

Two-Variable Data Analysis

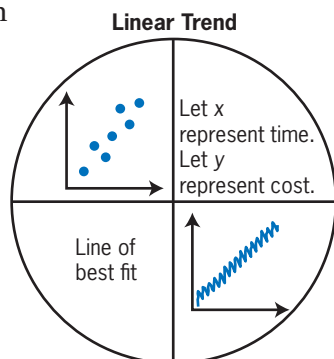
Climate change is an area of increasing concern. Shrinking arctic ice caps pose a danger to both animals and people. Polar bears are losing their habitat, while human residents in low-lying coastal regions are at risk of flooding. Some argue that climate change is not a real threat. The use of data is important when trying to assess such an important issue. Do you think climate change is a natural occurrence, or has human behaviour contributed? What are some of the variables that could be measured and compared to help answer this question?

Key Terms

line of best fit	presumed relationship
linear correlation	reverse cause and effect
correlation coefficient	relationship
linear regression	accidental relationship
cause and effect	residual plot
relationship	residual
common cause	outlier
relationship	hidden variable

Literacy Strategy

You can use concept circles to visually organize words and symbols that are related. As you work through the chapter, create concept circles to help you organize how different ideas relate to each other. An example is shown.



Career Link



Health Care Statistician

A health care statistician uses data management techniques to help advance medical treatments and educate the public. Responsibilities include analysing and interpreting health care data using statistical software. Managing and monitoring the integrity of data collection methods is also a significant part of this job. Health care statisticians usually have an advanced degree in statistics or mathematics and some leadership experience in health care. What are some factors that a health care statistician would be concerned about, and how could they be connected using two-variable data analysis?

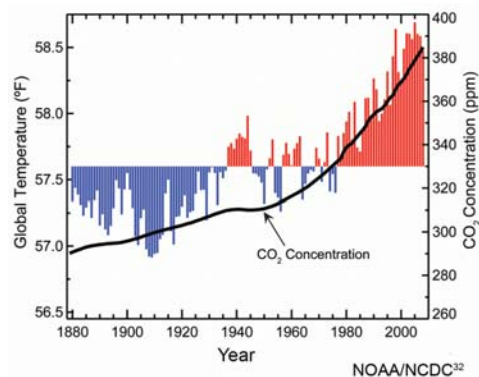


Chapter Problem

Climate Change

Scientists generally agree that there is sufficient evidence to declare climate change is a real phenomenon with dangerous implications that need to be addressed. The graph shows the mean yearly temperature in degrees Fahrenheit ($^{\circ}\text{F}$) and carbon dioxide (CO_2) concentration levels in parts per million (ppm) over time for Earth.

- The mean temperature over this time period is approximately 57.6°F . How is this represented in the graph?
 - What do the blue bars represent?
 - What do the red bars represent?
- Did Earth's temperature change by the same amount from year to year? Is there a long-term trend over several decades? Explain your reasoning.
- The solid black curve shows the carbon dioxide concentration level over time. Describe this trend.
- Perform some research on climate change. Describe some of the physical, ecological, and social effects of climate change.



Source: United States Global Change Research Program

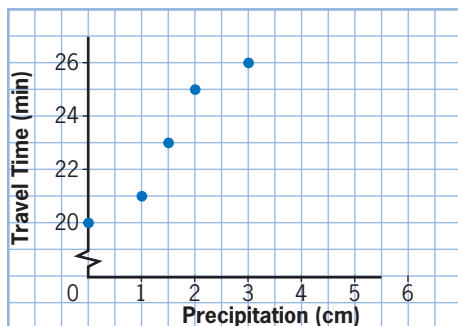
Literacy Link

In the United States, degrees Fahrenheit is used, so it is frequently seen in reports from the U.S., rather than degrees Celsius.

Prerequisite Skills

Scatter Plots

- The graph shows a commuter's travel time in minutes versus the amount of daily precipitation in centimetres.



- Describe the trend in the data.
 - Suggest why this trend may exist.
- The table shows Anna-Marie's average speed for a number of running races.

Distance (m)	Average Speed (m/s)
100	8.1
200	7.5
500	6.6
1000	6.1
1500	5.2

- Create a scatter plot of the data.
- Describe the trend.
- Estimate Anna-Marie's speed for a 750-m race. Explain your method.

Use a Graphing Calculator to Create a Scatter Plot

- Use a graphing calculator to create a scatter plot for the data.
 - Is there a trend in the scatter plot? If so, describe it.

Goals	Assists
21	30
18	24
35	30
12	13
27	37
6	9
40	55
32	31

Use a graphing calculator to create a scatter plot of the data.

Age	Height (cm)
14	140
14	146
14	150
15	148
15	160
16	157
17	171
18	170

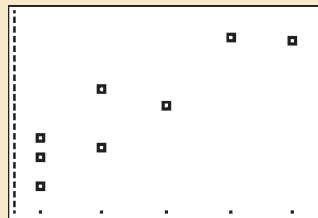
- Press **STAT** and choose **1:Edit**.
- Enter the data into lists **L1** and **L2**.

L1	L2	L3	2
14	150		
15	148		
15	160		
16	157		
17	171		
18	170		
---	-----		
L2(9) =			

- Press **2ND** **STAT PLOT**.
- Set the parameters as shown.

2ND	PLOT2	PLOT3
Off		
Type:		
Xlist:	L1	
Ylist:	L2	
Mark:	+	

- Press **ZOOM** and choose **ZoomStat**.



Press **TRACE** and use the left and right arrows to read the coordinates of the data.

Linear Models

Use this information to answer #4 and #5. Temperatures measured in Fahrenheit, F , and Celsius, C , are related by the following equation: $F = 1.8C + 32$.

4. a) Graph the relation for $C = 0^\circ$ to $C = 100^\circ$.
b) Identify the initial value. Explain what it means.
c) Identify the rate of change. What does it mean?
d) Use the graph to convert 25°C to degrees Fahrenheit.
e) Check your answer using the equation.
f) Which method do you prefer, and why?
5. a) Use the equation to convert 50°F to degrees Celsius.
b) Check your answer using the graph.
c) Which method do you prefer, and why?

Bias

Literacy Link

Bias occurs when data are collected or presented unfairly. It can lead to an inaccurate interpretation of the results of a statistical study.

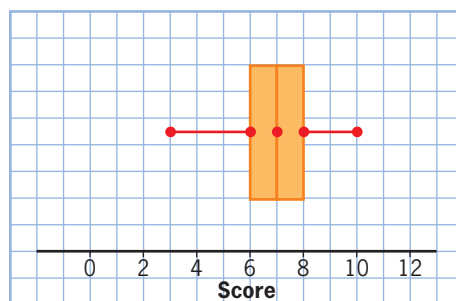
6. A student council surveys the football team to see if they should ask the principal for an increase in the football program's budget.
a) How is bias present in this study?
b) What could be done to remove the bias?
7. A radio talk show host asks callers if they think the current government should be unseated. Identify any potential sources of bias.

Summary Statistics for One-Variable Data

8. The chart shows the heights of the players of a high school basketball team.

Player	Height (m)
Sarah	1.8
Jessica	1.5
Tina	1.5
Latisha	1.5
Uma	1.4
Kyla	1.9
Mina	1.4
Luisa	1.6
Sangita	1.5
Caroline	1.9

- a) Determine the mean, median, and mode.
 - b) Explain why these measures of central tendency are not all equal.
 - c) Determine the standard deviation and z -score for a height of 1.9 m.
9. The box plot summarizes students' quiz scores out of 10.



- a) Determine the range of scores.
- b) Determine the median score.
- c) What is the interquartile range and what does it represent?

Line of Best Fit

Learning Goals

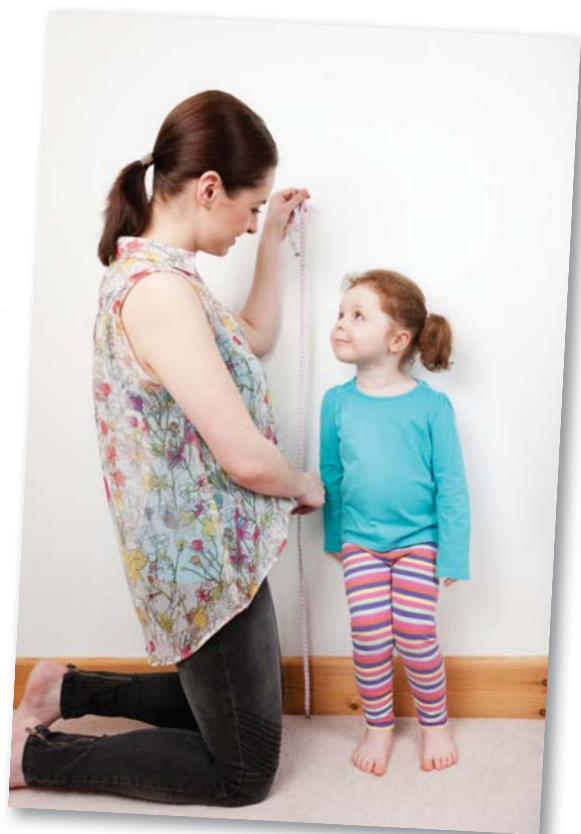
I am learning to

- classify a linear correlation between two variables
- determine a correlation coefficient using technology
- produce a line of best fit using linear regression

Minds On...

Some parents like to track their children's growth over time. One way to do this is to measure the child's height on every birthday and record it on a wall or door frame.

- What are some other ways to record these data?
- Do you think a child grows by the same amount every year?



Action!

Investigate Line of Best Fit

Materials

- transparent ruler

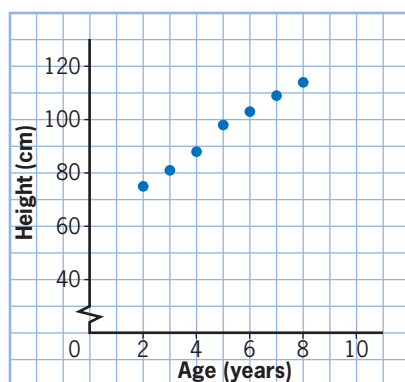
Literacy Link

A *time series* is a set of two-variable data in which a quantity is measured against time.

Literacy Link

An *independent variable* is a not affected by another variable. A *dependent variable* is affected by another variable.

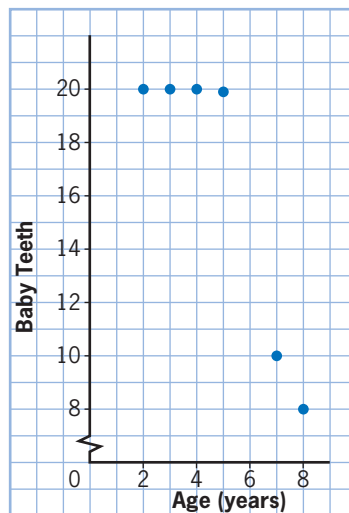
Darla's parents kept track of her height on her birthday every year from ages two through eight. The graph shows the measurements as a time series.



1. Identify the independent and dependent variables. How do you know?
2. Does there appear to be a relationship between Darla's height and her age? If so, describe it.

3. a) Sketch the graph and draw a **line of best fit**.
 b) Does it make sense to draw a solid or a dashed line in this case? Explain why.
4. Extend the line to answer the following questions.
 - a) How tall will Darla be at age 10?
 - b) Estimate Darla's height when she was a newborn.
5. **Reflect**
 - a) How accurate do you think these estimates are? Explain.
 - b) Will this trend continue forever? If not, when will this linear model no longer be useful? Explain your reasoning.

Darla's parents also kept track of the number of baby teeth she still had, except for one year when they forgot. The data, collected on her birthdays, are shown.



6. Identify the independent and dependent variables.
7. Does there appear to be a correlation between the number of baby teeth Darla had and her age? If so, describe it.
8. a) Sketch the graph and draw a line of best fit.
 b) Does it make sense to draw a solid or a dashed line in this case? Explain why.
 c) Use the line to estimate how many baby teeth Darla had on her sixth birthday.
9. **Reflect**
 - a) How accurate do you think this estimate is? Explain.
 - b) Will this trend continue forever? Why or why not?
10. **Reflect** Compare the two lines of best fit that you drew in this investigation.
 - a) Was one easier to draw than the other?
 - b) Was one a better model for the data than the other? Justify your answers.
11. **Extend Your Understanding** Create a data table relating two variables for which it would be nearly impossible to draw a line of best fit that makes sense. Create a scatter plot and explain why a line of best fit does not make sense for the graph you created.

line of best fit

- a straight line that represents a trend in the scatter plot as long as the pattern is more or less linear
- should pass through as many points as possible, with about half the points above and half below the line
- a solid line represents continuous data that are constantly changing
- a dashed line represents discrete data that change only in steps

Literacy Link

Interpolation and extrapolation are important tools for two-variable data analysis. To interpolate, read a linear model between given data points. To extrapolate, read beyond given data points.

linear correlation

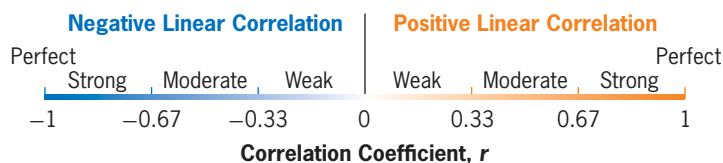
- a relationship in which a change in one variable tends to correspond to a proportional change in another variable

correlation coefficient

- a measure of how well a linear model fits a two-variable set of data
- values of r between -1 and 0 indicate a negative correlation, so the line of best fit has a negative slope
- an r value of 0 indicates that there is no linear correlation
- values of r between 0 and 1 indicate a positive correlation, so the line of best fit has a positive slope

In the Investigation, Darla's height changed largely in proportion to her age. As she got older, her height increased by an almost constant amount each year. This is an example of two variables that share a strong **linear correlation**. When two variables have a weaker linear correlation, a trend is still evident; however, a line of best fit is more difficult to recognize, as was the case in the number of Darla's baby teeth versus time. When two variables have no linear correlation, there is no recognizable linear pattern to the data.

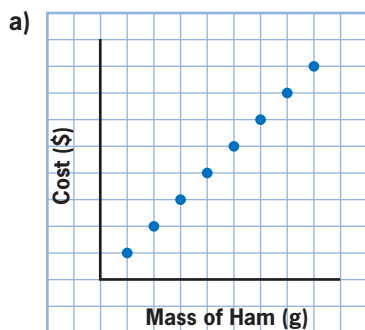
The **correlation coefficient**, r , is a measure of the strength of the linear correlation between two variables. For a positive correlation, r can have values between 0 , which represents no linear correlation, and 1 , which represents a perfect positive linear correlation. For a negative correlation, 0 signifies no linear correlation and -1 indicates a perfect negative linear correlation.



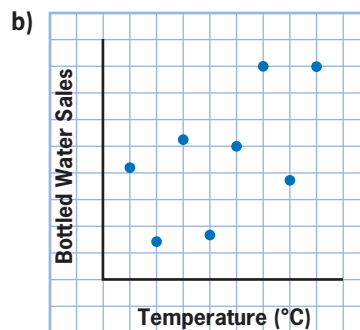
Example 1

Strength of Correlation

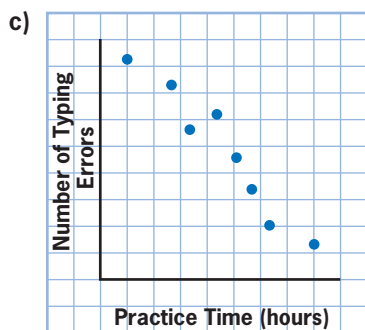
Do the graphs below show a linear correlation? Describe the correlation for each relationship.



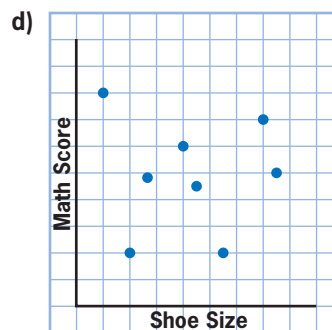
$$r = 1$$



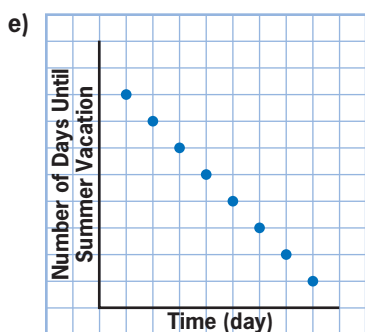
$$r = 0.6$$



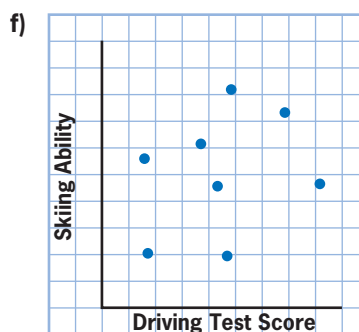
$$r = -0.96$$



$$r = 0.03$$



$$r = -1$$



$$r = 0.29$$

Solution

- As the mass of ham increases, the cost increases by a proportional amount. This is an example of a perfect positive correlation.
- As temperature increases, bottled water sales tend to increase; however, the trend is not obvious. The correlation coefficient of 0.6 suggests a correlation may exist.
- A clear pattern is evident here. As practice time increases, the number of typing errors decreases. The correlation coefficient of -0.96 indicates that this relationship has a strong negative linear correlation.
- There is little to indicate any linear pattern in this case. A correlation coefficient of 0.03 indicates there is definitely not a linear relation between math score and shoe size. The scatter plot does not appear to show any other type of relationship, so you can infer that math score is completely unrelated to shoe size.
- The number of days until summer vacation is decreasing over time at a proportional rate. This is an example of a perfect negative linear correlation.
- Careful inspection reveals a slight upward trend in the data. Looking from left to right, as driving test scores increase there is a slight tendency for skiing ability to increase. The correlation coefficient of 0.29 suggests a weak linear correlation.

Project Prep

Note the minor difference between the graphs in parts d) and f). If you encounter a weak linear correlation when researching your project, try to find data that offer stronger evidence before claiming a correlational relationship.

Your Turn

Sketch a scatter plot relating two variables that have:

- a strong positive correlation
- a moderate weak correlation
- no correlation

Indicate variables that could be correlated in this way for each case.

Example 2

Use Technology to Calculate the Correlation Coefficient

The table shows distance-time data for a student who is walking in front of a motion sensor.

d represents the distance between the walker and the motion sensor, in metres, after t seconds have passed.

Time, t (s)	Distance, d (m)
1	2.1
2	2.5
3	2.8
4	3.5
5	4.1

- Create a scatter plot relating distance, d , and time, t .
- Determine the strength of the linear correlation between these variables.
- Determine the equation of the line of best fit and explain what it means.

Solution

Method 1: Use a Graphing Calculator







Refer to the Prerequisite Skills on page 195 for help creating a scatter plot using a graphing calculator.

- Enter the data into a table.

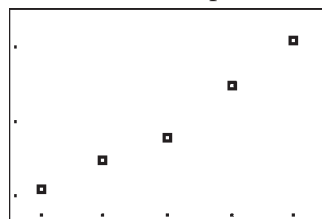
L1	L2	L3	2
1	2.1		
2	2.5		
3	2.8		
4	3.5		
5	4.1		

L2(6) =			

- Set the parameters as shown.

Plot1	Plot2	Plot3
On	Off	
Type: 		
Xlist: L1		
Ylist: L2		
Mark: 		

- Create a scatter plot of distance versus time.



- b) The scatter plot demonstrates a strong positive correlation between distance and time. Verify this by calculating the correlation coefficient.

- Press **STAT** and choose **CALC**.
- Select **4:LinReg(ax + b)**
- Press **2ND** **L1**, **2ND** **L2**, and then press **ENTER**.

The correlation coefficient, r , is approximately 0.99, confirming a near perfect positive linear correlation between distance and time.

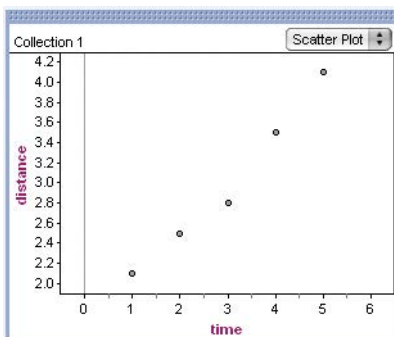
```
LinReg
y=ax+b
a=.5
b=1.5
r^2=.9765625
r=.9882117688
```

Method 2: Use Fathom™

- a) • Open Fathom™ and create a table as shown.

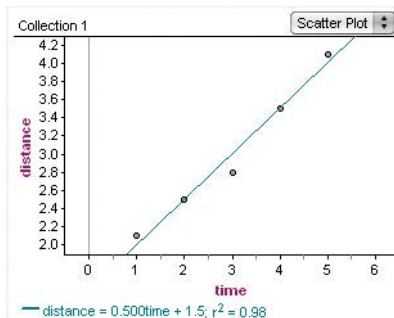
	time	distance
1	1	2.1
2	2	2.5
3	3	2.8
4	4	3.5
5	5	4.1

- Create a scatter plot of distance versus time.



- b) The scatter plot demonstrates a strong positive correlation between distance and time. Verify this by calculating the correlation coefficient. With the graph selected, choose the **Graph** menu and then choose **Least-Squares Line**.

The value of r^2 is 0.98. Take the square root to obtain $r = 0.99$. When you take the square root to determine the correlation coefficient, you must also analyse the graph to determine if it is positive or negative. The correlation coefficient, r , is approximately 0.99, confirming a strong positive linear correlation between distance and time.



Processes

Selecting Tools and Computational Strategies

If r does not appear, the Diagnostics may be turned off. To turn them on,

- press **2ND** **CATALOG**
- press **2ND** **ALPHA** **D** to jump to commands beginning with D
- select **Diagnostics On** and press **ENTER**.

Processes

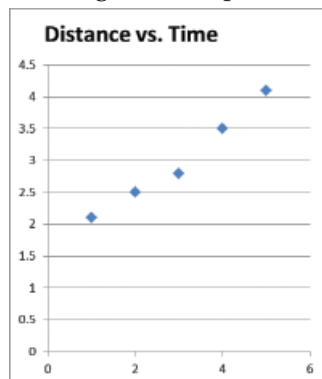
Connecting

Fathom™ and the TI-83/84+ graphing calculators both generate r^2 , which is called the coefficient of determination. While the correlation coefficient, r , measures how closely a set of data can be represented by a linear model, the coefficient of determination more generally measures how closely a set of data can be represented by either a linear model or a variety of non-linear models.

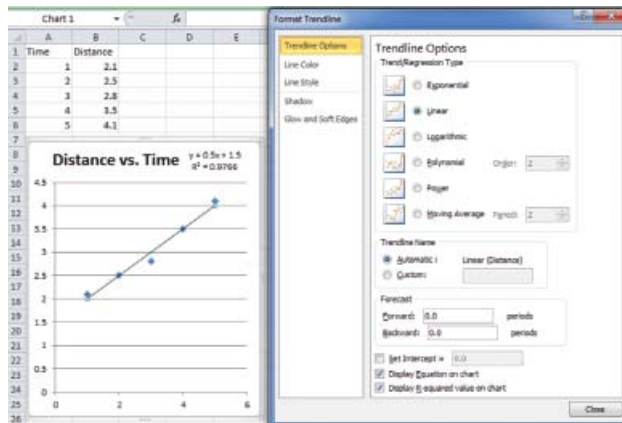
Method 3: Use a Spreadsheet

- a) • Open a new spreadsheet.
- Label two columns and enter the data as shown.
 - Create a scatter plot of distance versus time. Refer to the Prerequisite Skills on page 195 for help creating a scatter plot using a spreadsheet.

	A	B
1	Time	Distance
2	1	2.1
3	2	2.5
4	3	2.8
5	4	3.5
6	5	4.1



- b) The scatter plot demonstrates a strong positive correlation between distance and time. Verify this by calculating the correlation coefficient.
- Right click on one of the points and choose **Add Trendline**.
 - Choose **Linear**. Check the boxes to “Display Equation on chart” and “Display R-squared value on chart.” Click **Close**.



The value of r^2 is approximately 0.98. Take the square root to get $r = 0.99$.

The correlation coefficient, r , is approximately 0.99, confirming a strong positive linear correlation between distance and time.

- c) Each technology tool performed a **linear regression**. The equation of the line of best fit is $d = 0.5t + 1.5$, where d is the distance between the walker and the motion sensor after t seconds had passed. The fixed part of the equation, 1.5, gives the initial value of the walker's position. The variable part, $0.5t$, gives the rate of change. The equation shows that the walker started 1.5 m from the motion sensor and walked at a rate of 0.5 m/s away from the sensor.

Processes

Representing

The graphing calculator and spreadsheet used variables other than d and t . What variables were used in their place?

linear regression

- the formal process by which a line of best fit is mathematically determined

Your Turn

The table shows distance from home for a cyclist over time.

Time (min)	Distance (km)
10	9.8
20	8.1
30	5.8
40	4.2
50	2.3

- Create a scatter plot relating distance, d , and time, t .
- Determine the strength of linear correlation.
- Determine the equation of the line of best fit and explain what it means.
- Why do the actual data points not always fall exactly on the line?

Consolidate and Debrief

Key Concepts

- When a change in one variable is accompanied by a proportional change in another variable, the variables share a linear correlation.
- The correlation coefficient, r , is a measure of the strength of linear correlation between two variables. The value of r , which can be between -1 and 1 , gives an indication of how closely the data points relate to the line of best fit.
- The line of best fit can be used to model a linear correlation.
- Linear regression is the mathematical process that determines the line of best fit.

Reflect

R1. Two variables, X and Y , share a strong negative correlation.

- Sketch what a scatter plot of Y versus X could look like.
- Describe in words the correlation between X and Y .

R2. Repeat R1, assuming that X and Y have a moderate positive correlation.

R3. A student walking in front of a motion sensor generates distance-time data, where distance is in metres and time is in seconds. A linear regression on the data produces the information shown.

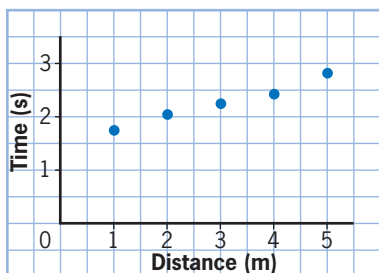
Describe everything you can about the motion of this walker and the relationship between distance and time.

```
LinReg
y=ax+b
a=.51
b=.49
r2=.9897260274
r=.9948497512
```

Practise

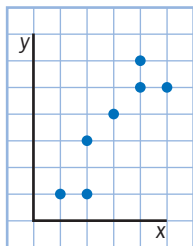
Choose the best answer for #1 to #3.

- Two variables have a linear correlation of -0.94 . Which of the following is true?
 - The variables share a strong, positive correlation.
 - The variables share a moderate, positive correlation.
 - The variables share a strong, negative correlation.
 - The variables share a moderate, negative correlation.
- The scatter plot shows the distance, d , in metres, and time, t , in seconds, for a student walking in front of a motion sensor.



Which equation represents the line of best fit?

- $d = -0.25t + 1.5$
 - $d = 0.25t + 1.5$
 - $d = 0.5t + 1.5$
 - $d = 1.5t + 0.25$
- The scatter plot shows the relationship between two variables, x and y . Which of the scenarios is most likely to have this relationship?
 - student's math score versus student's height
 - height above ground of a skydiver versus time
 - computer boot-up time versus unemployment rate
 - number of automobile accidents versus amount of snowfall



- Compare the linear regression data for two students walking in front of a motion sensor:

Bob's walk
 $d = 0.75t + 2$
 $r = 0.70$

Tracey's walk
 $d = 0.75t + 2$
 $r = 0.95$

- How are the movements of these walkers similar? different?
- Sketch possible distance versus time graphs for Bob and Tracey.

Apply

Use the table to answer #5 and #6.

Number of Books Owned	Hours per Week Watching TV	Hours per Week Using Internet
25	16	8
44	20	12
12	22	16
16	25	13
78	8	20
112	2	15
56	10	11

5. Application

- Create a scatter plot of time spent watching television versus number of books owned.
- Characterize the correlation.
- Perform a linear regression and record the correlation coefficient. Does this support your answer to part b)? Explain.

- Repeat the analysis performed in #5 for

- time spent on the Internet versus number of books owned
- time spent on the Internet versus time spent watching television

7. Thinking

- Use CANSIM or an alternate data source to find a time series that has a strong linear correlation.
- Create a scatter plot of the time series.
- Describe the correlation.

- d) Construct a line of best fit. Explain what the equation means.
- e) Why do you think the trend exists?

8. Communication In Fathom™, open the sample file found here: **Sample Documents/Statistics/Correlation and Regression/CorrelationPlay.ftm**. Drag the points to create each of the following. Then, describe or sketch the data pattern for each.

- a) $r^2 = 1$ for a positive trend
- b) $r^2 = 1$ for a negative trend
- c) a moderate positive linear correlation
- d) $r^2 = 0$

9. Open Question The result of a linear regression between two variables is:

$$y = -0.25x + 15 \quad r^2 = 0.75$$

- a) Sketch what the scatter plot of y versus x could look like.
- b) Explain the reasoning behind your sketch.

Achievement Check

10. Is there a linear correlation between body measures?

- a) Pick three or four body measures, such as height, foot length, head circumference, hand span, forearm length, and so on. Measure and record data for yourself and several classmates.
- b) Use technology such as Fathom™ or a spreadsheet to make two-variable comparisons using these strategies:
- scatter plot
 - correlation coefficient
 - line of best fit
- c) Identify two or three strong linear correlations. In each case:
- describe the nature of the correlation
 - identify and interpret the correlation coefficient
 - write the equation of the line of best fit and explain what it means

11. Thinking The table shows Marley's distance from home.

Time (h)	Distance (km)
0.0	280
0.5	230
1.0	210
1.5	150
2.2	140
2.5	90

- a) Create a scatter plot of distance versus time in Fathom™.
- b) From the **Graph** menu, choose **Add Movable Line**. Adjust the line so that it is close to being a line of best fit:
- Click and drag it near one of the endpoints to adjust the slope.
 - Click and drag it near the middle to translate the line vertically.
- c) From the **Graph** menu, choose **Show Squares**. What happens?
- d) Adjust the line so that the **Sum of Squares** value is as small as possible. Estimate when this occurs.
- e) From the **Graph** menu, choose **Least-Squares Line**. Test your estimate. Explain what you observe.

Extend

- 12.** The coefficient of determination, r^2 , is a measure of how well a regression curve fits a set of data. When Y is the dependent variable, and X is the independent variable, r^2 represents how much variance in Y can be explained by X. If the curve of best fit is a perfect fit, $r^2 = 1$. The closer r^2 gets to 0, the poorer the curve of best fit. State the values of r^2 in some of the questions in this section and interpret their meanings.
- 13.** Perform some research on the coefficient of determination, r^2 .
- a) How is it useful?
- b) What values can it have?
- c) How is it similar to the correlation coefficient? How is it different?

Cause and Effect

Learning Goals

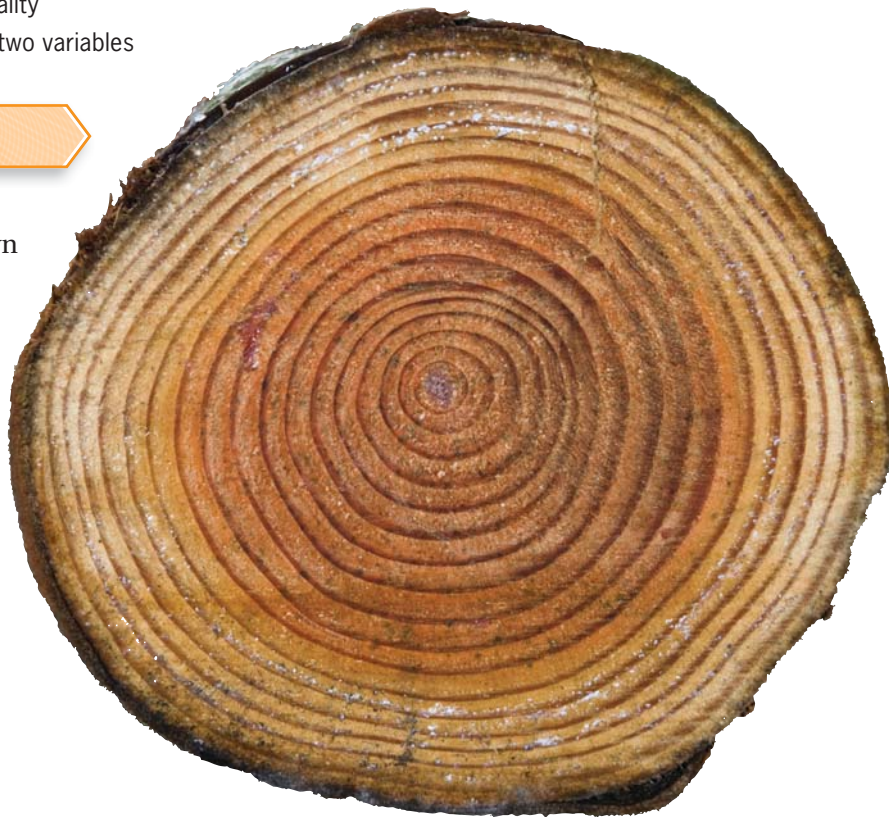
I am learning to

- distinguish between correlation and causality
- identify the type of relationship between two variables

Minds On...

When a lumberjack cuts down a tree, you can see its cross section, as shown in the image. Each ring in the cross section represents one year of the tree's growth.

- Do you think a tree grows by the same amount every year?
- What are some factors that might cause a tree to grow faster or slower?



Action!

Investigate Cause and Effect

Materials

- ruler
- graphing calculator or software

1. Measure the radius of the tree for each ring shown. Call the smallest ring Year 1. Record radius versus year in a table.

Year	Radius (cm)
1	
2	

2. a) Create a scatter plot of radius versus time.
b) Describe the correlation.

3.
 - a) Perform a linear regression.
 - b) What is the correlation coefficient? What does this suggest about the correlation?
4.
 - a) What is the equation of the line of best fit?
 - b) Interpret what this equation means.
5.
 - a) Use your linear model to predict the radius of the tree at an age of
 - 25 years
 - 50 years
 - b) Discuss any assumptions you must make and identify any factors that could affect the accuracy of your model.
6. **Reflect** Compare your linear models with those of your classmates. Are they identical? Are they close? Explain why there may be some inconsistency.
7. **Reflect** Identify the independent variable and the dependent variable in this study. Could they be reversed? Explain why or why not.
8. **Extend Your Understanding**
 - a) Calculate the cross-sectional area of the tree after each year.
 - b) Create a scatter plot of area versus time.
 - c) Does the relationship appear to have a strong linear correlation? Explain.
 - d) Use a spreadsheet or a graphing calculator to perform a quadratic regression. Comment on how well the curve that is generated fits the data. Why does this make sense?

Processes

Reflecting

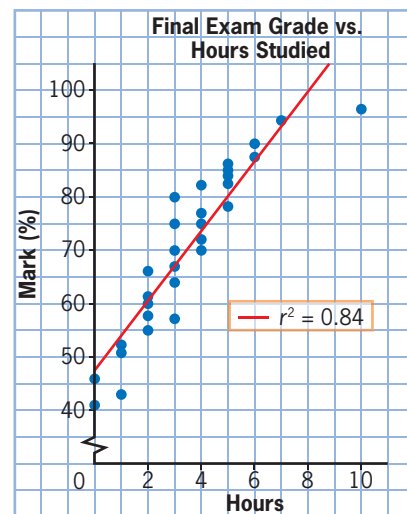
Think about the data collection process. Would everyone have exactly the same data? Why or why not?

Data analysis involves much more than fitting a line or a curve to a set of data points. Establishing a linear correlation between a dependent variable and an independent variable is just the first step in understanding the true nature of a relationship. Once you know there is a correlation, it is important to consider how and why such a correlation exists. Finding the meaning behind a linear correlation is what distinguishes true mathematical modelling from simply fitting a line to a set of data.

Example 1

Analyse a Cause and Effect Relationship

The scatter plot and line of best fit show the relationship between the mark achieved and number of hours studied for a grade 12 data management final exam.



cause and effect relationship

- the correlation between two variables in which a change in one directly causes a change in the other

- Does this correlation have a **cause and effect relationship**?
- Interpret the line of best fit.
- According to the linear model, for how many hours must a student study to achieve a perfect score of 100%? Comment on the validity of this answer.

Solution

- Calculate the correlation coefficient to determine the strength of linear correlation. Then, consider whether a causal relationship is reasonable.

$$r^2 = 0.84 \quad \text{Take the square root.}$$
$$r \approx 0.92$$

The correlation coefficient is about 0.92, suggesting a very strong correlation between mark and hours studied. Since most educational experts agree that strong study habits will result in higher achievement, and given the satisfactory sample size, it is reasonable to characterize this relationship as an example of cause and effect. The dependent variable is the student's mark, which is affected by the independent variable, number of hours studied.

- The value of the correlation coefficient suggests that the line of best fit for these data is relatively good for predicting student performance. The equation relating mark, m , to hours studied, h , is $m = 6.5h + 47$.

Generally, you can predict a student's mark by substituting the number of hours studied into h in the equation and then calculating m .

Determine the vertical intercept and rate of change to develop a deeper interpretation of this linear model. The vertical intercept occurs when h is set to 0:

$$m = 6.5(0) + 47$$
$$m = 47$$

The linear model predicts a failing grade of 47 if a student does not study at all.

The rate of change is 6.5, which means that a student's mark will increase by approximately six and a half percentage points for each additional hour studied, according to the linear model.

- c) The linear model can be applied to estimate the number of hours required to score a perfect exam. Substitute $m = 100$ and solve for h .

$$m = 6.5h + 47$$

$$100 = 6.5h + 47$$

$$100 - 47 = 6.5h$$

$$53 = 6.5h$$

$$\frac{53}{6.5} = h$$

$$8.2 = h$$

The linear model predicts that a score of 100% can be expected after a little more than eight hours of study. Care must be taken in making this prediction, however, particularly since one student who studied for 10 hours did not achieve a perfect exam. This example illustrates that linear models, while often useful, can have significant limitations.

Processes

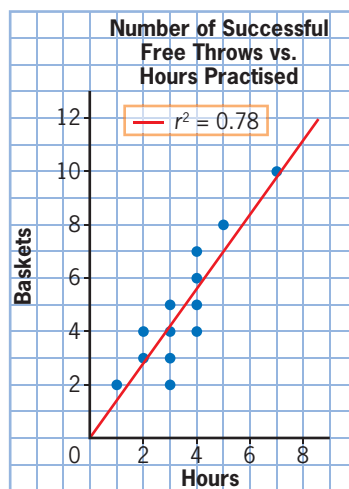
Reasoning and Proving

Does the result $h = 8.2$ suggest that a student who studies for nine or more hours can expect a final mark above 100%? Explain.

Your Turn

The scatter plot and line of best fit show the relationship between the number of successful free throws made out of 10 attempts and the number of hours spent practising for members of a basketball team.

- Characterize this correlation with regard to cause and effect.
- Interpret the line of best fit.
- Discuss any limitations of this linear model.



When analysing linear trends in data, it is important to distinguish between correlation and causality. Just because two variables share a strong linear correlation, it does not necessarily imply that a change in one variable is responsible for a change in the other. Inferring a cause and effect relation based strictly on correlational evidence is one of the most common errors in two-variable data analysis.

Sometimes two variables share a linear correlation because they both depend on another, third variable. This is known as a **common cause relationship**.

common cause relationship

- the correlation between two variables in which both variables change as a result of a third common variable

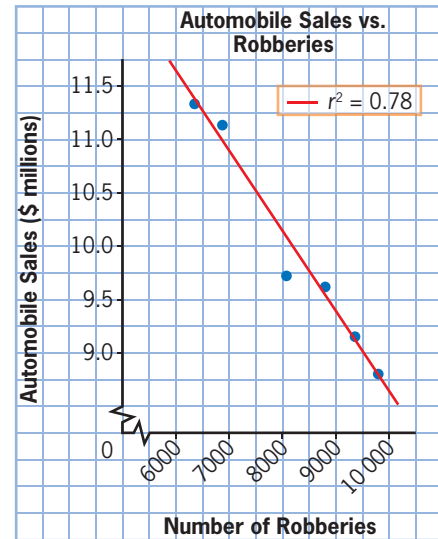
Example 2

Common Cause Relationships

The table and graph show a strong negative linear correlation between automobile sales and robberies from 1988 to 1993.

Year	Number of Robberies	Automobile Sales (\$)
1988	6375	11 335 615
1989	6899	11 140 918
1990	8101	9 736 777
1991	9823	8 808 249
1992	9370	9 156 456
1993	8828	9 623 595

Source: CANSIM Table 079-0003, New motor vehicle sales, Canada, provinces and territories, Statistics Canada, April 11, 2014; CANSIM Table 252-0001, Crimes, by actual offences, Statistics Canada, November 15, 2001



Is it likely that an increase in robberies would cause car sales to drop? Identify a possible common cause for these trends.

Solution

To conclude from these data that an increase in robberies could be responsible for a decrease in automobile sales does not make much sense. It is more likely that a common cause is involved.

What could cause both of these trends? When economic times are poor, families may be more likely to delay a major purchase such as an automobile. It is possible that robberies may be more likely to increase when people lose their jobs and become desperate to take care of their families. Is it possible that both of these trends are caused by high unemployment?

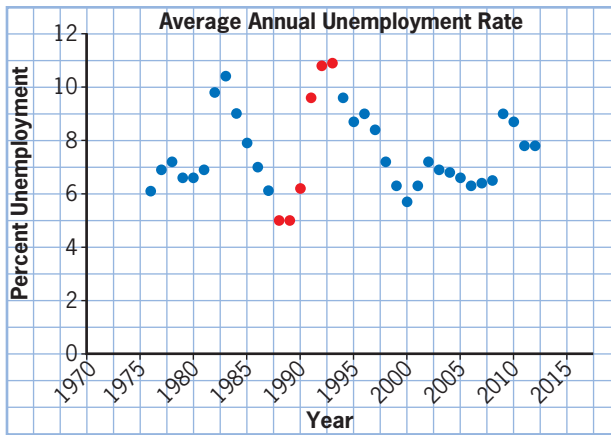
The table and graph show the average annual unemployment rate in Ontario.

Year	Percent Unemployment	Year	Percent Unemployment	Year	Percent Unemployment
1976	6.1	1988	5.0	2000	5.7
1977	6.9	1989	5.0	2001	6.3
1978	7.2	1990	6.2	2002	7.2
1979	6.6	1991	9.5	2003	6.9
1980	6.6	1992	10.8	2004	6.8
1981	6.9	1993	10.9	2005	6.6
1982	9.8	1994	9.6	2006	6.3
1983	10.4	1995	8.7	2007	6.4
1984	9.0	1996	9.0	2008	6.5
1985	7.9	1997	8.4	2009	9.0
1986	7.0	1998	7.2	2010	8.7
1987	6.1	1999	6.3	2011	7.8
				2012	7.8

Source: Labour Force Survey, Annual Average Unemployment Rate, Statistics Canada

Literacy Link

In economic terms, a *recession* refers to a poor time for business. Signs of a recession include low investments, low profits, and high unemployment. A *boom* is the opposite of a recession.



The unemployment rate spiked to a very high level during this same time period. It is quite possible that decreased automobile sales and increased counts of robbery could both be a result of a recession, or weak economy.

Your Turn

While performing research for his data management project, Aidan discovers a strong linear correlation between the number of forest fires per year and the yield of tomato harvest for the same region. He wonders if growing more tomatoes causes more forest fires. Do you agree with Aidan's line of reasoning? If so, explain why. If not, offer a more likely explanation for this correlation.

It is important not to jump to a conclusion too quickly when deciding on the type of relationship that exists between two variables. In the example above, a cause and effect relationship was ruled out because it did not make sense. A reasonable case was made for a common cause relationship; however, this still would not necessarily constitute proof.

Variables can be related in different ways. A **presumed relationship** can exist when it seems to make sense that two variables are related, and yet no causality can be inferred and it is also difficult to identify a clear common cause factor. An example of this would be a positive correlation between the number of books in children's homes and their math scores.

A **reverse cause and effect relationship** occurs when the assumed causation becomes reversed. Consider the positive correlation between severe illness and depression. A researcher might hypothesize that being severely ill is very emotionally difficult and so it causes depression. However, in reality, depressed people struggle to take care of themselves, so they are more likely to become severely ill.

Sometimes two variables share a strong linear correlation for no logical reason at all. This is called an **accidental relationship**. Suppose that a positive correlation was found between the local kitten birth rate and the price of eggs. It would be sensible to consider such a correlation purely coincidental.

Project Prep

The data set in Example 2 was relatively small. To build a stronger case of automobile sales dropping and robberies increasing as a direct cause of high unemployment, more data from other recession periods could be studied. If similar results were replicated many times over, then a stronger argument could be made for a common cause relationship. When collecting data for your project, consider whether the size of the data set is sufficient to support your analysis.

presumed relationship

- a relationship that makes sense but does not seem to have a causation factor

reverse cause and effect relationship

- a relationship in which the independent and dependent variable are reversed

accidental relationship

- a relationship that is based purely on coincidence

Example 3

Identifying Types of Relationships

Suggest the most likely type of relationship for each correlation.

- a) The number of fire stations in a city is positively correlated with the number of parks.
- b) The price of butter is positively correlated with fish population levels.
- c) Seat belt infractions are positively correlated with traffic fatalities.
- d) Self-esteem is positively correlated with vocabulary level.
- e) Charged crimes is positively correlated with the size of the police force.

Solution

- a) It does not seem to make sense that building more fire stations would cause an increase in the number of parks in a city. Both of these variables likely share a positive correlation with city population. This is most likely a common cause relationship.
- b) No clear connection exists between the price of butter and fish population. When two variables share a strong linear correlation for no logical reason at all, it is an accidental relationship.
- c) Seat belts are specifically designed to save lives in the event of an accident. A strong case can be made that this is a cause and effect relationship.
- d) It makes sense that someone with strong language skills would tend to also have higher self-esteem. It would be difficult to suggest, however, that one causes the other, and it would also be challenging to identify a clear common cause factor. This is most likely a presumed relationship.
- e) One hypothesis might be that when crimes increase, a police department responds by recruiting more officers. It could also be argued that as a police force is increased, there are more opportunities to catch and charge criminals. A reverse cause and effect relationship occurs when the assumed causation becomes reversed.

Your Turn

Classify the relationships and justify your choice in each case.

- a) A patient's stress level is negatively correlated with the amount of exercise performed.
- b) Student math scores are positively correlated with English scores.
- c) Pancake sales are negatively correlated with amount of rainfall.
- d) Job interview success rate is positively correlated with number of years a person has been married.

Project Prep

When searching for data for your project, you may discover a linear correlation between two variables. Could the correlation be explained by the presence of a common cause variable? Could a case be made for a cause and effect relationship? If no logical reason for the correlation can be inferred, the relationship may have to be dismissed as accidental. To answer these questions you will likely need to dig more deeply into the data.

Key Concepts

- A cause and effect relationship exists when one variable is directly responsible for a change in another variable.
- If two variables share a strong correlation, it does not imply that a cause and effect relationship exists.
- A common cause relationship exists when a common third variable is responsible for the correlation between two other variables.
- Several types of relationships can exist between two variables, including cause and effect, common cause, presumed, reverse cause and effect, and accidental.

Reflect

- R1.** Explain why correlational evidence alone does not imply a cause and effect relationship.
- R2.** A positive correlation was discovered with each pair of the following variables:
- cases of the flu
 - amount of severe winter weather
 - tissue sales
- What is the most likely type of relationship involved here? Explain.
- R3.** A study found that caffeine intake is positively correlated with nervousness.
- a) Suggest a cause and effect relationship that could explain this correlation.
 - b) Suggest a reverse cause and effect relationship that could be argued.
- R4.** a) Make up an accidental relationship between two variables.
b) Explain why the correlation is likely to be accidental.

Practise

Choose the best answer for #1 to #3.

1. As variable x increases, variable y decreases proportionately. Which of the following statements is definitely true?
 - A This is a cause and effect relationship in which x is the dependent variable.
 - B This is a cause and effect relationship in which y is the dependent variable.
 - C These variables share a positive correlation.
 - D These variables share a negative correlation.
2. Patients who participate in a new exercise program experience a drop in blood pressure over the same period of time. Which statement is most likely correct?
 - A This is a cause and effect relationship, in which the dependent variable is amount of exercise.
 - B This is a cause and effect relationship, in which the dependent variable is blood pressure.
 - C This is an accidental relationship.
 - D This is a presumed relationship.

3. In an economic study, average salaries were negatively correlated with the unemployment rate. Which of the following is most likely to be a common cause factor that accounts for this correlation?
 - A the price of eggs
 - B the strength of the economy
 - C the current birth rate
 - D movie industry revenue
4. At a ski resort, lift ticket sales were positively correlated with hot chocolate sales.
 - a) Is this likely a cause and effect relationship? Explain why or why not.
 - b) Suggest a common cause that could explain this relationship.

Apply

5. a) At a dance competition, how would you expect the relationship to look between dance performance score and number of hours practised? Sketch a graph to support your answer.
b) Do you think this is a cause and effect relationship? Explain.
6. Identify the most likely type of relationship between the two variables for each scenario. Assume the independent variable is mentioned first. Justify your answers.
 - a) Grass growth is positively correlated with amount of rainfall.
 - b) Arm length is positively correlated with leg length.
 - c) Sandwich sales are negatively correlated with dog bite incidents.
 - d) Interest in televised sport is positively correlated with fitness level.
 - e) Incidence of diabetes is negatively correlated with healthy eating habits.
 - f) Heart disease is positively correlated with lung cancer.
7. A researcher wonders if people who do not get enough sleep also eat a lot of fast food.
 - a) Explain why this is unlikely to be a cause and effect relationship.
 - b) Suggest a common cause that could explain this correlation.
8. **Communication** A student discovers that ice-cream sales are positively correlated with occurrences of heat stroke. He suggests that ice-cream consumption could be a cause of heat stroke.
 - a) Do you agree with the student? Why or why not?
 - b) What advice can you offer to improve his analysis?
9. **Application** The number of deer in a region is positively correlated with the number of wolves.
 - a) Explain how this could be a cause and effect relationship with the number of deer as the dependent variable.
 - b) Explain how this could also be described as a reverse cause and effect relationship.



Achievement Check

10. The table shows time series data for the population of two Ontario towns.

Year	Collingwood	Grimsby
1991	14 382	18 520
1996	15 596	19 585
2001	16 039	21 297
2006	17 290	23 937
2011	19 241	25 325

Source: Cities and Towns Table, City Populations

- a) Create a scatter plot of population of Collingwood versus population of Grimsby. Describe the correlation.
- b) Perform a linear regression. Interpret the equation of the line of best fit.
- c) Do you think this is a cause and effect relationship? Explain your thinking.
- d) What type of relationship do you think this is? Justify your answer with mathematical reasoning.

Use this information to answer #11 to #13.

The table shows the supply and demand for widgets at various selling price points. The demand represents the number of widgets expected to sell at a certain price. The supply represents the number that can be produced at a certain price.

Price (\$)	Quantity Demanded	Quantity Supplied
10	18	8
11	17	9
12	15	11
13	14	12
14	12	14
15	10	16
16	9	17

11. Application

- Create a scatter plot of price versus widget quantity demanded. Describe the correlation.

Processes

Representing

In economics, the independent variable is usually plotted on the vertical axis and the dependent variable on the horizontal axis, contrary to common mathematical convention.

- Is this likely a cause and effect relationship? Explain.
 - Identify the independent and dependent variables. Explain your thinking.
12. a) Create a scatter plot of price versus widget quantity supplied. Describe the correlation.
- Is this likely a cause and effect relationship? Explain.
 - Identify the independent and dependent variables. Explain your thinking.

13. Thinking

- Perform a linear regression for the graphs in #11 and #12, and plot both functions on the same grid.
- Identify the point of intersection. Explain what it signifies.
- What will likely happen if the widget price is set
 - above the intersection point?
 - below the intersection point?
 Explain your thinking.

Extend

14. The table shows a time series for the population of Brampton.

Year	1991	1996	2001	2006	2011
Population	234 445	268 251	325 428	433 806	523 911

Source: Cities and Towns Table, City Populations

- Create a scatter plot of population versus time using technology.
- Does the correlation appear to be linear? Explain.
- Perform a linear regression. Describe the goodness of fit.
- Perform an exponential regression on the data. Describe the shape of the curve that appears.
- Compare the goodness of fit of the curve of best fit to that of the line of best fit.
- Why might an exponential model be appropriate for this relationship?

Literacy Link

An exponential regression is the process of fitting an exponential curve of best fit to a set of data. To perform an exponential regression, use technology to carry out the same steps as a linear regression, but choose exponential regression instead of linear regression.

15. Investigate the population of a city or town in Ontario that is of interest to you. Graph the time series and compare linear and exponential regression models. Decide which is appropriate and justify your choice using mathematical reasoning.

Dynamic Analysis of Two-Variable Data

Learning Goals

I am learning to

- identify outliers and account for their impact on a data trend
- recognize the presence of extraneous variables
- identify a hidden variable and account for its impact on a correlation

Minds On...

Suppose two athletes are competing for a position on the track and field team in the long jump event. The chart shows their tryout distances, in metres.

	Jump 1	Jump 2	Jump 3	Jump 4	Jump 5	Mean
Hank	4.5	4.0	4.3	4.2	4.0	4.2
Vito	4.9	5.1	4.8	1.2	4.7	4.1

- Which athlete do you think should make the team?
- Could you make an argument for either athlete?
- Is there anything that seems unusual in the data?
- Would more information be helpful?



Action!

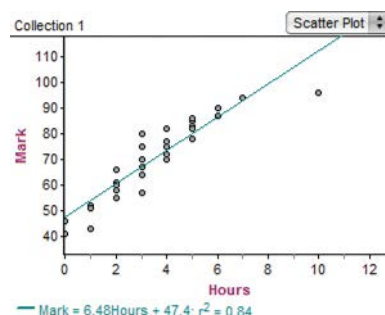
Investigate A Residual Plot

Materials

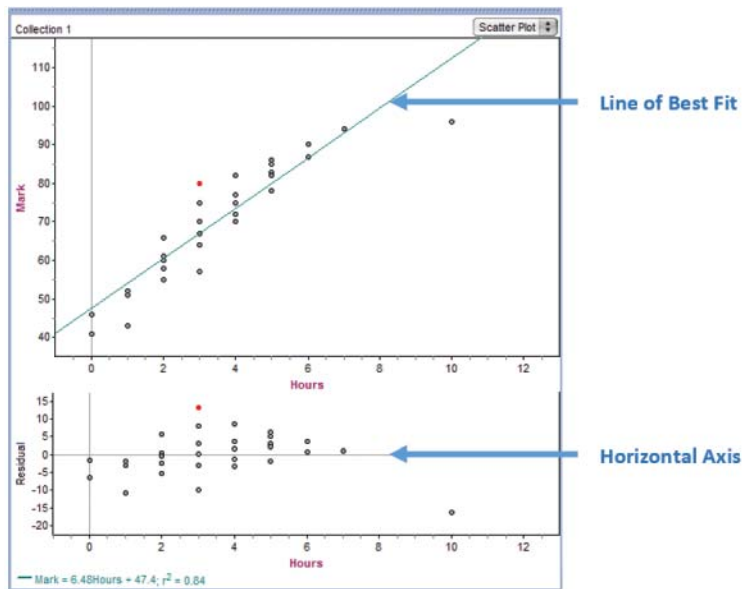
- computer with Fathom™ software

In the previous section, you compared the number of hours studied and final exam grades using a scatter plot and line of best fit. In this activity you will use that data set to perform a deeper analysis of the relationship between the line of best fit and the data that the line is fitted to.

- Your teacher will provide you with a Fathom™ file called **InvestigateAResidualPlot.ftm**.
 - Examine the scatter plot and line of best fit:
 - Click on a point to read its coordinates.
 - Repeat for a new point.



2. a) Identify the coordinates of some points that are right on the line or very close to it.
b) How good a predictor of performance is the linear model for these students? Explain your reasoning.
3. a) What are the coordinates of some points that are relatively far away from the line?
b) Is the linear model a good predictor for these students? Explain.
4. a) Create a **residual plot**. With the scatter plot selected, choose **Graph** and then **Residual Plot**.
b) Enlarge the graph by clicking and dragging one of the corners of the window frame.
c) Click on a point in the scatter plot and identify its corresponding point in the residual plot. The point (3, 80) is selected below:



residual plot

- shows the value of each **residual** graphically as the vertical distance from a horizontal axis

residual

- the difference between a data point's actual dependent value and the dependent value predicted by the line of best fit

5. a) Click on a point in the scatter plot. Is it above or below the line of best fit? Is it close to or far from it?
b) Identify the corresponding point in the residual plot. Is it above or below the horizontal axis? Is it close to or far from it?
c) Compare your answers for a) and b).
d) Repeat for several other points.
6. **Reflect** How does a residual plot appear to be related to a scatter plot?
7. **Extend Your Understanding** Calculate the difference between the actual exam score and the score predicted by the linear model for several students. Explain how this difference is represented in the residual plot.

Example 1

Construct and Analyse a Residual Plot

Will an increase in recycling result in a reduction of landfill? The table compares the mass of garbage and recycling for a town during a recycling campaign.

Amount Recycled (kg)	Amount of Garbage (kg)
120	200
144	175
160	190
175	156
200	142
210	167
224	140
236	150

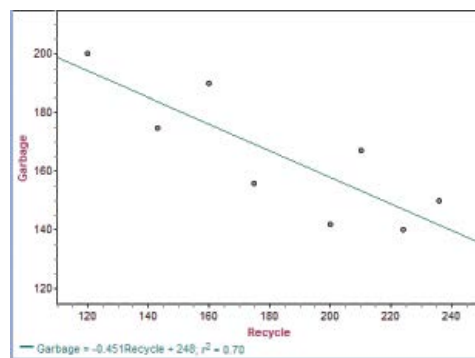
- Create a scatter plot and perform a linear regression. Describe the trend. Is a linear model reasonable in this case?
- Interpret the residuals for (120, 200) and (200, 142).
- Construct a residual plot. Describe the pattern.

Solution

- The scatter plot and linear regression show that as the mass of recycling increases, the mass of garbage decreases. The correlation coefficient of

$$r = -\sqrt{0.70} \\ \approx -0.84$$

confirms a strong negative correlation. A linear model is appropriate for describing this relationship.



- The residual of (120, 200) is the difference between the actual mass of garbage (200 kg) and the amount predicted by the linear model. Use the equation of the line of best fit to calculate the amount predicted by the linear model. Let m represent the mass of recycled material and g the amount of garbage, both in kilograms.

$$\begin{aligned} g &= -0.451m + 248 \quad \text{Substitute } m = 120. \\ &= -0.451(120) + 248 \\ &= 193.88 \end{aligned}$$

The linear model predicts that approximately 194 kg of garbage would be produced when the recycling amount is 120 kg. Subtract this value from the actual amount of garbage produced to determine the residual.

$$200 - 194 = 6$$

The residual for this datum is 6, which means that the actual mass of recycled material is 6 kg higher than what the linear model predicts.

Repeat this process for the datum (200, 142).

$$\begin{aligned} g &= -0.451(200) + 248 \\ &= 157.8 \end{aligned}$$

Processes

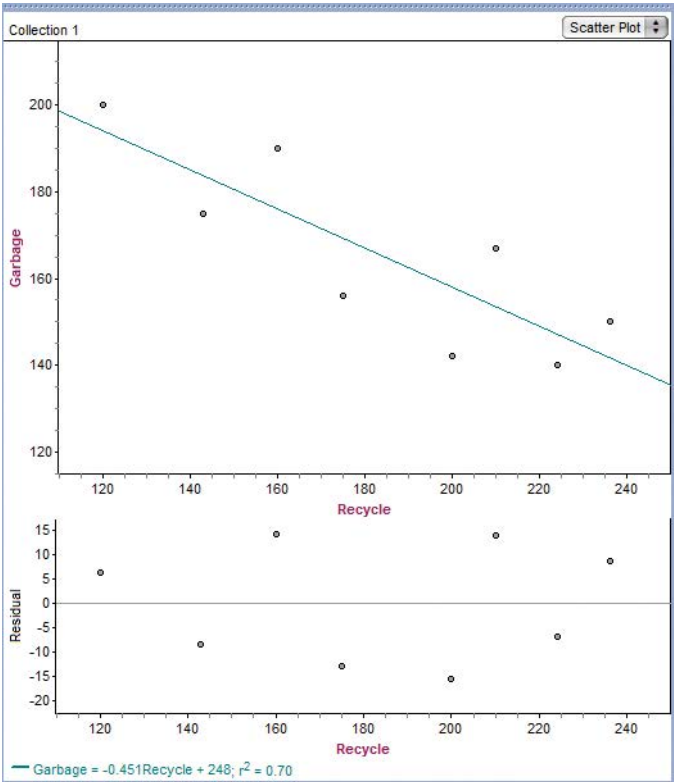
Reasoning and Proving

How do you know to include a negative sign with the correlation coefficient?

The actual value is 142. The linear model predicts approximately 158. So, the residual is $142 - 158 = -16$.

The residual for this datum is -16 , which means that the actual mass of recycled material is 16 kg lower than what the linear model predicts.

- c) To construct a residual plot using Fathom™, select the scatter plot and choose **Graph** and then **Make Residual Plot**.



There is no clear pattern in the residual plot. Points seem to be randomly located above and below the horizontal axis. This is normal for a good linear model because there should be roughly the same number of data points above as below the line of best fit.

Your Turn

The table compares a jewellery store's weekly diamond ring sales to the minutes of radio advertising purchased for the same week.

- a) Create a scatter plot and perform a linear regression. Describe the trend. Is a linear model reasonable in this case?
- b) Interpret the residuals for (20, 7.9) and (18, 4.7).
- c) Construct a residual plot. Describe the pattern in it.

Advertising (min)	Diamond Ring Sales (\$1000s)
7	1.6
12	4.4
20	7.9
32	10.6
22	6.0
18	4.7
26	7.6

A residual plot can be helpful for deciding how well a linear model fits a set of data. If there is a pattern or irregularity in a residual plot, the linear model may have to be re-evaluated.

Example 2

Account for Outliers

The table shows repair costs over the course of a year for several cars of the same model.

Age of Car (years)	Repair Costs (\$1000s)
4	0.6
4	0.8
4	3.2
5	1.1
6	1.2
6	1.5
7	1.4
8	1.8
9	1.9
10	2.1

- Create a scatter plot of repair costs versus age.
- Perform a linear regression and discuss the goodness of fit.
- Construct a residual plot and identify any outliers.
- Repeat the regression with the outlier removed.
- Compare the two linear models.

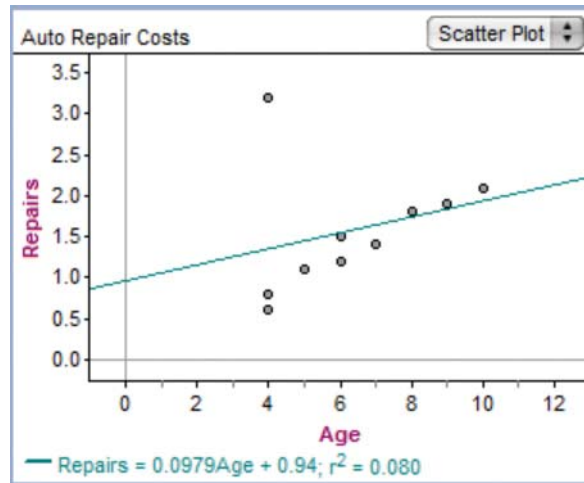
outlier

- a data point that does not fit an otherwise clear trend
- in a scatter plot, the outlier is relatively far from the line of best fit
- in a residual plot, the outlier is either relatively far above or below the horizontal line

Solution

Method 1: Use Fathom™

- Import or enter the data into Fathom™. Select the case table and drag a **New Graph** into the workspace. Drag the **Age** attribute onto the horizontal axis and the **Repairs** attribute onto the vertical axis.
- With the graph selected, choose **Graph** and then **Least-Squares Line**.



This line of best fit does not appear to serve as a very good model for the data. The relatively high extrapolated vertical intercept suggests around a thousand dollars of repairs for a brand-new vehicle, which does not sound right.

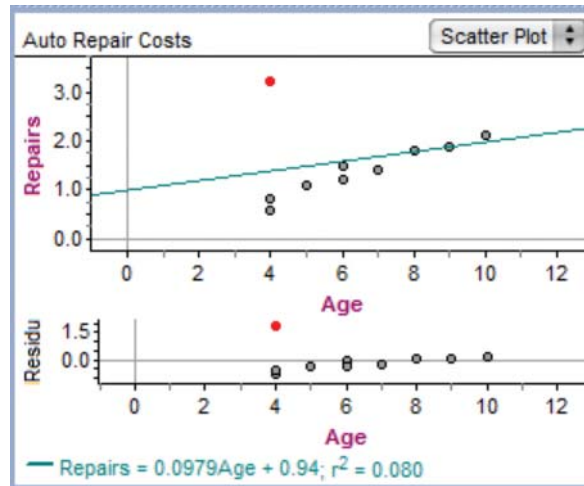
The correlational coefficient,

$$r = \sqrt{0.08}$$

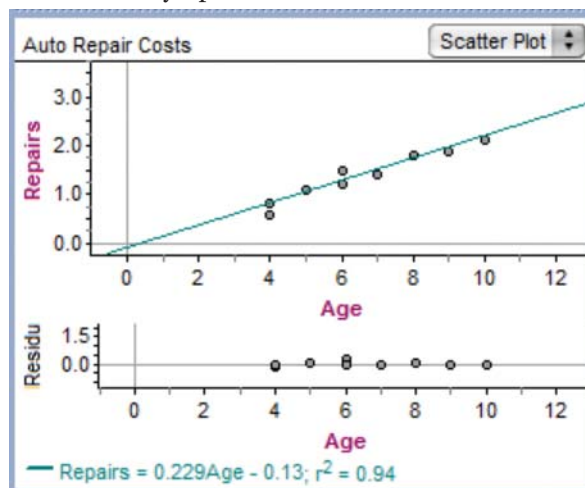
$$\approx 0.28,$$

suggests a weak linear correlation, but, except for one point, the data appear to form a strong linear trend. Could the point (4, 3200) be an outlier? Could it be responsible for the construction of a weak linear model?

- c) Construct a residual plot and click on the suspected outlier. The residual of (4, 3200) is much farther from the residual line than the other residuals, which appear to have a pattern. This suggests that the outlier has a strong influence on the linear model.



- d) Repeat the analysis with the outlier removed. With the point (4, 3200) selected, choose **Edit** and then **Cut Case**. The linear model will be automatically updated.



The new line of best fit fits the remaining data very well. The predicted initial repair cost is nearly zero. The correlational coefficient,

$$r = \sqrt{0.94}$$

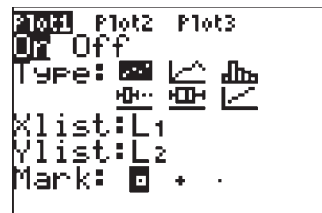
$$= 0.97,$$

suggests a very strong linear correlation. There is no obvious pattern to the residuals and none appear overly far from the residual line. This appears to be a strong linear model for predicting repair costs for this type of vehicle.

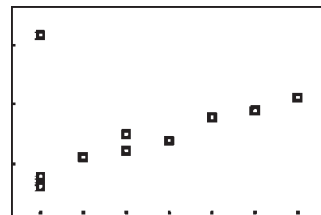
Method 2: Use a Graphing Calculator

a) Create a scatter plot of repair costs versus age. Refer to the Prerequisite Skills on page 195 for help.

- Enter the data in the lists.
- Set up the parameters as shown.

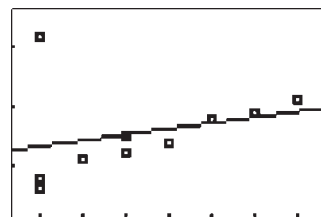


- The scatter plot will appear.

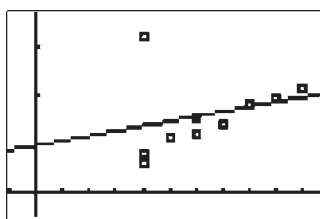


b) Perform a linear regression and store it as function Y1:

- Press **2ND** **QUIT** to go to the home screen.
- Press **STAT**. Then, choose **CALC** and **4:LinReg(ax+b)**.
- Press **2ND** **L1**, **2ND** **L2**, **VARΣ**. Choose **Y-VARS**. Then, choose **1:Function** and then **Y1**. Press **ENTER**.
- Press **GRAPH** to see the line of best fit.



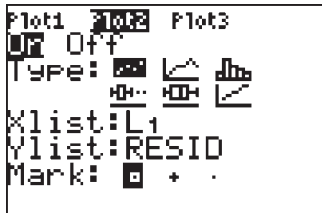
Press **ZOOM** and choose **3:Zoom Out** to see where the line of best fit crosses the vertical axis.



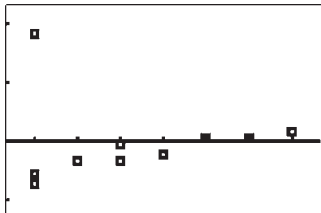
This line of best fit does not appear to serve as a very good model for the data. The relatively high extrapolated vertical intercept suggests around a thousand dollars of repairs for a brand-new vehicle, which does not sound right. The correlational coefficient, $r = 0.28$, suggests a weak linear correlation, but, except for one point, the data appear to form a strong linear trend. Could the point (4, 3200) be an outlier? Could it be responsible for the construction of a weak linear model?

c) Construct a residual plot and look for the suspected outlier:

- Press **STAT** and choose **1:Edit**.
- Move to the top of L3. Press **2ND** **INS**. Type **RESID** and press **ENTER**.
- Press **2ND** **STAT PLOT**. Turn Plot1 off and turn Plot2 on.
- Set the parameters as shown.



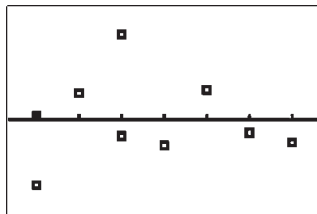
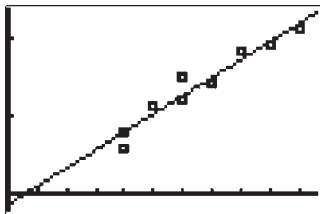
- Press **ZOOM** and choose **9:ZoomStat**.



The residual of (4, 3200) is much farther from the residual line than the other residuals which appear to have a pattern. This suggests that the outlier has a strong influence on the linear model.

d) Repeat the analysis with the outlier removed.

- Press **STAT** and choose **1:Edit**.
- Delete the table entries corresponding to (4, 3200).
- Repeat the linear regression analysis.



The new line of best fit fits the remaining data very well. The predicted initial repair cost is nearly zero. The correlational coefficient, $r = 0.97$, suggests a very strong linear correlation. There is no obvious pattern to the residuals and none appear overly far from the residual line. This appears to be a strong linear model for predicting repair costs for this type of vehicle.

Processes

Selecting Tools and Computational Strategies

You will need to use the **ALPHA** key to access the letter **RESID**.

Literacy Link

In the automotive industry, a *lemon* refers to a defective vehicle that requires many repairs.

Processes

Reasoning and Proving

If you identify an outlier in a set of data, consider why it may be present and the impact it has on the trend. It may be helpful to analyse the data both with and without the outlier present.

- e) The original linear model was heavily influenced by the presence of an outlying point: a four-year-old car needing \$3200 in repairs in one year. This seems quite high compared to the other data; however, there may be other factors involved.

- Was the car involved in a serious collision?
- Did the owner neglect to perform preventive maintenance?
- Was the vehicle just a lemon?

After the outlier was removed from the analysis, the linear model appeared to be quite strong, suggesting that it was the superior model for predicting repair costs for this type of vehicle. The size of the data set, however, is quite small in terms of statistical reliability. A much larger sample size not only would reduce the likelihood of sample bias, but also would lessen the influence of a single outlier.

Your Turn

The table shows the sale price for several used motorcycles of the same model and their age.

Age (years)	Price (\$1000s)
1	15
2	13
2	12
3	2
3	10
4	8
4	7.5
4	7
5	6.5
5	6
6	5
7	4

- Create a scatter plot of sale price versus age.
- Perform a linear regression and discuss the goodness of fit.
- Construct a residual plot and identify any outliers. Suggest reasons why any outliers may exist.
- Repeat the regression with the outlier removed.
- Compare the two linear models.

Example 3

Account for a Hidden Variable

Is absenteeism among professional workers on the rise in Canada? Are there similar trends for males and females? The contingency table shows the average number of absences for all Canadian males and females who have at least one university degree. Perform dynamic statistical analysis to determine any absenteeism trends.

Literacy Link

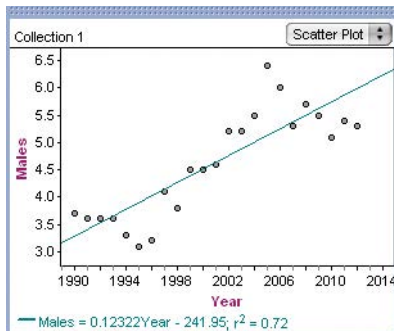
A contingency table subdivides data into two or more categories.

Absences					
Year	Males	Females	Year	Males	Females
1990	3.7	11	2002	5.2	8.6
1991	3.6	11.1	2003	5.2	8.5
1992	3.6	10.3	2004	5.5	8.7
1993	3.6	11.3	2005	6.4	8.8
1994	3.3	11.5	2006	6.0	9.4
1995	3.1	12.1	2007	5.3	9.9
1996	3.2	12.0	2008	5.7	9.2
1997	4.1	7.7	2009	5.5	9.3
1998	3.8	7.6	2010	5.1	8.8
1999	4.5	7.0	2011	5.4	9.5
2000	4.5	6.8	2012	5.3	8.7
2001	4.6	7.7			

Source: CANSIM Table 279-0036, Absence rates of full-time employees, by sex and education, Canada, Statistics Canada, September 18, 2013

Solution

- In Fathom™, open a new **Collection** and paste the cases. Drag a **New Graph** into the workspace and plot **Males** versus **Year**. From the **Graph** menu, choose **Least-Squares Line**.



There appears to be an upward trend between male absenteeism and year, with a correlation coefficient of $\sqrt{0.72} \approx 0.85$. Male absenteeism appears to be strongly correlated with time. Professional male worker absenteeism is increasing at a rate of approximately 0.12 days per year.

Processes

Reflecting and Reasoning and Proving

When a data trend appears fragmented, what can you do to try to explain the difference? How can you discover if a hidden variable is present?

hidden variable

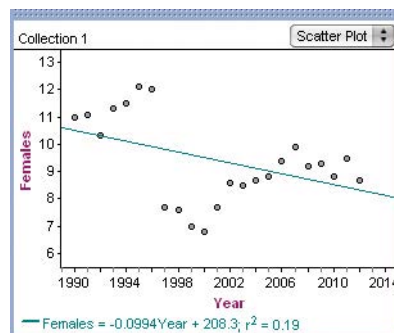
- a variable that affects or obscures the relationship between two other variables
- can sometimes result in a false correlation or a fragmented trend

Project Prep

When searching for data for your project, be watchful for irregularities in data trends. Consider the presence of outliers or hidden variables and take them into account when performing your analysis.

- Drag a **New Graph** into the workspace and plot **Females** versus **Year**. Then, construct a line of best fit. Examine the data for females.

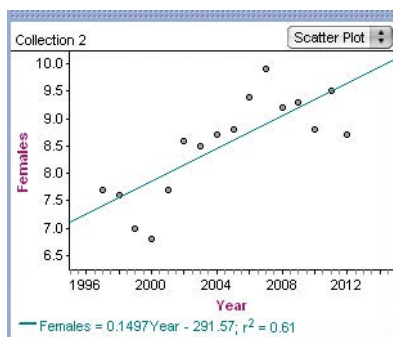
There appears to be a slight downward trend between female absenteeism and year. The correlation coefficient of $-\sqrt{0.19} \approx -0.44$ suggests, however, that this correlation is barely moderate. A closer inspection of this graph reveals two distinct trends, between 1990 and 1996 and between 1997 and 2012. There appear to be two distinct positive trends. What could possibly account for this strange pattern? The following excerpt was taken from the data source:



Footnotes:

1. Data from 1987 to 1996 include maternity leave. Also, men using paid paternity (in Quebec only) and parental leave are included in the calculation till 2006.

Maternity leave is included in absence tallies until 1996 only. This **hidden variable** could account for the fragmented pattern in the data. Remove the data points from 1990 to 1996 and repeat the analysis.



Now the trend for females more closely resembles the male data. A correlational coefficient value of $\sqrt{0.61} \approx 0.78$ confirms that female absence is strongly correlated with time. Absenteeism among female professional workers is increasing at a rate of approximately 0.15 days per year.

The data in this analysis suggest that absenteeism among Canadian professionals is on the rise. A desire to identify possible causes for this could prompt further study.

Your Turn

Do you think men using paid paternity leave in Quebec until 2006 is a significant hidden variable in this analysis? Why or why not?

Consolidate and Debrief

Key Concepts

- A residual is the difference between the actual dependent value of a datum and the value predicted by a line of best fit.
- A residual plot shows how close each data point is to a line of best fit.
- An outlier is a data point that does not fit well in an otherwise linear trend.
- An outlier can have a strong impact on a linear regression model if the number of data points is relatively small.
- A hidden variable can distort or obscure a linear correlation between two other variables.
- It is important to consider the impact of outliers and hidden variables when conducting a correlational study. It may help to remove or account for them when analysing the data.

Reflect

R1. Explain what a residual plot shows for a set of two-variable data. Draw a sketch to support your answer.

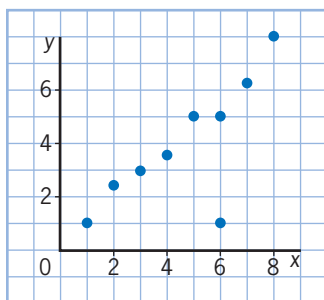
R2. a) What is an outlier?

b) Describe two ways of identifying an outlier in a set of two-variable data.

Practise

Choose the best answer for #1 to #3.

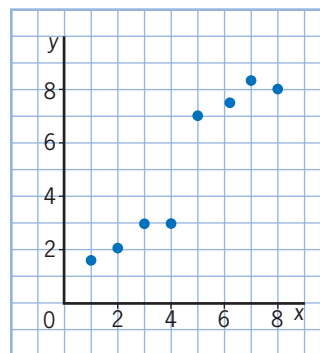
1. Consider the correlation.



Which statement is most accurate?

- A** There is a strong positive correlation.
- B** There is a moderate positive correlation.
- C** There is a strong positive correlation with an outlier.
- D** There is a strong positive correlation with a possible hidden variable.

2. Consider the correlation.



Which statement is most accurate?

- A** There is a strong positive correlation.
- B** There is a moderate positive correlation.
- C** There is a strong positive correlation if the outlier is disregarded.
- D** There is a strong positive correlation with a hidden variable.

3. What impact can a hidden variable have on a linear trend?
- A It can hide or obscure the linearity.
 - B It can cause an irregularity in an otherwise linear trend.
 - C Both A and B are possible.
 - D It cannot have an impact on the trend.

Use this information to answer #4 to #6. Is science ability related to math ability? The table shows a set of final grades for a number of intermediate students.

Math Grade	Science Grade
80	75
72	76
84	52
67	70
58	62
90	88
75	77

4. a) Create a scatter plot of science marks versus math marks. Perform a linear regression.
- b) Is this a good linear model? Explain why or why not.
5. a) Create a residual plot.
- b) Determine the residual for (84, 52).
- c) How does this residual compare to the others?
6. a) Repeat the analysis of the previous two questions after removing (84, 52).
- b) Compare the new linear model to the original. Which do you think is better and why?

Apply

7. **Communication** Jonathon's test scores are 80%, 84%, 83%, 40%, and 83%.
- a) Which score appears to be an outlier? Explain.
 - b) Determine Jonathon's mean, median, and mode scores.
 - c) Remove the outlier. Discuss the impact this has on Jonathon's
 - mean score
 - median score
 - mode score

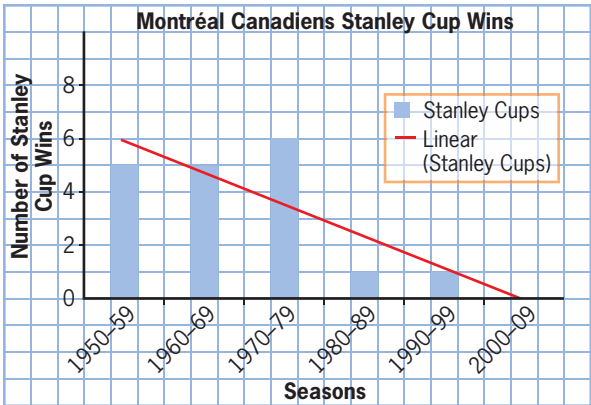
Use this information to answer #8 and #9. The table shows the weekly earnings of a restaurant server, including tips.

Time Worked (h)	Earnings (\$)
30	540
25	510
33	605
26	780
35	620
29	525

8. a) Construct a scatter plot of earnings versus time worked. Describe the correlation.
- b) Perform a linear regression. Interpret the meaning of the equation of the line of best fit.
- c) Is this a useful linear model? Explain.
9. a) Construct a residual plot.
- b) Identify an outlier in the data. What could account for the unusual data point?
- c) Repeat the analysis of #8 with the outlier removed.
10. **Open Question**
- a) A set of two-variable data has no outliers. Draw a sketch that shows what its residual plot could look like.
 - b) Repeat part a) for a data set that has an outlier.

Use this information to answer #11 to #13.

The graph illustrates the number of Stanley Cup wins by the Montréal Canadiens over time, measured in decades. For these questions, 1950 refers to the 1949–50 season, and so on.



11. a) What does this graph suggest about the performance trend of the Montréal Canadiens over the 50-year period?
- b) Any team in the National Hockey League (NHL) is eligible to win the Stanley Cup. Consider the table, which shows how the number of teams in the NHL changed over time. Identify a possible hidden variable related to the correlation shown in the graph.

Year	Number of Teams	Year	Number of Teams
1940	7	1980	21
1943	6	1992	22
1968	12	1993	24
1971	14	1995	26
1973	16	1999	27
1975	18	2000	28
1979	17	2001	30

12. The Stanley Cup was not awarded in the 2004–05 season due to a labour disruption. Discuss how this could also represent a hidden variable in this study.
13. Based on the given data, could you make an argument that the Montréal team of the 1970s was a better hockey team than those of the 1950s or 1960s? Explain.

Achievement Check

14. The table shows the average annual attendance for a minor league baseball team.

Year	Attendance (thousands)
2001	5.6
2002	5.8
2003	6.3
2004	6.5
2005	6.7
2006	6.8
2007	4.9
2008	5.3
2009	6.5
2010	7.0
2011	7.2
2012	7.4

- a) Construct a scatter plot of this time series. Is baseball interest on the rise?
- b) Perform a linear regression. Describe the strength of correlation.
- c) What graphical evidence is there of a hidden variable?
- d) In 2007, a large factory shut down due to the poor economy. How do you think this affected this correlational study?
- e) What do you think happened over the next few years following the plant closure?
- f) Repeat the linear regression with the data points for 2007–2009 removed. Compare this linear model to the previous one.
- g) Reflect on the interest in baseball now.

Extend

15. Conduct some research on residuals.
- a) How can you tell if a point will have a positive or negative residual?
- b) What is the sum of all residuals for a set of data? Why does this make sense?
16. Choose a set of data in this section. Calculate the residuals manually. Then, Construct a residual plot without using technology. Explain your method.

Uses and Misuses of Data

Learning Goals

I am learning to

- recognize that the same data can be presented in different ways
- see that the way data are presented can have an impact on how the data are interpreted
- recognize when, how, and why data are deliberately distorted in order to influence the perception of the reader



Minds On...

Data can help us make wise decisions. But data can also cause damage if not used properly. The Education Quality and Assurance Office (EQAO) provides detailed data related to the performance of Ontario students in reading, writing, and mathematics. Educators can use these data to identify areas of strength and weakness in their students' achievement.

- In what ways could educators use these data to help children with their learning?
- Are there any ways that this type of data could be misused?

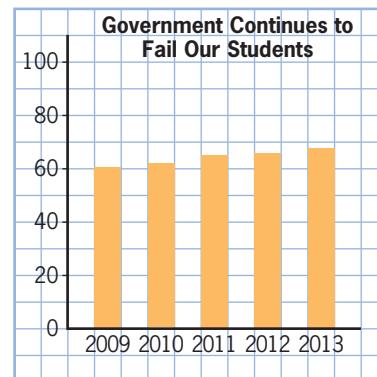
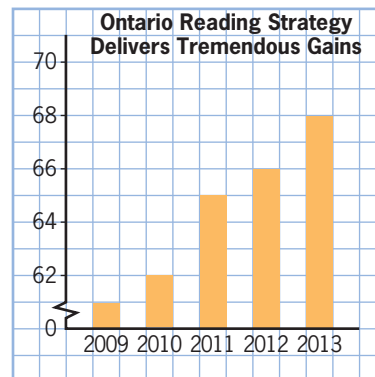
Action!

Investigate Bias in Two-Variable Data

Materials

- ruler

The graphs represent Ontario's grade 3 EQAO reading scores for the school years from 2008–09 to 2012–13. Each bar represents the number of students who met or exceeded the provincial standard in reading.



Source: Education Quality and Accountability Office

1. What is the same about the numerical information in these graphs? What is different? Explain.
2. Identify any elements of bias present in these two graphs.
3.
 - a) Do you think the title of the first graph is appropriate? Why or why not?
 - b) Do you think the title of the second graph is appropriate? Why or why not?
 - c) Write a more appropriate title for these graphs. Explain why your title is a better one.
4. **Reflect** Suggest a particular group that might have constructed each graph, and explain why that group might have presented the graph in the way that it did.
5. **Extend Your Understanding** Read and reflect on the following quotes. What do you think these authors are suggesting or implying?

“Be able to analyze statistics, which can be used to support or undercut almost any argument.”

Marilyn vos Savant

“There are two ways of lying. One, not telling the truth, and the other, making up statistics.”

Josefina Vazquez

“Statistics: the only science that enables different experts using the same figures to draw different conclusions.”

Evan Esar

Processes

Connecting

The authors of these quotes are interesting people. Use the Internet to research them. Also look up other interesting quotes on statistics.

How you choose to display data can have a significant impact on how a reader is likely to interpret them. By changing the vertical scale on a graph, a relatively small trend can be made to appear much larger, and vice versa. The intentional use of biased words can influence the interpretation of data. As a critical user of data, it is important to recognize when, how, and why data are being distorted.

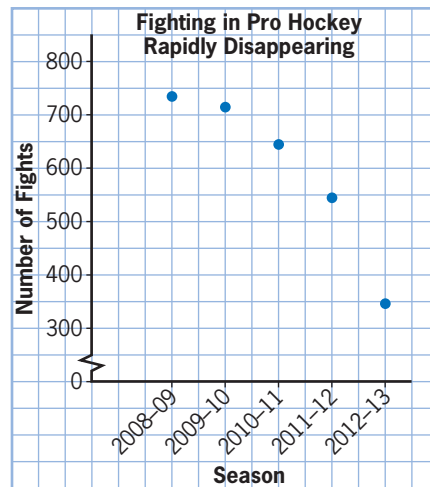
Example 1

Distorting Data to Sway Opinion

Is fighting in hockey on its way out in the NHL? The graph and table illustrate the total number of fights per season in NHL hockey games over time.

Season	Games	Fights
2012–2013	720	347
2011–2012	1230	546
2010–2011	1230	645
2009–2010	1230	714
2008–2009	1230	734

Source: NHL Fight Stats Table, hockeyfights



- Identify the sources of bias in the graph.
- Present the data in an unbiased way and reinterpret the data.

Solution

- Bias exists in both the data and the title of the graph.

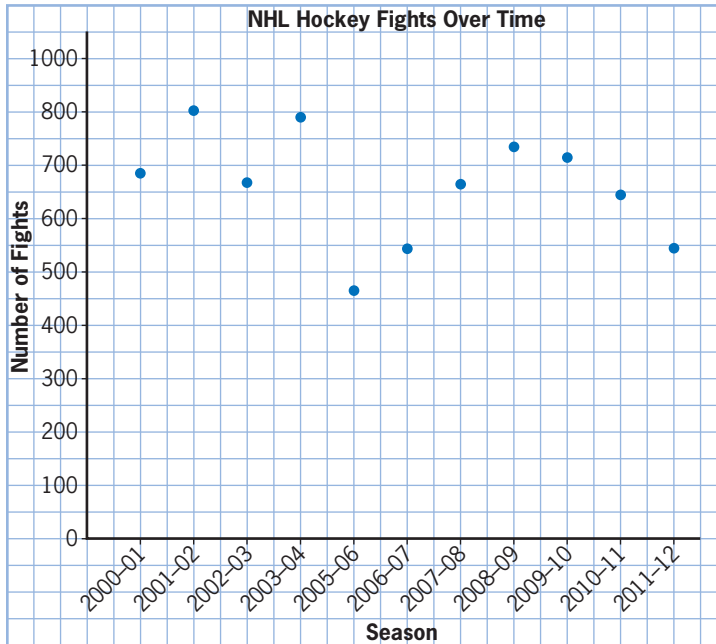
Processes

Reflecting

Look at the dramatic drop in fights from 2011–12 to 2012–13. What happened there?

Source of Bias	Explanation	Probable Intended Effect
Title	The language is not neutral.	<ul style="list-style-type: none"> to convince that fighting is on the decline
Vertical scale	The scale does not start at zero.	<ul style="list-style-type: none"> to exaggerate the downward trend
Outlier (2012–13, 720) due to a hidden variable	Only 720 games were played during this season due to a labour disruption.	<ul style="list-style-type: none"> to exaggerate the downward trend
Sample size	Only five points were chosen.	<ul style="list-style-type: none"> to hide that the trend is less downward over a longer period of time

- b) To present the data more fairly, remove the outlier, use a larger sample size, start the vertical scale at zero, and give the graph a title that has neutral language.



This graph illustrates the data in an unbiased way. While there still seems to be a slight downward trend in the number of fights over time, it appears to be much more gradual than the way it was originally presented. There is little in this graph to suggest that fighting will disappear from the NHL any time soon.

Your Turn

Is the number of multiple-fight NHL games on the decline? The table shows the number of games per season with more than one fight over time.

Season	Games With More Than One Fight
2012-13	66
2011-12	98
2010-11	117
2009-10	171
2008-09	173

Source: NHL Fight Stats Table, hockeyfights

- Create a scatter plot that shows the number of games with more than one fight as a time series.
- Is there an outlier in the data? If so, should it be removed? Explain.
- Identify any other sources of bias that could be distorting the linear correlation.

Example 2

Sensational Use of Data

The following article was taken from an Internet news blog.

The aliens are getting braver: UFO sightings in Canada doubled last year

By Lindsay Jolivet | [Daily Buzz](#)

The alien invasion has arrived at Earth's doorstep.

Canadian reports of UFO sightings more than doubled between 2011 and 2012, according to a report released this week by Ufology Research.

Sure, the skies have become more and more crowded with space junk and debris. And yes, it's possible that we're all getting just a little bit crazier.



People look at the night sky using night vision goggles during a UFO tour (Reuters)

However it's safest for everyone if we interpret the data as irrefutable evidence of an imminent attack from extra terrestrials that are swarming our skies in greater numbers than ever.

Canadians reported 1,981 UFO sightings in 2012, according to the report, which analyzed data of reported sightings from researchers and websites that monitor UFOs.

About 40 per cent of last year's reported sightings took place in Ontario.

Library and Archives Canada keeps a database of historical UFO sightings, including a case in 1969 when residents of Prince George, B.C. saw a round, glowing object ascending into the sky. An official RCMP investigation, detailed in a report, found the object was a plastic laundry bag and candles converted into a makeshift hot air balloon.

Well, you can't be too careful with these things.

The report says only a tiny fraction of cases involved close encounters with UFOs and in fact, most were merely sightings of lights in the sky. The report speculates on several explanations, including an increase of secret military exercises or a lack of knowledge about objects frequenting our skies that are not exactly "unidentified."

Chris Rutkowski, Ufology's research director, told the *Winnipeg Sun* last year there was no definitive evidence that any of the people reporting sightings have been in contact with aliens.

But what if he's one of *them*?

Source: Yahoo! News Canada, October 12, 2013

- Discuss how data are used in this article in a questionable way.
- What inappropriate conclusions are drawn?
- What is the probable intent of the article?

Solution

- a) The author does not explain how the data were collected, nor is it made clear what constitutes a “UFO sighting.” It may be possible that people who claim sightings are more likely to believe in extra-terrestrial beings and therefore more likely to misinterpret a normal event as a UFO sighting. The data may be subject to voluntary response bias.
- b) The title of the article draws an inappropriate conclusion that UFO sightings imply the existence and nearby presence of extra-terrestrial beings. This false conclusion is further established in the third paragraph where the author infers that an invasion is imminent.
- c) The article does go on to offer alternative explanations for the reported sightings, and offers a reputable quote suggesting no definite evidence of alien contact. This, combined with the playful opening and final sentences and overall sensational tone of the article, suggest that this was written as a piece of satire, more to entertain than inform.

Your Turn

Consider the following statement taken from the article:

About 40 per cent of last year’s reported sightings took place in Ontario.

Does this suggest that aliens are more interested in Ontario than they are in other provinces? Explain your thinking.

Literacy Link

In literature, *satire* refers to a written piece that uses sarcasm and other forms of humour to make fun of something. *Sensationalism* is a type of bias where a piece of work uses tactics such as over-exaggeration to provoke an emotional response. The author probably hopes she will increase her readership by generating controversy in these ways.

Consolidate and Debrief

Key Concepts

- You can display data in multiple ways. Sometimes data are deliberately distorted to make an argument more convincing.
- The media often sensationalize data to generate public interest.

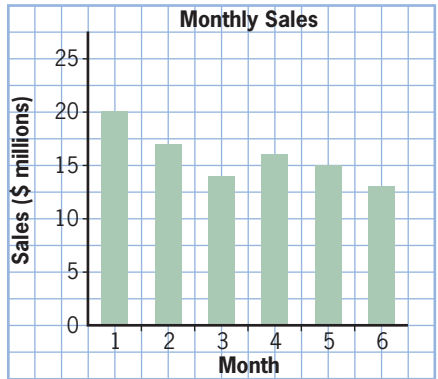
Reflect

- R1. a) In what ways can data be deliberately distorted?
b) List some motives someone could have to deliberately distort data.
- R2. A juice company shares data in a television commercial that suggest its brand of juice is twice as popular as a competing brand.
a) Why would this company want to share this information?
b) What information would you like to know before making an informed interpretation of the study’s findings?
- R3. a) What is meant by the sensational use of data?
b) What is the purpose of using data in a sensational way?

Practise

Choose the best answer for #1 to #3.

1. The graph shows a sales trend over the past six months. Which of the following is the most appropriate title for this graph?



- A Sales in sharp decline
B Sales holding steady
C Sales poised to bounce back
D Sales trend over the past six months
2. On a television commercial, three enthusiastic parents claim that a new home study program has done wonders for their children’s progress in math. This correlational evidence is likely subject to which of the following biases?
- A sample size bias
B random sample bias
C both
D neither
3. Which statement is least likely to have been made for a sensational purpose?
- A Human and chimpanzee DNA differ by 0.7%.
B 80% of all automobile accidents occur within 40 kilometres of home.
C The unemployment rate rose this month by 0.5%.
D Three out of four people are easily impressed by statistics.

4. In the supermarket, you see a tabloid newspaper with the following headline: “20% of America convinced the King still lives!” This headline is in reference to Elvis Presley, commonly known as the King of Rock ’n Roll, who died in 1977.



- a) How reliable do you consider the data referenced in this headline to be?
b) What questions would you like answered about how the data were collected?
c) What is the likely intended use of the data?

Apply

Use this information to answer #5 to #7.

The data show monthly reported UFO sightings for the period from January 1, 2010 to October 12, 2013.

Reports	Count	Reports	Count
Oct-13	34	Nov-11	440
Sep-13	723	Oct-11	634
Aug-13	850	Sep-11	546
Jul-13	947	Aug-11	634
Jun-13	609	Jul-11	749
May-13	515	Jun-11	393
Apr-13	405	May-11	315
Mar-13	383	Apr-11	315
Feb-13	275	Mar-11	328
Jan-13	369	Feb-11	270
Dec-12	654	Jan-11	325
Nov-12	766	Dec-10	302
Oct-12	664	Nov-10	358
Sep-12	747	Oct-10	465
Aug-12	884	Sep-10	448
Jul-12	919	Aug-10	524
Jun-12	741	Jul-10	833
May-12	511	Jun-10	372
Apr-12	495	May-10	327
Mar-12	527	Apr-10	293
Feb-12	387	Mar-10	261
Jan-12	574	Feb-10	186
Dec-11	530	Jan-10	291

Source: National UFO Reporting Center Report Index by Month

5. Application

- There is a clear outlier in this data set. Identify it and explain why it should be excluded from the analysis.
- Create a scatter plot of UFO sightings versus time.
- Does the number of UFO sightings appear to be on the rise? Explain.

6. Thinking

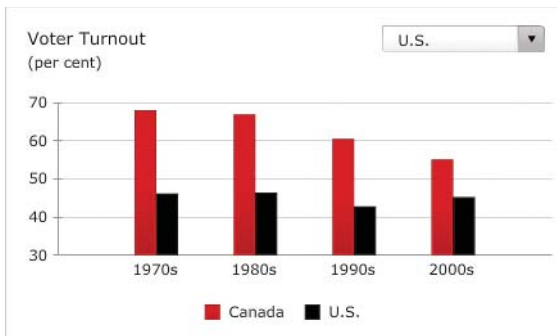
- Describe any seasonal trends in the data.
- Does this suggest that we are more or less likely to be visited by extra-terrestrial beings at different times of the year? Explain your thinking.

7. Communication

- Do you think there is bias in these data? Why or why not?
- Identify some questions you would like answered about how the data were collected.

Achievement Check

8. Are Canadian citizens politically involved? The graph below compares voter turnout for Canadian and US citizens over time.



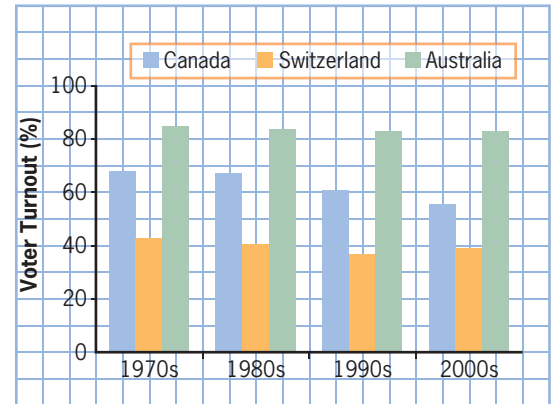
Source: Conference Board of Canada

- Open Question** Write a title for this graph that suggests Canada's political interest is consistently strong.
- How could you adjust the vertical scale of the graph to emphasize your point?

- Open Question** Write a title for this graph that suggests Canada's political interest is declining.
- How could you adjust the vertical scale of the graph to support your point?

Extend

9. The graphs below compare voter turnout for Canadian, Swiss, and Australian citizens over time.



Source: Conference Board of Canada

- Identify any trends that you see. Suggest some possible reasons for these.
 - Perform some research on the political situation of these countries. Identify and report on any hidden variables that you discover. Your teacher can direct you to some helpful websites.
10. Search the Internet or print media for an example of a sensational use of data.
- What statistical claim(s) are being made?
 - What is the likely purpose of the article?
 - Why would you consider the data to be unreliable?

Advanced Techniques for Data Analysis

Learning Goals

I am learning to

- organize and display data using a variety of tools and strategies
- analyse various representations of data

Minds On...

People who exercise regularly get many health benefits such as lower body fat, higher muscle mass, and better respiratory health.

- How does regular exercise affect resting heart rate?
- How could you use data analysis to find out?



Action!

Example 1

Use a Filter to Compare Groups

The heart rates for two groups of women in their thirties are shown below. The women in Group A train to run marathons. The women in Group B do not exercise regularly.

Group	Resting Heart Rate (bpm)	Working Heart Rate During Exercise (bpm)	Heart Rate 1 Min After Exercise (bpm)	Heart Rate 30 Min After Exercise (bpm)
A	55	175	117	77
B	76	187	125	96
A	59	184	115	85
B	74	189	128	92
B	67	182	134	87
B	80	193	131	90
A	64	179	126	72
A	52	173	121	80
B	65	198	137	96
A	57	186	114	74
B	71	186	134	87
A	59	181	119	79

- Summarize and compare the resting heart rate and working heart rate of Group A and Group B.
- Identify any potential bias in the analysis of these data.

Solution

- You are comparing quantitative, or numerical, data across different categorical groups. One strategy to do this is to run the quantitative data through a categorical filter, training.

Method 1: Analyse Side-by-Side Box Plots Using Fathom™

Import or enter the data into Fathom™. Construct and analyse box plots to compare aggregate resting heart rate and maximum heart rate scores.

- Drag a **New Graph** into the workspace.
- Click and drag the Resting Heart Rate attribute onto the horizontal axis.
- Click on the pull-down menu to change the **Dot Plot** to a **Box Plot**.
- Click and drag the Group attribute onto the vertical axis.
- Repeat these steps for Working Heart Rate.
- Align the graphs vertically.

Adjust the vertical scales to eliminate data distortion.

- Drag the ends of the scale to stretch or compress it.
- Drag the middle of the scale to slide it.

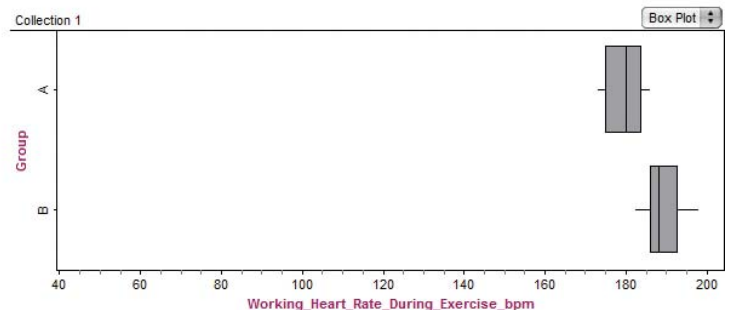
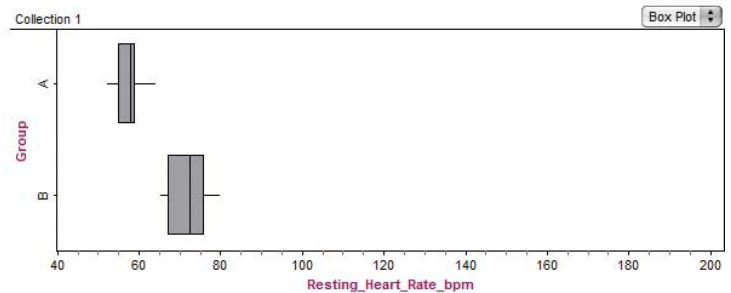
Processes

Representing

In chapter 6, box plots were called box and whisker plots. Both names are commonly used.

Collection 1

Collection 1	Group	Resting...	Workin...	Heart_R...	Heart_R...
1	A	55	175	117	77
2	B	76	187	125	96
3	A	59	184	115	85
4	B	74	189	128	92
5	B	67	182	134	87
6	B	80	193	131	90
7	A	64	179	126	72
8	A	52	173	121	80
9	B	65	198	137	96
10	A	57	186	114	74
11	B	71	186	134	87
12	A	59	181	119	79



The side-by-side box plots show that Group A had both a lower resting heart rate and a lower working heart rate than Group B.

Literacy Link

A contingency table shows the frequency distribution of variables in different categories.

Filter the table data to compare the summary statistics by group.

- Drag a new **Summary Table** into the workspace.
- Drag the Resting Heart Rate attribute to the right arrow on the summary table.
- Repeat for Working Heart Rate.
- Click and drag the Group attribute to the down arrow of the summary table.

Collection 1		Resting_Heart_Rate_bpm	Working_Heart_Rate_During_Exercise_bpm
Group	A	57.6667 6	179.667 6
	B	72.1667 6	189.167 6
Column Summary		64.9167 12	184.417 12

S1 = mean ()

S2 = count ()

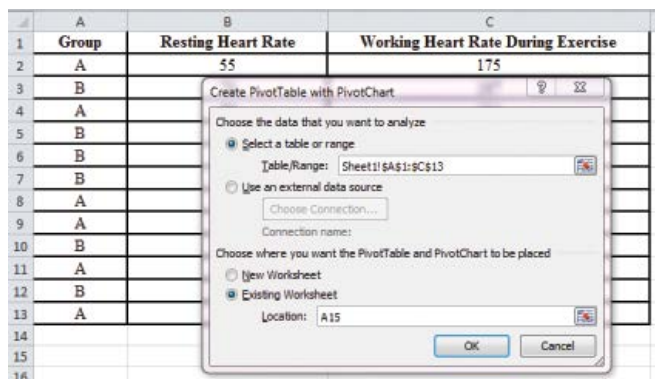
The contingency tables show that the marathon runners had a mean resting heart rate of 57.7 bpm compared to a rounded mean of 72.2 bpm for those who did not exercise. The maximum heart rate of the marathon runners had a mean of 179.7 bpm compared to a mean of 189.2 bpm for the group that did not exercise.

Method 2: Analyse a Pivot Table Using a Spreadsheet

A pivot table summarizes data by sorting, giving count totals, or averaging data found in large data tables. Pivot charts provide a visual representation of the summary in the pivot table.

Import or enter the data into a spreadsheet. Use cells A1 to C13. Create a pivot chart and pivot table to compare resting heart rates and working heart rates.

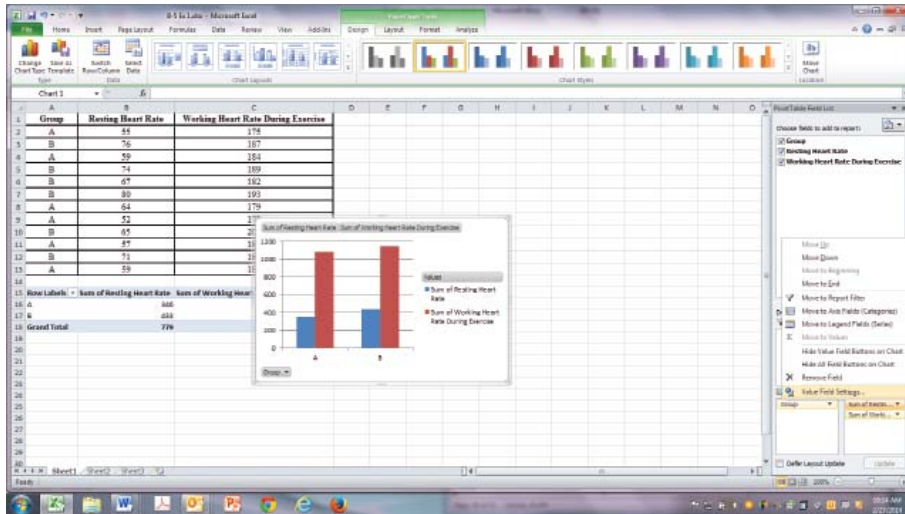
- Select the entire data set.
- From the **Insert** ribbon, click on **PivotTable** and choose **PivotChart** from the drop-down menu.
- Choose **Existing Worksheet** and type A15 into the Location field. Click **OK**.



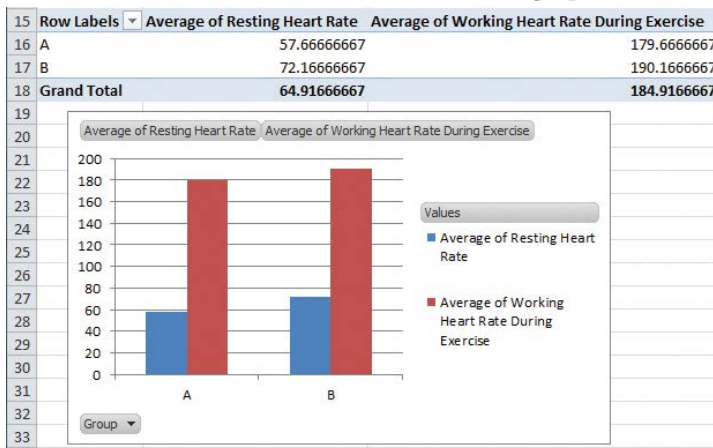
Customize the pivot table to calculate and display the mean heart rates.

- In the **PivotTable Field List**, check the boxes for Group, Resting Heart Rate, and Working Heart Rate During Exercise.

- In the Σ VALUES section, click on **Sum of Resting Heart Rate** and then choose **Value Field Settings**.



- Change **Sum** to **Average** and click **OK**. Repeat for **Sum of Working Heart Rate**.
- Click **OK**.
- Close the **Pivot Chart Fields** window. The graph will look like this:



Both the pivot table and chart show that marathon runners had a lower average resting heart rate than non-runners, with an average of 57.7 bpm compared to 72.2 bpm. They also show that marathon runners had a lower average working heart rate than non-runners with an average of 179.7 bpm compared to 190.2 bpm.

- b) The analysis of the data in this example is subject to sample size bias. This could be addressed by including additional cases in the study.

Your Turn

Open a new spreadsheet. Summarize and compare the performance of both groups in this example in heart rate after 30 min of exercise and heart rate after 1 min of exercise.

Example 2

Use Graphing Tools to Uncover a Hidden Variable

The table shows distance-time data for a number of bicycle runs. The altitude data represent the net change in altitude over the course of the run.

Distance (km)	Time (min)	Altitude (m)
14.9	37	-80
13.6	28	-200
12.1	42	110
15.7	57	170
20.0	35	-350
15.7	42	100
12.6	37	170
15.3	50	80
14.8	40	-80
11.2	25	-100
14.6	50.3	340

Determine if there is a correlation between distance and time.

Solution

Method 1: Use a Bubble Plot in a Spreadsheet

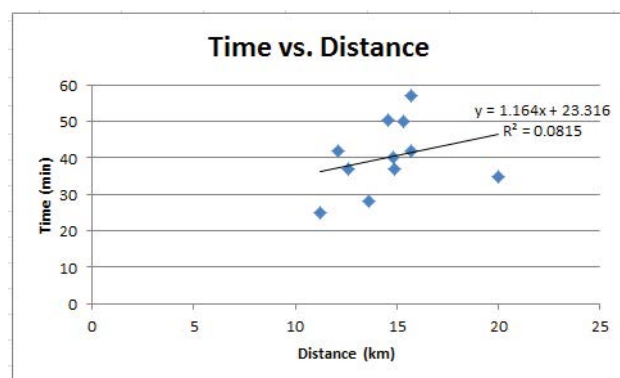
A bubble plot can show the influence of a third variable in a correlational study. In a bubble plot, each point becomes a bubble whose size corresponds to the value of the third variable. In the case of positive and negative values, the colour of the bubble becomes significant.

Enter the data in a spreadsheet. Create a scatter plot of time versus distance and perform a linear regression.

Processes

Reasoning and Proving

Usually distance is plotted against time. Why does it make sense to plot time versus distance in this case?

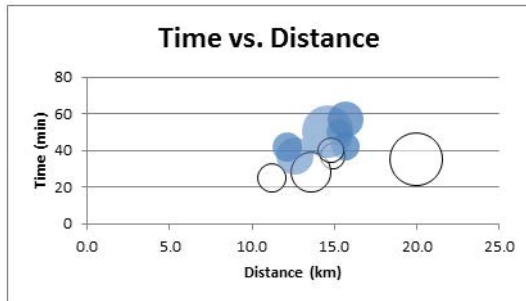


There does not appear to be a strong linear correlation between run time and distance, which seems unusual. The correlation coefficient $r = 0.29$ suggests a weak linear correlation.

Consider the altitude data. Could this be a hidden variable obscuring the linear correlation?

Construct a bubble plot to see if altitude change has an impact on the trend.

- Select the entire data set. From the **Insert** ribbon, click on **Other Charts** and choose **Bubble Chart**.
- Right click on one of the bubbles and choose **Format Data Series**.
- Under **Series Options**, check **Show negative bubbles**.
- Close the **Format Data Series** window.

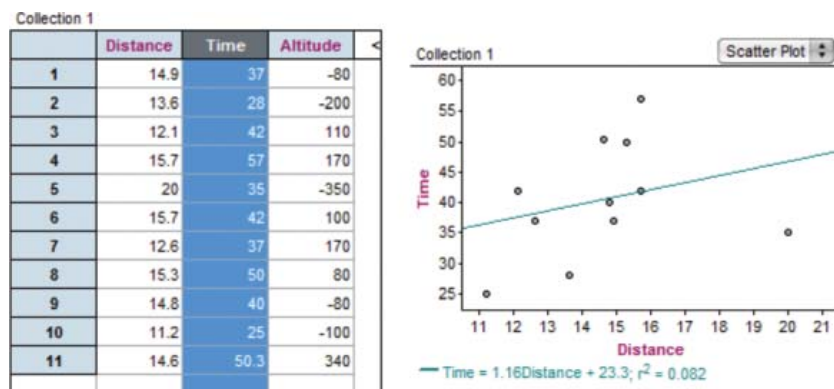


The bubble plot confirms that change in altitude is obscuring the linear correlation between run time and distance. The uphill runs, represented by the blue bubbles, tend to have longer run times than the downhill runs, represented by the white bubbles.

Method 2: Use a Legend Attribute in Fathom™

In Fathom™, you can identify the impact of a third variable by analysing a legend attribute. A legend attribute adds a colour scale to the data points, which corresponds to the value of the third variable being measured.

Enter the data into a Fathom™ table. Create a scatter plot of time versus distance and perform a linear regression.

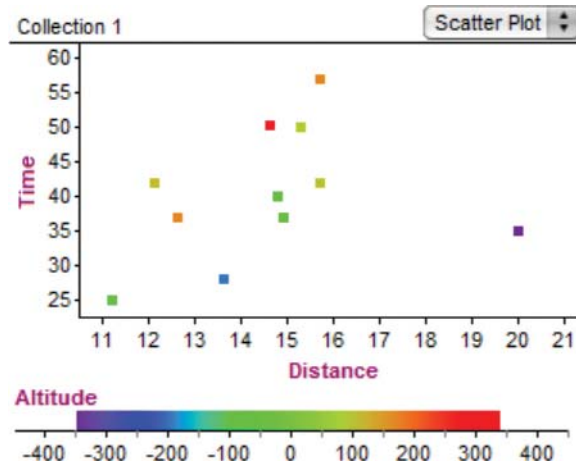


There does not appear to be a strong linear correlation between run time and distance, which seems unusual. The correlational coefficient $r = 0.29$ suggests a weak linear correlation.

Consider the altitude data. Could this be a hidden variable obscuring the linear correlation?

Add a legend attribute scale to see if altitude change has an impact on the trend.

- From the **Graph** menu, uncheck **Least-Squares Line**.
- Click and drag the Altitude attribute onto the middle of the graph.



The altitude attribute scale confirms that altitude is obscuring the linear correlation between run time and distance. The uphill runs, represented by the yellow, orange, and red points, tend to have longer run times than the downhill runs, represented by the purple, blue, and bright green points.

To account for this hidden variable, it would make sense to collect and analyse more distance-time data for runs with identical altitude changes.

Your Turn

- a) In Fathom™, open the sample file found here: **Sample Documents/Sports/Bicycling.ftm**. Repeat the analysis performed in this example using the data. A sample of the data is shown.

Distance (km)	Time (min)	Altitude (m)
14.5	33.7	-80
9.7	15.2	-270
8.9	14.4	-130
12.0	58.0	600
14.4	20.0	-600
17.9	31.0	-350
10.3	27.0	0
5.6	15.0	70
10.3	31.0	155
9.5	31.0	125
9.2	17.5	-125
5.4	15.8	45
7.9	36.2	380
6.4	19.2	-160
14.5	38.4	-140

- b) Are your results consistent with those in the example? Discuss.

Consolidate and Debrief

Key Concepts

- Raw data can be awkward to work with. Software programs have tools to filter, organize, present, and analyse data.
- Contingency tables, side-by-side box plots, pivot tables, and pivot charts are methods for comparing quantitative data across different categories.
- Bubble plots and legend attributes are tools that can be used to recognize the possible impact of a hidden variable in a correlational study.

Reflect

- R1. a)** What is a contingency table?
b) How can a contingency table be useful?
- R2.** Refer to Example 1.
a) How were side-by-side box plots used?
b) What did they illustrate?
- R3.** Refer to Example 2. What information did the bubble plot and the legend attribute provide that the scatter plot did not?
- R4.** Consider the tools and strategies used in this section. Write a summary of their advantages and disadvantages. Use examples to support your points.

Practise

Choose the best answer for #1 to #3.

Use this information to answer #1 and #2.

The contingency table shows the unemployment rate for non-student youths aged 15 to 24.

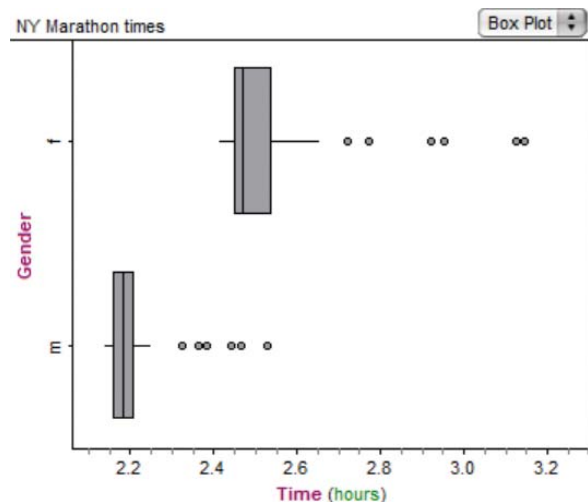
Year	Male Unemployment Rate (%)	Female Unemployment Rate (%)
2008	14.9	11.3
2009	19.8	11.6
2010	19.8	13.2
2011	15.6	12.2
2012	17.5	13.3

Source: CANSIM Table 282-0095, Labour force survey estimates (LFS), by full- and part-time students during school months, sex and age group, Statistics Canada, January 9, 2014

1. What are the mean unemployment rates for the time period shown?
A 15.6 for males and 12.2 for females
B 17.5 for males and 12.3 for females
C 19.8 for males and 13.25 for females
D 14.9 for males and 11.6 for females
2. Which of the following characterizes the linear correlation between male and female unemployment rates?
A strong positive
B strong negative
C moderate positive
D moderate negative

Use this information to answer #3 and #4.

The graph shows the winning times for the New York Marathon from 1970–1999.



Source: Fathom™: Sample Documents/Sports/MarathonTimes

3. Which of the following statements is true?

- A The fastest female time is about 2.5 hours and the fastest male time is about 3.2 hours.
- B The mean male winning time is about 2.2 hours and the mean female time is about 2.5 hours.
- C The median male winning time is about 0.3 hours faster than the median female time.
- D None of the above is true.

4. What are the fastest winning times shown for males and females?

Apply

Use the data in Example 1 on page 424 for #5 and #6.

5. **Application** In Fathom™, select a summary table. From the **Summary** menu, choose **Add Basic Statistics**. What summary statistics are generated?

6. **Application** In Fathom™, select a summary table. From the **Summary** menu, choose **Add Five Number Summary**. What summary statistics are generated?

Use the data in Example 2 Your Turn on page 430 for #7 and #8.

7. a) Use a filter in Fathom™ to separate the downhill runs from the uphill runs.
 - Create a scatter plot of time versus distance, as in the example.
 - From the **Object** menu, choose **Add Filter**.
 - Choose **Attributes** and double click on **Altitude**.
 - Type < 0 m and click **OK**.
 The uphill runs will be filtered out leaving only the downhill runs.
 - b) Perform a linear regression and interpret the equation of the line of best fit.
8. a) Repeat the analysis filtering out the downhill runs.
 - b) Compare the two linear models. Does it make sense for them to differ in this way? Explain.

Use this information to answer #9 to #11.

Your teacher will provide you with a Fathom™ file called **Hotdogs.ftm**.

9. **Thinking** Is there a correlation between sodium and fat content in hot dogs?
 - a) Create a scatter plot of number of calories versus amount of sodium.
 - b) Perform a linear regression. Describe the correlation.
 - c) Repeat the analysis separately for each type of meat. Compare the linear models and the strength of correlation to the original model in each case.
 - d) Is the type of meat a hidden variable in the correlation between calories and sodium? Explain why or why not.

Processes

Selecting Tools and Computational Strategies

To analyse only beef hot dogs, you can use a filter:

- With the scatter plot selected, click on **Object** and choose **Add Filter**.
- Choose **Attribute** and double-click on **Type**.
- Type = "Beef" and click on **OK**.

10. Communication Which types of hot dogs have the least amount of sodium?

- Create side-by-side box plots to compare the amount of sodium in each type.
- Create summary statistics and interpret the results.

11. Which types of hot dogs have the lowest number of calories? Repeat the analysis from the previous question for number of calories.

✓ Achievement Check

12. The table provides preparation and performance data for a number of students at a driving school.

Gender	Study Hours	Road Hours	Written Test (%)	Road Test (%)
M	4	9	80	95
M	6	8	70	90
F	6	5	75	80
F	7	6	80	85
F	8	5	95	85
F	7	7	85	85
M	5	8	80	85
M	4	6	75	70
F	8	7	90	85
M	3	8	75	90
M	5	7	85	80
F	9	4	70	75
M	2	5	85	80
F	8	7	80	85
M	4	9	90	95
F	7	9	90	95

- Use tools and strategies from this section to compare the average number of hours studied and written test score for males and females.
- Compare the performance of the two gender groups.
- Repeat the analysis for average number of road hours and road test score for both genders.

13. Application A group of elementary students were chosen at random to write a literacy test. Their results are shown.

Grade	Score
4	7
3	6
5	8
2	3
2	4
6	6
5	7

- Use Fathom™ to construct a scatter plot. Describe the correlation.
- Construct a dot plot with a grade legend attribute.
 - Drag a **New Graph** into the workspace.
 - Drag the **Score** attribute onto the horizontal axis.
 - Drag the **Grade** attribute onto the middle of the graph.
- How does the second graph represent the correlation between grade and score?

Extend

Use this information to answer #14 and #15.

The table shows the popularity of four Ontario political parties over the course of five surveys.

Conservative	34	35	35	35	34
Liberal	31	29	31	30	47
NDP	29	28	27	25	16
Green	6	8	5	6	2

Source: Ontario Voter Support Separated by Party, Laurier Institute for the Study of Public Opinion

14. Application

- Enter the data into a spreadsheet. Create a column of sparklines.
 - Highlight the numerical data.
 - From the **Insert** ribbon, choose **Line** from the Sparklines menu.
 - To enter a Location Range, highlight four empty cells in a column beside the data table. Click OK.
- Describe what appears.

15. Communication

- What do the sparklines illustrate?
- What do they not show clearly?
- Do any sparklines distort the data? Explain.

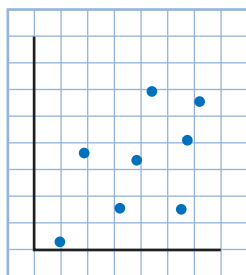
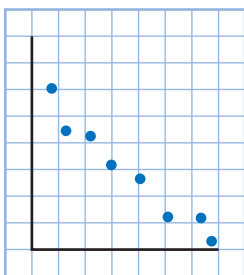
Chapter 8 Review

Learning Goals

Section	After this section, I can
8.1	<ul style="list-style-type: none">• classify a linear correlation between two variables• determine a correlation coefficient using technology• produce a line of best fit using linear regression
8.2	<ul style="list-style-type: none">• distinguish between correlation and causality• identify the type of relationship between two variables
8.3	<ul style="list-style-type: none">• identify outliers and account for their impact on a data trend• recognize the presence of extraneous variables• identify a hidden variable and account for its impact on a correlation
8.4	<ul style="list-style-type: none">• recognize that the same data can be presented in different ways• see that the way data are presented can have an impact on how the data are interpreted• recognize when, how, and why data are deliberately distorted in order to influence the perception of the reader
8.5	<ul style="list-style-type: none">• organize and display data using a variety of tools and strategies• analyse various representations of data

8.1 Line of Best Fit, pages 382–391

1. Match each scatter plot with its correct correlation coefficient.



Correlation coefficients:
 -0.97 , -0.56 , 0.56 , 0.97

2. The table shows a trucker's distance from home over time.

Time, t (h)	Distance, d (km)
0.5	500
1.0	425
1.5	360
2.0	280

- a) Use graphing technology to create a scatter plot of the data.

- b) Determine the strength of linear correlation between these variables.
- c) Perform a linear regression. Interpret the meaning of the equation of the line of best fit.

8.2 Cause and Effect, pages 392–401

3. Children's self-esteem is positively correlated with their level of achievement.
- a) Suggest a cause and effect relationship that could account for the results.
- b) What reverse cause and effect relationship could also account for the results?
4. Characterize each of the relationships. The independent variable is listed first.
- a) Computer sales are negatively correlated with the unemployment rate.
- b) The price of gas is positively correlated with the performance of a football team.
- c) Running speed is positively correlated with heart rate.

8.3 Dynamic Analysis of Two-Variable Data, pages 402–415

Use this information to answer #5 and #6.

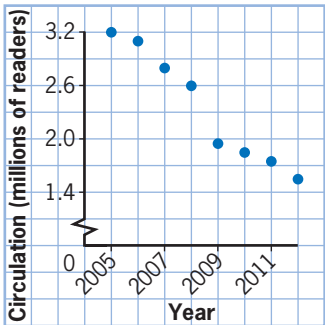
The table shows student population data for a new high school. The school wants to project the school's population growth over time.

Year	Population
2007	778
2008	984
2009	998
2010	1010
2011	1018
2012	1026

5. a) Create a scatter plot of population versus time. Call 2007 year 0. Describe the correlation.
b) Perform a linear regression. Interpret the equation of the line of best fit.
c) Create a residual plot. Does this appear to be a good linear model? Explain.
6. Grade 12 was not offered until 2008.
 - a) How does this information affect the correlational study? Does it make sense to remove the 2007 datum? Explain.
 - b) Repeat the analysis with the outlier removed. Compare the two models.
 - c) Use both models to predict the school's population in 2016. Which model should the principal rely on and why?

Use this information to answer #7 to #9.

The graph shows a newspaper's annual average circulation data.



7. a) Describe the trend in sales.
b) Is there evidence of a hidden variable?
c) When do you think the newspaper raised its price from \$1 to \$1.50? Explain.
d) Explain how the price change represents a hidden variable in this correlation.

8.4 Uses and Misuses of Data, pages 416–423

8. How does the vertical scale in the newspaper circulation graph distort the linear trend?
9. Suppose this graph were published with the headline "Newspaper circulation in free fall."
 - a) Explain how this title is biased.
 - b) Write an unbiased title for this graph.

8.5 Advanced Techniques for Data Analysis, pages 424–433

Use this information to answer #10 to #12.

The table shows a group of students' grade 12 calculus marks and first year university marks. Half took a summer prep course.

Yes Prep.		No Prep.	
Grade 12	First Year	Grade 12	First Year
80	77	77	65
70	72	74	68
92	86	68	60
84	84	70	64
85	84	81	72

10. a) Create a scatter plot that compares first year marks to grade 12 marks.
b) Perform a linear regression. Interpret the correlation coefficient.
c) Was the summer prep course helpful?
11. a) Create a contingency table and side-by-side box plots or a pivot table and pivot chart to compare the two groups.
b) Use summary statistics to determine if the summer prep course is helpful.
12. a) Use a bubble plot or a legend attribute to compare the two groups.
b) Does the graph show that the prep course is helpful? Explain.

Chapter 8 Test Yourself

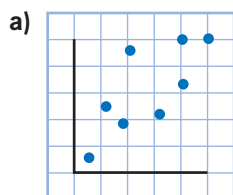
Achievement Chart

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

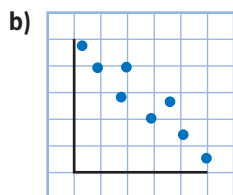
Multiple Choice

Choose the best answer for #1 to #3.

1. Classify the nature of each linear correlation.



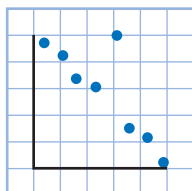
- A moderate positive correlation
- B strong positive correlation
- C moderate negative correlation
- D strong negative correlation



- A moderate positive correlation
- B strong positive correlation
- C moderate negative correlation
- D strong negative correlation

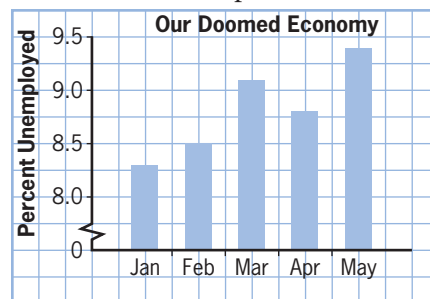
2. Consider the scatter plot relating two variables.

Which of the following best characterizes this correlation?



- A strong positive correlation
- B strong negative correlation
- C strong negative correlation with a hidden variable
- D strong negative correlation with an outlier

3. The graph shows the unemployment rate over a five-month period.



Which of the following contribute to bias in this graph?

- A the title
- B the sample size
- C the choice of vertical scale
- D all of the above

Short Answer

4. Classify each of the relationships. The independent variable is listed first.
- a) Snow tire sales are positively correlated with hot chocolate sales.
 - b) Movie box office sales are negatively correlated with ticket price.
 - c) Driving test scores are negatively correlated with driver height.
 - d) Cheeseburger sales are negatively correlated with pita sales.
5. Attendance at football games is positively correlated with a team's position in the league standings.
- a) Suggest a cause and effect relationship that could account for the results.
 - b) Pose and defend an argument for a reverse cause and effect relationship.

Extended Response

6. The table shows the speed of a skydiver as she falls from the instant she leaves a plane.

Time (s)	Speed (m/s)
0	0.0
1	5.0
2	9.1
3	12.5
4	15.8

- a) Construct a scatter plot of speed versus time. Describe the correlation.
- b) Perform a linear regression. Interpret the equation of the line of best fit.
7. The number of users of a social networking website is shown as a time series.

Time (months)	Users (millions)	Time (months)	Users (millions)
1	1.2	7	3.3
2	1.5	8	1.7
3	2.0	9	1.6
4	2.4	10	1.3
5	2.7	11	1.1
6	3.1	12	0.9

- a) Construct a scatter plot of this time series. Describe the trend.
- b) Perform a linear regression. Interpret the equation of the line of best fit.
- c) Construct a residual plot. Do you think this is a good model for this correlation? Explain why or why not.
- d) Is there evidence of a hidden variable? Explain.
- e) Initially this website was free. When did the website start charging a fee? How do you know? Explain the effect this had on the linear trend.
- f) Use a graph to illustrate the impact of this hidden variable on the time series correlation.
- g) Construct a new linear model that will do a better job of predicting the future popularity of this website. Discuss your reasoning, including any assumptions you make.

Chapter Problem

Climate Change

Present your findings on climate change in one of the following formats:

- written report
- electronic slideshow
- podcast
- poster or infographic

Consider your culminating project. What elements of two-variable data could be relevant? Will your project involve exploring multiple data sets? What tools and strategies from this chapter will you use to organize, filter, display, and analyse your data? How will you identify and account for outliers and hidden variables?

Chapters 7 and 8 Cumulative Review

1. The chart shows systolic blood pressure measurements for a sample of 30 athletes.

Systolic Blood Pressure (mmHg)					
120	118	118	115	120	119
124	120	115	122	121	117
121	124	116	123	118	121
117	123	116	116	123	122
119	115	124	117	122	119

- a) Is the distribution uniform?
- b) What is the height of the probability distribution?
- c) What is the probability that an athlete's blood pressure is 120 mmHg or less?
- d) How would you expect the distribution to differ if the general population was measured instead of a select group? Give reasons for your answer.
2. At a provincial high school track meet, 40 contestants participated in the discus event. The table shows the results.

Distance (m)	Frequency
25–30	1
30–35	3
35–40	5
40–45	9
45–50	6
50–55	6
55–60	5
60–65	3
65–70	1
70–75	1

- a) Sketch a frequency histogram for these data.
- b) Sketch a frequency polygon.
- c) Estimate the mean distance thrown.
- d) Add a relative frequency column to the table.
- e) What is the probability that a contestant throws the discus more than 60 m?
- f) How would you expect the distribution to differ if the data were taken from students in a physical education class instead of those who were participating in a track meet? Give reasons for your answer.

3. A large number of airplanes in Canada are built from kits. A company advertises that building its kit takes a mean time of 300 h, with a standard deviation of 40 h. The company has sold 120 such kits.

- a) What percent of the builders are expected to take longer than 400 h to complete their kits?
- b) How many builders are expected to finish their kits in less than 250 h?
- c) If 120 kits were given to people selected at random from the community, how would the distribution of times change? Give reasons for your answer.

4. An industrial wind turbine is designed to produce a maximum of 1.8 MW of power in a 60 km/h wind. Lab tests of 120 turbines showed that seven did not meet this standard.

- a) Determine a 99% confidence interval for the proportion of turbines from the plant that do not meet the standard.
- b) Use a standard statistical format suitable for a report to state the results of the lab tests.

5. A drug prescribed to prevent gout produces side effects in 2% of patients who take it. A new formulation is tested on 1500 patients, and 20 suffer side effects.

- a) Could this distribution be reasonably modelled using a normal approximation? Give reasons for your answer.
- b) Determine the mean and standard deviation of the normal approximation.
- c) What is the probability that the new formulation is no better than the original? Use the normal approximation to determine the probability that at most 20 patients suffered side effects.
- d) Is the company justified in claiming that the new formulation results in side effects in about 1.3% of the patients who take it? Give reasons for your answer.

6. The table gives the distance of a hiker from camp over time.

Time (min)	Distance (km)
10	0.5
20	1.1
30	1.4
40	1.7
50	2.4

- Create a scatter plot of distance versus time. Describe the correlation.
 - Perform a linear regression. Interpret the equation of the line of best fit.
7. A study found that worker absenteeism is negatively correlated with income level.
- Do you think this is a cause and effect relationship? Explain.
 - Suggest a possible common cause factor that could account for this relationship.
8. Characterize each type of relationship.
- Automotive sales is positively correlated with amount of rainfall.
 - Number of hours practised is negatively correlated with number of musical errors.

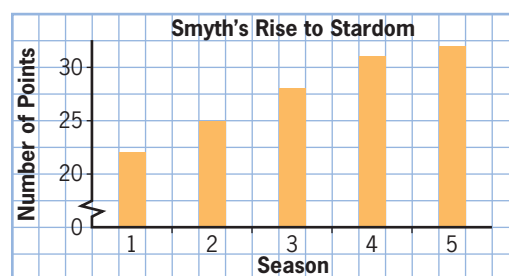
9. The table shows student achievement for the grade 9 applied EQAO assessment over time.

Year	Percent Achieving at or Above Provincial Standard
2001–2002	21
2002–2003	21
2003–2004	25
2004–2005	27
2005–2006	35
2006–2007	35
2007–2008	34
2008–2009	38
2009–2010	40
2010–2011	42
2011–2012	44
2012–2013	44

Source: Education Quality and Accountability Office

- Create a scatter plot for this time series. Call 2001–2002 year 0. Describe the correlation.
- Perform a linear regression. Interpret the equation of the line of best fit.
- Construct a residual plot. Does there appear to be any evidence of a hidden variable in the data? Explain.
- In 2005, the math curriculum was revised. Could this fact be considered a hidden variable? Why or why not?
- Repeat the analysis performed in parts a) to c) for 2005–06 to 2012–2013.
- Is there evidence that this linear model is better than the original one? Explain.

10. The graph shows the point totals for a hockey player's first five seasons.



- Identify any sources of bias in the graph.
 - Smyth's contract is up for renewal. Do you think the graph was made by the team manager or by Smyth's agent? Explain your thinking.
 - How could you remove all bias in the graph?
11. Research the gender income gap (GIG) and report on the following:
- What initiatives have been put in place to try to eliminate the GIG in Canada?
 - Has the Canadian GIG been increasing, decreasing, or stayed constant over time?
 - How does the Canadian GIG compare to that of other countries?
- Use graphs, summary statistics, and the tools from this section to support your points.