** Display the line of best fit for each graph.
- Use a **Model** object to view the correlation coefficient for each line.
  - Which relationship shows a stronger correlation?
  - What does this suggest about the data?
- 15.** What other factors could be affecting the number of days parents with young children take off for personal and family matters?



### Reflect

- Suppose you want to create a scatter plot relating two variables for a table of data in *Fathom*. Explain how to decide which variable to plot along the horizontal axis.
- Why is it useful to know the equation of the line of best fit? Include an example.
- Researchers often make a conjecture before they begin to investigate a possible correlation between variables. What are advantages and disadvantages of making a conjecture before collecting and analysing data?

# 3.7

## Conducting an Experiment to Collect Two-Variable Data

College students in fishery and aquaculture programs look for correlations between variables such as water temperature and egg hatching, water depth and water quality, or types of food and fish growth.



### Inquire

#### Materials

- *Fathom*, *Microsoft Excel*, TI-83 or TI-84 graphing calculator, or grid paper
- materials for the experiment

### Designing and Conducting Experiments

Work with a partner or in a group.

#### ■ Planning a question or conjecture

When you pose a question or conjecture for an experiment to collect two-variable data, you should follow these steps.

- Find a topic that interests you.
  - Identify the two variables for the topic.
  - Pose a question or make a conjecture about a possible relationship between the variables.
1. Identify the variables that are being studied in each experiment.
    - a) Kasey wants to prove that a person's height is positively related to the number of basketball free throws he or she can successfully make in 10 attempts.
    - b) Ramon wants to determine if there is a relationship between a person's height and walking speed.
  2. Write a question or conjecture for each experiment in question 1.



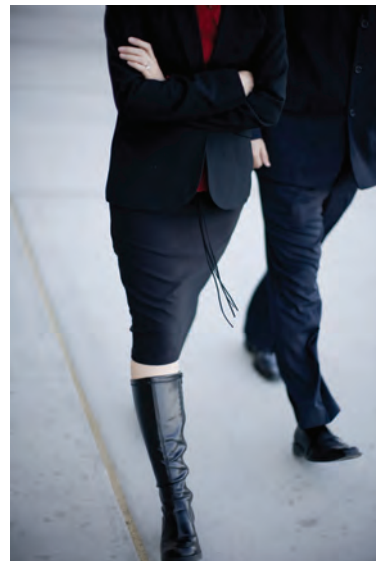
### ■ Developing a method

When you develop a method for measuring or collecting data, think about these issues.

- Reduce the effects of other variables by keeping all other factors in the experiment the same, as much as possible.
- If you cannot think of a way to measure each variable, plan a different experiment.

**3.** Refer to Ramon's experiment in question 1.

- a) What method could Ramon use to measure height?
- b) What method could he use to measure walking speed?
- c) What can be different about the people he includes in his experiment?
- d) Explain why the people should be about the same in these ways: age, body type, and fitness level.



### ■ Determining what materials you will need

As you develop a method, plan what you will need for the experiment. Keep the availability and the cost of materials in mind.

**4.** Refer to question 3. Describe materials for Ramon's experiment.

### ■ Writing a plan

A plan should include these items.

- A statement outlining the question you want to answer or the conjecture you want to prove
- Details of your method, including the expected length of observation time or the number of observations or trials
- A strategy to limit the influence of other variables
- A list of the materials
- Any tables or checklists needed for recording your findings

You should have at least one person read your plan. Ask the reader if he or she understands your plan, if anything has been left out, and if the plan seems possible.

5. Safety/health issues, expense issues, privacy issues, and sensitivity to people's feelings are reasons an experiment may not be appropriate. Explain how one or more of these issues might apply to each experiment.
- Faria wanted to conduct a blind taste test where people of various ages had to identify what they had just eaten as quickly as possible.
  - Davian's experiment was aimed at finding a connection between hours spent doing homework and the mark earned in a course.
  - Travis likes to work out. His experiment involved looking for a correlation between the circumference of a person's bicep and the maximum weight he or she can bench-press.

■ **Avoiding problems when collecting experimental data**

- Except for the variables you are measuring, keep the characteristics of your sample as similar as possible.
- Collect a large sample of data that includes a range of values.
- Measure or count as accurately as possible.



6. Explain the problem with how each person collected data.
- Alisha investigated how height is related to successful basketball free throws. To find people of different heights, she asked a student from each of Grades 2 to 12.
  - For a ball-rolling experiment, Dexter needed to mark a distance of 20 feet. He estimated that one of his paces is 4 feet long, then took five paces, and marked the distance.

- c) For an experiment comparing leg length to jump height, Alex collected data from his parents and three sisters. He plotted all five points, and concluded there was no correlation.
- d) Madison wanted to determine if there is a relationship between a person's age and the ability to hear high-pitched frequencies. She collected data from many students in her grade.



- 7.** The student-run store at Bernadette's school sells coffee, tea, and hot chocolate. She wants to determine if there is a relationship between the outside temperature and the number of hot drinks the store sells before school.
- a) What factors other than temperature might affect the sales? How can Bernadette work around these factors?
  - b) For how many days should Bernadette collect data? Explain your reasoning.
  - c) Write a plan for Bernadette's experiment. Include tables or checklists she would need.

**■ Planning and conducting your own experiment**

- 8.** Choose one of these topics or think of your own topic.

Is there a correlation between these variables?

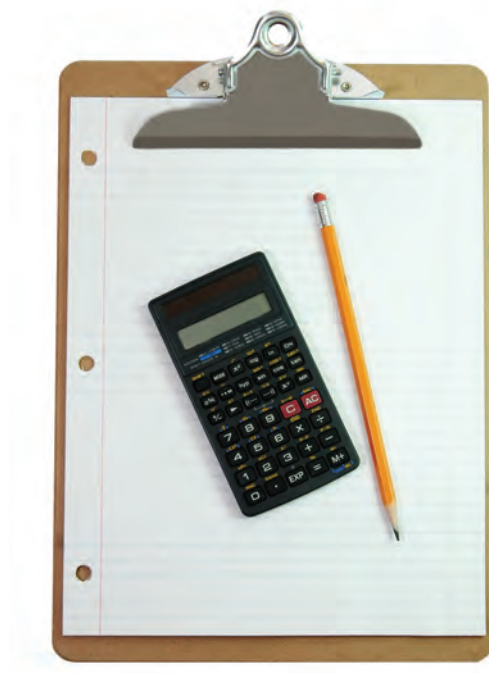
- The circumference of a straw and how long it takes to drink a specific volume of liquid
- Hand span and the number of small items a person can pick up at once
- The number of students in the cafeteria during a period and the number of students in the library
- The number of sit-ups and number of push-ups that a person can do in 1 min
- Age and a person's ability to remember objects that are displayed for 1 min, then hidden
- Height and a person's lung capacity
- The type size of a page of text and how long it takes to read the text

- Identify the variables for your topic. Develop a question or conjecture about a possible relationship between the variables.
- Write a plan for an experiment to test your question or conjecture. Include the elements outlined in this lesson.
- Get approval from your teacher.
- Gather the materials.
- Conduct the experiment.

If computers or graphing calculators are available, use them to plot your data.

## ■ Displaying and analysing your data

9. Make decisions about how you will display your data in a scatter plot.
- Create a scatter plot for your data.
  - Is there a correlation? If so, describe it. If not, tell how the scatter plot shows this.
  - If appropriate, include a line of best fit and its equation. If a linear model is not appropriate, explain why.
  - Write a conclusion about your question or conjecture. Tell how your scatter plot supports this.



### *Reflect*

- What was the most challenging part of planning or conducting your experiment? How did you deal with this challenge?
- What would you do differently if you were to repeat the experiment?
- If you modelled your data with a line of best fit, how confident are you in the model? Justify your answer.

# Study Guide

## Scatter Plots and Correlations

A correlation indicates a relationship between two variables.

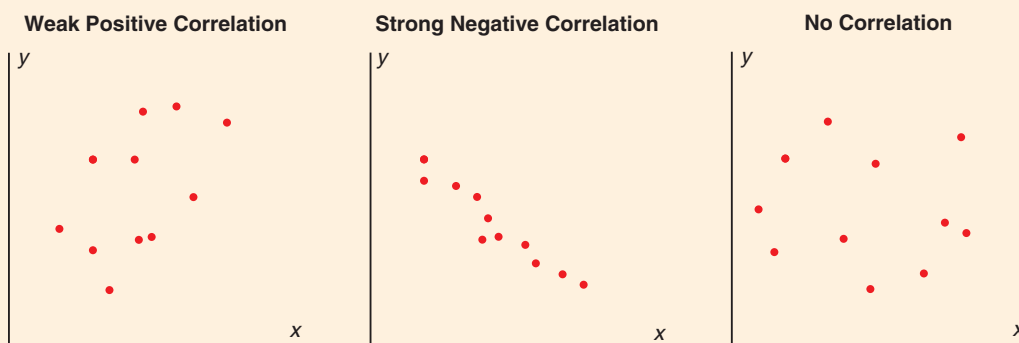
- In a positive correlation, points on the scatter plot go up to the right.
- In a negative correlation, points go down to the right.

A correlation does not necessarily indicate a cause-and-effect relationship.

A linear correlation is weak if the points are spread out.

A linear correlation is strong if the points appear to lie along a line.

If there is no trend, there is no correlation.



A point that does not follow the trend shown by the rest of the data is an outlier.

## Line of Best Fit

Other names for the line of best fit are *regression line* or *trend line*.

To determine an equation of the line of best fit, find its slope,  $m$ , and  $y$ -intercept  $b$ :

$$y = mx + b$$

Predictions for data values between known points are called **interpolation**.

Predictions for data values beyond known points are called **extrapolation**.

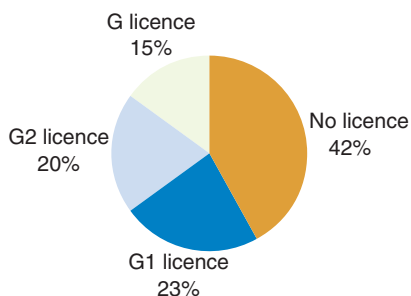
These factors can reduce the reliability of a linear model:

- A weak linear correlation
- A non-linear correlation
- Data that are too clustered
- Outliers in the data
- A data sample that is too small
- A data sample that is biased

## Chapter Review

- 3.1** 1. State whether each situation involves one-variable or two-variable data.

a) **Driving Status of Students in Course**



b)

Monthly Precipitation	
Rainfall (mm)	Snowfall (cm)
19	32
21	26
35	20
56	7
69	0

- c) According to data Hayden collected, his classmates watch an average of 11.4 h of television per week.

2. Each situation below involves two-variable data. Identify the two variables.

- a) In science class, students learn that air pressure decreases as height above Earth's surface increases.
- b) Assuming that rainfall stays within the seasonal range, higher amounts of rainfall increase crop yield.
- c) The time taken to cook a turkey increases as the turkey's mass increases.

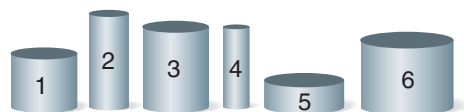
- 3.2** 3. Predict whether the variables in each situation would have a negative correlation, a positive correlation, or no correlation.

- a) The number of people playing a board game and the likelihood that you will win
- b) The mass of a car and its selling price
- c) The number of people in a household and the monthly household water use
- d) The size of a pizza and the number of toppings on it
- e) The speed of a river's current and the time it takes to travel downstream

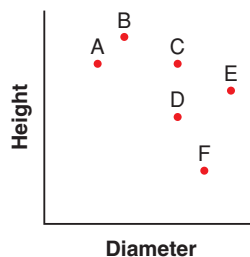
4. State the independent variable and dependent variable for each situation.

- a) Aquariums that hold more water are more expensive.
- b) The more rain we have, the less time people spend watering their lawns.

5. Identify which cylinder is represented by each point in the scatter plot.



**Cylindrical Containers**



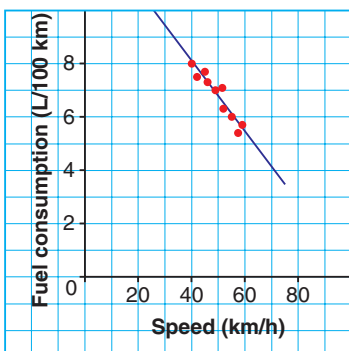
6. This table compares the number of jigsaw puzzle pieces and the manufacturer's recommended minimum age.

Number of pieces in puzzle	Recommended minimum age
100	5
500	7
48	3
65	4
300	9
150	6
550	12
300	6
200	8
100	4

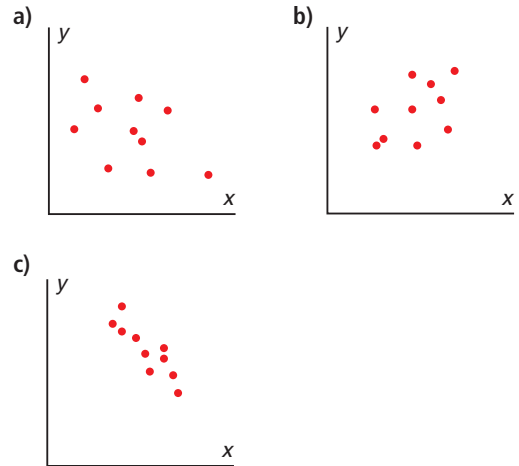
- a) Graph the data and sketch the line of best fit.  
 b) What conclusion can you make from the graph?
7. Explain why the following data analysis is not reasonable. There may be more than one reason.

A car manufacturer tested one of their vehicle models and concluded that the linear correlation between speed and fuel consumption is strong.

Effect of Vehicle Speed on Fuel Consumption



8. Describe each correlation.



9. The Body Mass Index (BMI) relates a person's weight and height. A clinician gathered age and BMI data for a group of people aged 8 to 17 years.

Age (years)	BMI
8	15.5
8	17.0
10	16.0
10	18.0
12	20.0
13	19.0
13	20.0
14	20.0
15	21.0
15	20.0
16	22.5
16	21.0
17	22.0
17	21.0

- a) Graph the data. Identify any outliers, if they occur.  
 b) Sketch the line of best fit and determine its equation.  
 c) Describe any correlation you see between the variables. What conclusion could someone make based on the scatter plot?  
 d) How many people were included in the sample? Do you think these data are enough to justify a conclusion about people in this age group? Explain.

**3.4 10.** Refer to question 6.

- a) Create a scatter plot of the data.
- b) Use technology to determine the equation of the regression line.
- c) How many pieces would be in a puzzle with a recommended minimum age of 10?
- d) What would be the recommended minimum age for a jigsaw puzzle with 400 pieces?



**11.** Open the file *daytona.xls*, which gives data on average speeds of the winning drivers of the Daytona 500.

	A	B	C
1	Year	Winning driver	Average speed (mph)
2	1960	Junior Johnson	125
3	1965	Fred Lorenzen	142
4	1970	Pete Hamilton	150
5	1975	Benny Parsons	154
6	1980	Buddy Baker	178
7	1985	Bill Elliot	172
8	1990	Derrike Cope	166
9	1995	Sterling Marlin	142
10	2000	Dale Jarrett	156
11	2005	Jeff Gordon	135

- a) Create a scatter plot for the data. Describe the correlation.
- b) Create a line of best fit. Use it to predict the speed of the winning driver in 2010.
- c) How reliable is the prediction you made in part b? Explain.
- d) Do you think it is acceptable to include only every 5th year? Justify your answer.

**3.6 12.** The table shows the rate of injury among young workers compared to the actual number of injury claims that were submitted between 1996 and 2005.

Number of claims	Injury rate (%)
11 657	5.0
11 612	5.0
11 006	4.8
9980	4.4
11 040	4.4
10 595	4.1
8630	3.5
8962	3.4
9460	3.5
10 280	3.4

- a) Open *Fathom*. Create a table and scatter plot for these data.
- b) Draw a least-squares line and determine the value of the correlation coefficient.
- c) What does the *r*-value suggest about the type of correlation that exists between the variables?
- d) How might you explain that in 2005, the injury rate decreased but the number of injury claims increased?

**3.7 13.** Esin wants to determine if there is a correlation between a person's weight and the number of chin-ups the person can perform in 1 min.

- a) Write a plan for an experiment Esin could perform to collect data.
- b) Esin wants volunteers with a variety of weights. How could he keep other physical factors from affecting his results?
- c) What ethical issues will Esin have to consider?



## Practice Test

Multiple Choice: Choose the best answer for questions 1 and 2. Justify each choice.

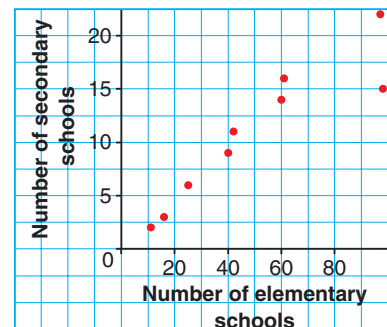
- Which is another name for the line of best fit?
  - Response line
  - Explanatory line
  - Trend line
  - Predictor line
- Which phrase best describes the correlation between the number of people attending a movie and the number of empty seats remaining in the theatre?
  - Strong and positive
  - Strong and negative
  - Weak and positive
  - Weak and negative

Show your work for questions 3 to 6.

- Knowledge and Understanding** This table shows data for ten players from the Toronto Blue Jays for the 2007 season.
  - Create a scatter plot of the data.
  - Does there appear to be a correlation between the number of times at bat and the number of hits? If so, describe it. If not, explain why not.
  - Draw a line of best fit. Use it to estimate the number of hits a player might have after 100 times at bat.
- Communication** Rosa wants to use data she has collected to investigate the correlation between people's leg strength and arm strength, then use the data to make predictions. Describe the steps she should follow.
- Thinking** Zaki heard that if you drop a piece of buttered bread, it lands with the buttered side down. Zaki decided to test this theory.
  - Briefly describe an experiment Zaki could conduct.
  - What are some issues Zaki should consider to ensure that his experimental results are valid?
  - Is Zaki's experiment about a correlation between two variables? Explain.
- Application** Use this scatter plot.
  - Describe the correlation.
  - Would it be reasonable to draw a line of best fit for these data? Justify your answer.
  - Identify a point that might be considered an outlier. What might the outlier suggest about this particular school district?

At bats	Hits
643	191
385	101
608	177
584	143
531	147
327	82
357	103
425	102
290	69
331	80

Schools in Nine Ontario School Districts



# Chapter Problem

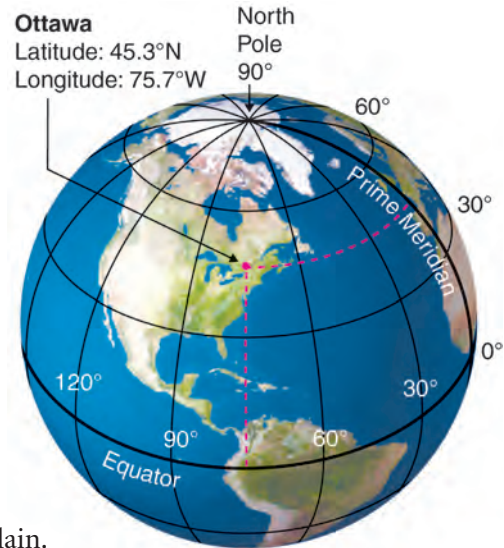
# Temperatures around the Globe

## Materials

- *Microsoft Excel* or *Fathom* or grid paper
- *jantemps.xls* or *jantemps.ftm*

*Latitude* and *longitude* describe the location of places on Earth.

- Latitude describes the location in degrees north or south of the Equator.
- Longitude describes the location in degrees east or west of the Prime Meridian.



1. Predict whether there is a relationship between the mean January temperature of a North American city and the city's position north of the Equator. Explain.



2. Open *jantemps.xls* or *jantemps.ftm*. Create a scatter plot for January temperature and latitude or use grid paper to plot the data shown in this table. Describe the correlation.

Plot latitude on the horizontal axis.

Temperatures				
	City	JanuaryTemp_deg_F	Latitude_deg_N	Longitude_deg_W
1	Key West, FL	65	25.0	82.0
2	New Orleans, LA	45	30.8	90.2
3	Atlanta, GA	37	33.9	85.0
4	San Francisco, CA	42	38.4	123.0
5	St. Louis, MO	24	39.3	90.5
6	Denver, CO	15	40.7	105.3
7	New York, NY	27	40.8	74.6
8	Toronto, ON	23	43.6	79.6
9	Ottawa, ON	12	45.3	75.7
10	St. John's, NF	22	47.9	52.5
11	Vancouver, BC	37	49.2	123.2
12	Winnipeg, MB	-1	49.9	97.2
13	Edmonton, AB	9	53.5	113.5
14	Whitehorse, YT	-1	60.6	135.6
15	Yellowknife, NT	-18	62.5	114.5

3. Draw a line of best fit and determine the equation of the line.

4. Predict whether there is a relationship between the mean January temperature of a city in North America and the city's position west of the Prime Meridian. Check your prediction.
5. Make your own prediction about the data. Check your prediction.