CHAPTER 3

Statistics of Two Variables

Specific Expectations	Section
Define the correlation coefficient as a measure of the fit of a scatter graph to a linear model.	3.1, 3.2, 3.3, 3.5
Calculate the correlation coefficient for a set of data, using graphing calculators or statistical software.	3.1, 3.2, 3.3, 3.5
Demonstrate an understanding of the distinction between cause-effect relationships and the mathematical correlation between variables.	3.1, 3.2, 3.3, 3.4, 3.5
Describe possible misuses of regression.	3.2, 3.3, 3.5
Explain examples of the use and misuse of statistics in the media.	3.5
Assess the validity of conclusions made on the basis of statistical studies, by analysing possible sources of bias in the studies and by calculating and interpreting additional statistics, where possible.	3.2, 3.3, 3.4, 3.5
Demonstrate an understanding of the purpose and the use of a variety of sampling techniques.	3.4, 3.5
Organize and summarize data from secondary sources, using technology.	3.1, 3.2, 3.3, 3.4, 3.5
Locate data to answer questions of significance or personal interest, by searching well-organized databases.	3.1, 3.2, 3.4, 3.5
Use the Internet effectively as a source for databases.	3.1, 3.2, 3.4, 3.5

Chapter Problem

Job Prospects

Gina is in her second year of business studies at university and she is starting to think about a job upon graduation. She has two primary concerns—the job market and expected income. Gina does some research at the university's placement centre and finds employment statistics for graduates of her program and industry surveys of entry-level salaries.

Year	Number of Graduates	Number Hired Upon Graduation	Mean Starting Salary ($000)
1992	172	151	26
1993	180	160	27
1994	192	140	28
1995	170	147	27.5
1996	168	142	27
1997	176	155	26.5
1998	180	160	27
1999	192	162	29
2000	200	172	31
2001	220	180	34

1. How could Gina graph this data to estimate
   - **a)** her chances of finding a job in her field when she graduates in two years?
   - **b)** her starting salary?
2. What assumptions does Gina have to make for her predictions? What other factors could affect the accuracy of Gina's estimates?

This chapter introduces statistical techniques for measuring relationships between two variables. As you will see, these techniques will enable Gina to make more precise estimates of her job prospects.

Two-variable statistics have an enormous range of applications including industrial processes, medical studies, and environmental issues—in fact, almost any field where you need to determine if a change in one variable affects another.

Review of Prerequisite Skills

If you need help with any of the skills listed in purple below, refer to Appendix A.

1. **Scatter plots** For each of the following sets of data, create a scatter plot and describe any patterns you see.

a)

x	y
3	18
5	15
8	12
9	10
12	8
15	4
17	1

b)

x	y
4	6
7	2
13	17
14	5
23	19
24	11
25	30
33	21
36	29
40	39
42	26
46	32

2. **Scatter plots** For each plot in question 1,

i) graph the line of best fit and calculate its equation

ii) estimate the x- and y-intercepts

iii) estimate the value of y when $x = 7$

3. **Graphing linear equations** Determine the slope and y-intercept for the lines defined by the following equations, and then graph the lines.

a) $y = 3x - 4$ **b)** $y = -2x + 6$

c) $12x - 6y = 7$

4. **Graphing quadratic functions** Graph the following functions and estimate any x- and y-intercepts.

a) $y = 2x^2$

b) $y = x^2 + 5x - 6$

c) $y = -3x^2 + x + 2$

5. **Graphing exponential functions**

a) Identify the base and the numerical coefficient for each of the following functions.

i) $y = 0.5(3)^x$ **ii)** $y = 2^x$ **iii)** $y = 100(0.5)^x$

b) Graph each of the functions in part a).

c) Explain what happens to the value of x as the curves in part b) approach the x-axis.

6. **Sigma notation** Calculate each sum without the use of technology.

a) $\sum_{i=1}^{8} i$ **b)** $\sum_{i=1}^{5} i^2$

7. **Sigma notation** Given $\bar{x} = 2.5$, calculate each sum without the use of technology.

a) $\sum_{i=1}^{6} (i - \bar{x})$ **b)** $\sum_{i=1}^{4} (i - \bar{x})^2$

8. **Sigma notation**

a) Repeat questions 6 and 7 using appropriate technology such as a graphing calculator or a spreadsheet.

b) Explain the method that you chose.

9. **Sampling (Chapter 2)** Briefly explain each of the following terms.

a) simple random sample

b) systematic sample

c) outlier

10. **Bias (Chapter 2)**

a) Explain the term *measurement bias*.

b) Give an example of a survey method containing unintentional measurement bias.

c) Give an example of a survey method containing intentional measurement bias.

d) Give an example of sampling bias.

3.1 Scatter Plots and Linear Correlation

Does smoking cause lung cancer? Is job performance related to marks in high school? Do pollution levels affect the ozone layer in the atmosphere? Often the answers to such questions are not clear-cut, and inferences have to be made from large sets of data. Two-variable statistics provide methods for detecting relationships between variables and for developing mathematical models of these relationships.

The visual pattern in a graph or plot can often reveal the nature of the relationship between two variables.

INVESTIGATE & INQUIRE: Visualizing Relationships Between Variables

A study examines two new obedience-training methods for dogs. The dogs were randomly selected to receive from 5 to 16 h of training in one of the two training programs. The dogs were assessed using a performance test graded out of 20.

Rogers Method		Laing System	
Hours	Score	Hours	Score
10	12	8	10
15	16	6	9
7	10	15	12
12	15	16	7
8	9	9	11
5	8	11	7
8	11	10	9
16	19	10	6
10	14	8	15

1. Could you determine which of the two training systems is more effective by comparing the mean scores? Could you calculate another statistic that would give a better comparison? Explain your reasoning.
2. Consider how you could plot the data for the Rogers Method. What do you think would be the best method? Explain why.
3. Use this method to plot the data for the Rogers Method. Describe any patterns you see in the plotted data.
4. Use the same method to plot the data for the Laing System and describe any patterns you see.
5. Based on your data plots, which training method do you think is more effective? Explain your answer.

6. Did your plotting method make it easy to compare the two sets of data? Are there ways you could improve your method?

7. a) Suggest factors that could influence the test scores but have not been taken into account.

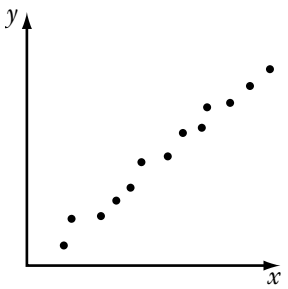
   b) How could these factors affect the validity of conclusions drawn from the data provided?

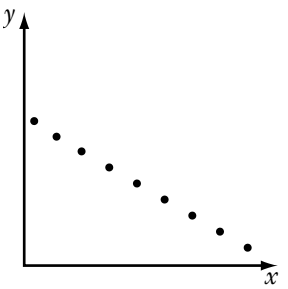
In data analysis, you are often trying to discern whether one variable, the **dependent** (or **response**) **variable**, is affected by another variable, the **independent** (or **explanatory**) **variable**. Variables have a **linear correlation** if changes in one variable tend to be proportional to changes in the other. Variables X and Y have a **perfect positive** (or **direct**) **linear correlation** if Y increases at a constant rate as X increases. Similarly, X and Y have a **perfect negative** (or **inverse**) **linear correlation** if Y decreases at a constant rate as X increases.

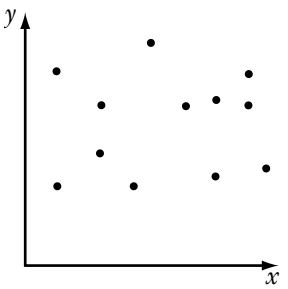
A **scatter plot** shows such relationships graphically, usually with the independent variable as the horizontal axis and the dependent variable as the vertical axis. The **line of best fit** is the straight line that passes as close as possible to all of the points on a scatter plot. The stronger the correlation, the more closely the data points cluster around the line of best fit.

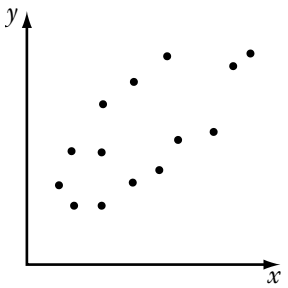
Example 1 Classifying Linear Correlations

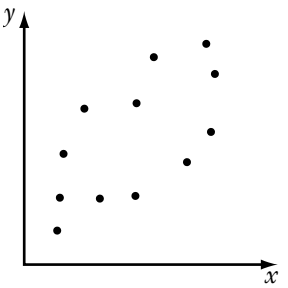
Classify the relationship between the variables X and Y for the data shown in the following diagrams.

a) 

b) 

c) 

d) 

e) 

f) 

Solution

a) The data points are clustered around a line that rises to the right (positive slope), indicating definitely that Y increases as X increases. Although the points are not perfectly lined up, there is a *strong positive linear correlation* between X and Y.

b) The data points are all exactly on a line that slopes down to the right, so Y decreases as X increases. In fact, the changes in Y are *exactly* proportional to the changes in X. There is a *perfect negative linear correlation* between X and Y.

c) No discernible linear pattern exists. As X increases, Y appears to change randomly. Therefore, there is *zero linear correlation* between X and Y.

d) A definite positive trend exists, but it is not as clear as the one in part a). Here, X and Y have a *moderate positive linear correlation*.

e) A slight positive trend exists. X and Y have a *weak positive linear correlation*.

f) A definite negative trend exists, but it is hard to classify at a glance. Here, X and Y have a *moderate or strong negative linear correlation*.

As Example 1 shows, a scatter plot often can give only a rough indication of the correlation between two variables. Obviously, it would be useful to have a more precise way to measure correlation. Karl Pearson (1857–1936) developed a formula for estimating such a measure. Pearson, who also invented the term *standard deviation*, was a key figure in the development of modern statistics.

The Correlation Coefficient

To develop a measure of correlation, mathematicians first defined the **covariance** of two variables in a sample:

$$s_{XY} = \frac{1}{n-1}\sum (x-\overline{x})(y-\overline{y})$$

where n is the size of the sample, x represents individual values of the variable X, y represents individual values of the variable Y, $\overline{x}$ is the mean of X, and $\overline{y}$ is the mean of Y.

Recall from Chapter 2 that the symbol $\sum$ means "the sum of." Thus, the covariance is the sum of the *products* of the deviations of x and y for all the data points divided by $n - 1$. The covariance depends on how the deviations of the two variables are related. For example, the covariance will have a large positive value if both $x - \overline{x}$ and $y - \overline{y}$ tend to be large at the same time, and a negative value if one tends to be positive when the other is negative.

The **correlation coefficient**, ***r***, is the covariance divided by the product of the standard deviations for X and Y:

$$r = \frac{s_{XY}}{s_X \times s_Y}$$

where s_X is the standard deviation of X and s_Y is the standard deviation of Y.

This coefficient gives a quantitative measure of the strength of a linear correlation. In other words, the correlation coefficient indicates how closely the data points cluster around the line of best fit. The correlation coefficient is also called the **Pearson product-moment coefficient of correlation (PPMC)** or **Pearson's *r***.

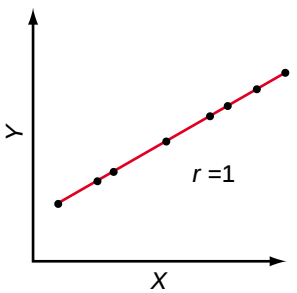
The correlation coefficient always has values in the range from –1 to 1. Consider a perfect positive linear correlation first. For such correlations, changes in the dependent variable Y are directly proportional to changes in the independent variable X, so $Y = aX + b$, where a is a positive constant. It follows that

$$\begin{aligned} s_{XY} &= \frac{1}{n-1}\sum (x-\bar{x})(y-\bar{y}) \\ &= \frac{1}{n-1}\sum (x-\bar{x})[(ax+b)-(a\bar{x}+b)] \\ &= \frac{1}{n-1}\sum (x-\bar{x})(ax-a\bar{x}) \\ &= \frac{1}{n-1}\sum a(x-\bar{x})^2 \\ &= a\frac{\sum (x-\bar{x})^2}{n-1} \\ &= as_X^2 \end{aligned}$$

$$\begin{aligned} s_Y &= \sqrt{\frac{\sum (y-\bar{y})^2}{n-1}} \\ &= \sqrt{\frac{\sum [(ax+b)-(a\bar{x}+b)]^2}{n-1}} \\ &= \sqrt{\frac{\sum (ax-a\bar{x})^2}{n-1}} \\ &= \sqrt{\frac{a^2\sum (x-\bar{x})^2}{n-1}} \\ &= a\sqrt{\frac{\sum (x-\bar{x})^2}{n-1}} \\ &= as_X \end{aligned}$$

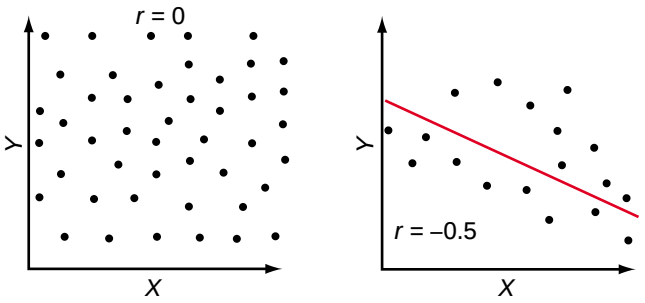
Substituting into the equation for the correlation coefficient gives

$$\begin{aligned} r &= \frac{s_{XY}}{s_X s_Y} \\ &= \frac{as_X^2}{s_X(as_X)} \\ &= 1 \end{aligned}$$

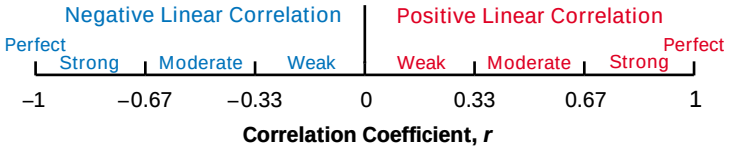


Similarly, $r = -1$ for a perfect negative linear correlation.

For two variables with no correlation, Y is equally likely to increase or decrease as X increases. The terms in $\sum(x-\bar{x})(y-\bar{y})$ are randomly positive or negative and tend to cancel each other. Therefore, the correlation coefficient is close to zero if there is little or no correlation between the variables. For moderate linear correlations, the summation terms partially cancel out.



The following diagram illustrates how the correlation coefficient corresponds to the strength of a linear correlation.



Using algebraic manipulation and the fact that $\sum x = n\bar{x}$, Pearson showed that

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

where n is the number of data points in the sample, x represents individual values of the variable X, and y represents individual values of the variable Y. (Note that $\sum x^2$ is the sum of the squares of all the individual values of X, while $(\sum x)^2$ is the square of the sum of all the individual values.)

Like the alternative formula for standard deviations (page 150), this formula for r avoids having to calculate all the deviations individually. Many scientific and statistical calculators have built-in functions for calculating the correlation coefficient.

It is important to be aware that increasing the number of data points used in determining a correlation improves the accuracy of the mathematical model. Some of the examples and exercise questions have a fairly small set of data in order to simplify the computations. Larger data sets can be found in the e-book that accompanies this text.

Example 2 Applying the Correlation Coefficient Formula

A farmer wants to determine whether there is a relationship between the mean temperature during the growing season and the size of his wheat crop. He assembles the following data for the last six crops.

Mean Temperature (°C)	Yield (tonnes / hectare)
4	1.6
8	2.4
10	2.0
9	2.6
11	2.1
6	2.2

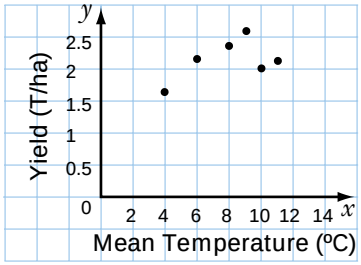
a) Does a scatter plot of these data indicate any linear correlation between the two variables?

b) Compute the correlation coefficient.

c) What can the farmer conclude about the relationship between the mean temperatures during the growing season and the wheat yields on his farm?

Solution

a) The farmer wants to know whether the crop yield depends on temperature. Here, temperature is the independent variable, X, and crop yield is the dependent variable, Y. The scatter plot has a somewhat positive trend, so there appears to be a moderate positive linear correlation.



b) To compute r, set up a table to calculate the quantities required by the formula.

Temperature, x	Yield, y	x^2	y^2	xy
4	1.6	16	2.56	6.4
8	2.4	64	5.76	19.2
10	2.0	100	4.00	20.0
9	2.6	81	6.76	23.4
11	2.1	121	4.41	23.1
6	2.2	36	4.84	13.2
$\sum x = 48$	$\sum y = 12.9$	$\sum x^2 = 418$	$\sum y^2 = 28.33$	$\sum xy = 105.3$

Now compute r, using the formula:

$$r = \frac{n\Sigma(xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$
$$= \frac{6(105.3) - (48)(12.9)}{\sqrt{[6(418) - (48)^2][6(28.33) - (12.9)^2]}}$$
$$= \frac{631.8 - 619.2}{\sqrt{(2508 - 2304)(169.98 - 166.41)}}$$
$$= \frac{12.6}{26.99}$$
$$= 0.467$$

The correlation coefficient for crop yield versus mean temperature is approximately 0.47, which confirms a moderate positive linear correlation.

c) It appears that the crop yield tends to increase somewhat as the mean temperature for the growing season increases. However, the farmer cannot conclude that higher temperatures *cause* greater crop yields. Other variables could account for the correlation. For example, the lower temperatures could be associated with heavy rains, which could lower yields by flooding fields or leaching nutrients from the soil.

Data in Action

From 1992 to 2001, Canada produced an average of 27 million tonnes of wheat a year. About 70% of this crop was exported.

The important principle that a correlation does not prove the existence of a cause-and-effect relationship between two variables is discussed further in section 3.4.

Example 3 Using Technology to Determine Correlation Coefficients

Determine whether there is a linear correlation between horsepower and fuel consumption for these five vehicles by creating a scatter plot and calculating the correlation coefficient.

Vehicle	Horsepower, x	Fuel Consumption (L/100 km), y
Midsize sedan	105	6.7
Minivan	170	23.5
Small sports utility vehicle	124	5.9
Midsize motorcycle	17	3.4
Luxury sports car	296	8.4

Solution 1 Using a Graphing Calculator

Use the ClrList command to make sure lists L1 and L2 are clear, then enter the horsepower data in L1 and the fuel consumption figures in L2.

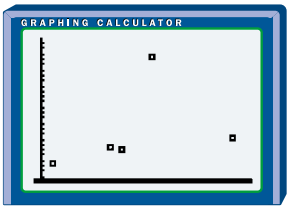
To display a scatter plot, first make sure that all functions in the Y= editor are either clear or turned off. Then, use STAT PLOT to select PLOT1.

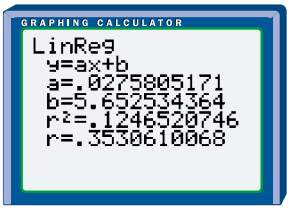
See Appendix B for more details on the graphing calculator and software functions used in this section.

Turn the plot on, select the scatter-plot icon, and enter L1 for XLIST and L2 for YLIST. (Some of these settings may already be in place.) From the ZOOM menu, select 9:ZoomStat. The calculator will automatically optimize the **window settings** and display the scatter plot.

To calculate the correlation coefficient, from the CATALOG menu, select **DiagnosticOn**, then select the **LinReg(ax+b) instruction** from the STAT CALC menu. The calculator will perform a series of statistical calculations using the data in lists L1 and L2. The last line on the screen shows that the correlation coefficient is approximately 0.353.

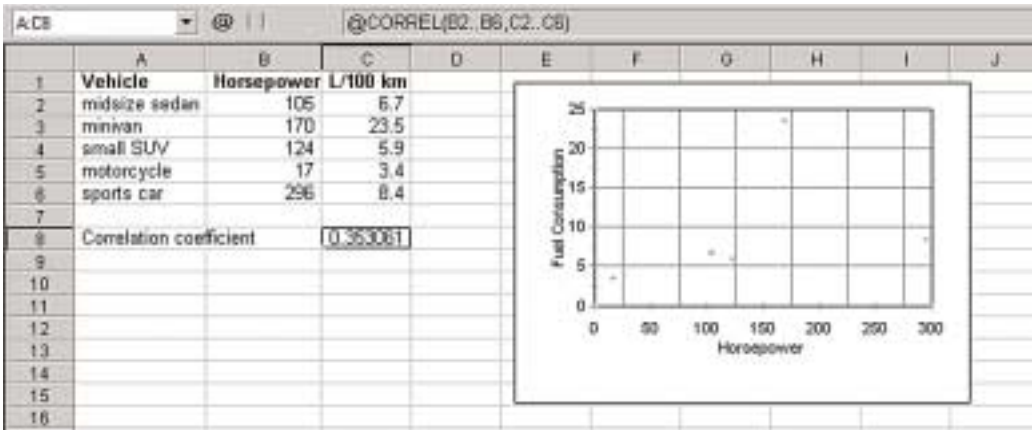
Therefore, there is a moderate linear correlation between horsepower and fuel consumption for the five vehicles.





Solution 2 Using a Spreadsheet

Set up three columns and enter the data from the table above. Highlight the numerical data and use your spreadsheet's **Chart feature** to display a scatter plot. Both Corel® Quattro® Pro and Microsoft® Excel have a **CORREL function** that allows you to calculate the correlation coefficient easily. The scatter plot and correlation coefficient indicate a moderate correlation between horsepower and fuel consumption.

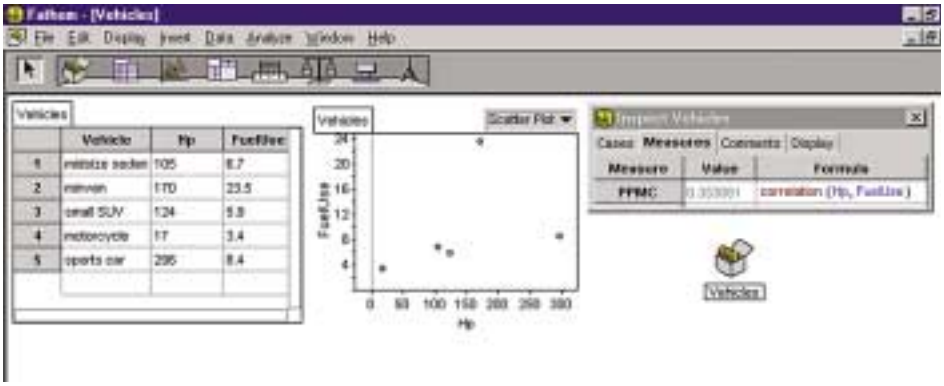


Solution 3 Using Fathom™

Create a new **collection** by setting up a **case table** with three attributes: Vehicle, Hp, and FuelUse. Enter the data for the five cases. To create a **scatter plot**, drag the **graph icon** onto the work area and drop the Hp attribute on the *x*-axis and the FuelUse attribute on the *y*-axis.

To calculate the **correlation coefficient**, right-click on the **collection** and select Inspect Collection. Select the Measures tab and name a new measure PPMC. Right-click this measure and select Edit Formula, then Functions / Statistical / Two Attributes / correlation. When you enter the Hp and FuelUse attributes in the correlation function, Fathom™ will calculate the correlation coefficient for these data.

Again, the scatter plot and correlation coefficient show a moderate linear correlation.



Project Prep

For your statistics project, you may be investigating the linear correlation between two variables. A graphing calculator or computer software may be a valuable aid for this analysis.

Notice that the scatter plots in Example 3 have an outlier at (170, 23.5). Without this data point, you would have a strong positive linear correlation. Section 3.2 examines the effect of outliers in more detail.

Key Concepts

- Statistical studies often find linear correlations between two variables.
- A scatter plot can often reveal the relationship between two variables. The independent variable is usually plotted on the horizontal axis and the dependent variable on the vertical axis.
- Two variables have a linear correlation if changes in one variable tend to be proportional to changes in the other. Linear correlations can be positive or negative and vary in strength from zero to perfect.
- The correlation coefficient, r, is a quantitative measure of the correlation between two variables. Negative values indicate negative correlations while positive values indicate positive correlations. The greater the absolute value of r, the stronger the linear correlation, with zero indicating no correlation at all and 1 indicating a perfect correlation.
- Manual calculations of correlation coefficients can be quite tedious, but a variety of powerful technology tools are available for such calculations.

Communicate Your Understanding

1. Describe the advantages and disadvantages of using a scatter plot or the correlation coefficient to estimate the strength of a linear correlation.

2. **a)** What is the meaning of a correlation coefficient of
   - **i)** -1?
   - **ii)** 0?
   - **iii)** 0.5?

   b) Can the correlation coefficient have a value greater than 1? Why or why not?

3. A mathematics class finds a correlation coefficient of 0.25 for the students' midterm marks and their driver's test scores and a coefficient of -0.72 for their weight-height ratios and times in a 1-km run. Which of these two correlations is stronger? Explain your answer.

Practise

A

1. Classify the type of linear correlation that you would expect with the following pairs of variables.
   - **a)** hours of study, examination score
   - **b)** speed in excess of the speed limit, amount charged on a traffic fine
   - **c)** hours of television watched per week, final mark in calculus
   - **d)** a person's height, sum of the digits in the person's telephone number
   - **e)** a person's height, the person's strength

2. Identify the independent variable and the dependent variable in a correlational study of
   - **a)** heart disease and cholesterol level
   - **b)** hours of basketball practice and free-throw success rate
   - **c)** amount of fertilizer used and height of plant
   - **d)** income and level of education
   - **e)** running speed and pulse rate

Apply, Solve, Communicate

B

3. For a week prior to their final physics examination, a group of friends collect data to see whether time spent studying or time spent watching TV had a stronger correlation with their marks on the examination.

Hours Studied	Hours Watching TV	Examination Score
10	8	72
11	7	67
15	4	81
14	3	93
8	9	54
5	10	66

   - **a)** Create a scatter plot of hours studied versus examination score. Classify the linear correlation.
   - **b)** Create a similar scatter plot for the hours spent watching TV.
   - **c)** Which independent variable has a stronger correlation with the examination scores? Explain.

d) Calculate the correlation coefficient for hours studied versus examination score and for hours watching TV versus examination score. Do these answers support your answer to c)? Explain.

4. **Application** Refer to the tables in the investigation on page 159.

a) Determine the correlation coefficient and classify the linear correlation for the data for each training method.

b) Suppose that you interchanged the dependent and independent variables, so that the test scores appear on the horizontal axis of a scatter plot and the hours of training appear on the vertical axis. Predict the effect this change will have on the scatter plot and the correlation coefficient for each set of data.

c) Test your predictions by plotting the data and calculating the correlation coefficients with the variables reversed. Explain any differences between your results and your predictions in part b).

5. A company studied whether there was a relationship between its employees' years of service and number of days absent. The data for eight randomly selected employees are shown below.

Employee	Years of Service	Days Absent Last Year
Jim	5	2
Leah	2	6
Efraim	7	3
Dawn	6	3
Chris	4	4
Cheyenne	8	0
Karrie	1	2
Luke	10	1

a) Create a scatter plot for these data and classify the linear correlation.

b) Calculate the correlation coefficient.

c) Does the computed r-value agree with the classification you made in part a)? Explain why or why not.

d) Identify any outliers in the data.

e) Suppose possible reasons for any outliers identified in part d).

6. **Application** Six classmates compared their arm spans and their scores on a recent mathematics test as shown in the following

Arm Span (m)	Score
1.5	82
1.4	71
1.7	75
1.6	66
1.6	90
1.8	73

a) Illustrate these data with a scatter plot.

b) Determine the correlation coefficient and classify the linear correlation.

c) What can the students conclude from their data?

7. **a)** Use data in the table on page 157 to create a scatter plot that compares the size of graduating classes in Gina's program to the number of graduates who found jobs.

b) Classify the linear correlation.

c) Determine the linear correlation coefficient.

8. **a)** Search sources such as E-STAT, CANSIM II, the Internet, newspapers, and magazines for pairs of variables that exhibit

i) a strong positive linear correlation

ii) a strong negative linear correlation

iii) a weak or zero linear correlation

b) For each pair of variables in part a), identify the independent variable and the dependent variable.

9. Find a set of data for two variables known to have a perfect positive linear correlation. Use these data to demonstrate that the correlation coefficient for such variables is 1. Alternatively, find a set of data with a perfect negative correlation and show that the correlation coefficient is −1.

10. **Communication**

a) Would you expect to see a correlation between the temperature at an outdoor track and the number of people using the track? Why or why not?

b) Sketch a typical scatter plot of this type of data.

c) Explain the key features of your scatter plot.

11. **Inquiry/Problem Solving** Refer to data tables in the investigation on page 159.

a) How could the Rogers Training Company graph the data so that their training method looks particularly good?

b) How could Laing Limited present the same data in a way that favours their training system?

c) How could a mathematically knowledgeable consumer detect the distortions in how the two companies present the data?

C

12. **Inquiry/Problem Solving**

a) Prove that interchanging the independent and dependent variables does not change the correlation coefficient for any set of data.

b) Illustrate your proof with calculations using a set of data selected from one of the examples or exercise questions in this section.

13. **a)** Search sources such as newspapers, magazines, and the Internet for a set of two-variable data with

i) a moderate positive linear correlation

ii) a moderate negative correlation

iii) a correlation in which $|r| > 0.9$

b) Outline any conclusions that you can make from each set of data. Are there any assumptions inherent in these conclusions? Explain.

c) Pose at least two questions that could form the basis for further research.

14. **a)** Sketch scatter plots of three different patterns of data that you think would have zero linear correlation.

b) Explain why r would equal zero for each of these patterns.

c) Use Fathom™ or a spreadsheet to create a scatter plot that looks like one of your patterns and calculate the correlation coefficient. Adjust the data points to get r as close to zero as you can.

3.2 Linear Regression

Regression is an analytic technique for determining the relationship between a dependent variable and an independent variable. When the two variables have a linear correlation, you can develop a simple mathematical model of the relationship between the two variables by finding a line of best fit. You can then use the equation for this line to make predictions by **interpolation** (estimating between data points) and **extrapolation** (estimating beyond the range of the data).

INVESTIGATE & INQUIRE: Modelling a Linear Relationship

A university would like to construct a mathematical model to predict first-year marks for incoming students based on their achievement in grade 12. A comparison of these marks for a random sample of first-year students is shown below.

Grade 12 Average	85	90	76	78	88	84	76	96	86	85
First-Year Average	74	83	68	70	75	72	64	91	78	86

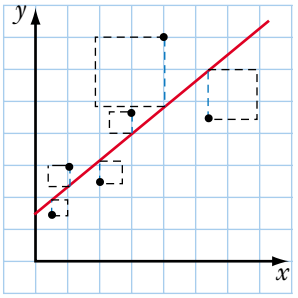
1. a) Construct a scatter plot for these data. Which variable should be placed on the vertical axis? Explain.
   b) Classify the linear correlation for this data, based on the scatter plot.
2. a) Estimate and draw a line of best fit for the data.
   b) Measure the slope and y-intercept for this line, and write an equation for it in the form $y = mx + b$.
3. Use this linear model to predict
   a) the first-year average for a student who had an 82 average in grade 12
   b) the grade-12 average for a student with a first-year average of 60
4. a) Use software or the linear regression instruction of a graphing calculator to find the slope and y-intercept for the line of best fit. (Note that most graphing calculators use a instead of m to represent slope.)
   b) Are this slope and y-intercept close to the ones you measured in question 2? Why or why not?

c) Estimate how much the new values for slope and y-intercept will change your predictions in question 3. Check your estimate by recalculating your predictions using the new values and explain any discrepancies.

5. List the factors that could affect the accuracy of these mathematical models. Which factor do you think is most critical? How could you test how much effect this factor could have?

It is fairly easy to "eyeball" a good estimate of the line of best fit on a scatter plot when the linear correlation is strong. However, an analytic method using a **least-squares fit** gives more accurate results, especially for weak correlations.

Consider the line of best fit in the following scatter plot. A dashed blue line shows the **residual** or vertical deviation of each data point from the line of best fit. The residual is the difference between the values of y at the data point and at the point that lies on the line of best fit and has the same x-coordinate as the data point. Notice that the residuals are positive for points above the line and negative for points below the line. The boxes show the squares of the residuals.



For the line of best fit in the least-squares method,

- the sum of the residuals is zero (the positive and negative residuals cancel out)
- the sum of the squares of the residuals has the least possible value

Although the algebra is daunting, it can be shown that this line has the equation

$$y = ax + b, \text{ where } a = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2} \text{ and } b = \overline{y} - a\overline{x}$$

Recall from Chapter 2 that $\overline{x}$ is the mean of x and $\overline{y}$ is the mean of y. Many statistics texts use an equation with the form $y = a + bx$, so you may sometimes see the equations for a and b reversed.

Example 1 Applying the Least-Squares Formula

This table shows data for the full-time employees of a small company.

Age (years)	Annual Income ($000)
33	33
25	31
19	18
44	52
50	56
54	60
38	44
29	35

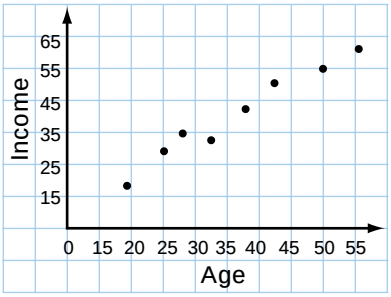
a) Use a scatter plot to classify the correlation between age and income.

b) Find the equation of the line of best fit analytically.

c) Predict the income for a new employee who is 21 and an employee retiring at age 65.

Solution

a) The scatter plot suggests a strong positive linear correlation between age and income level.



b) To determine the equation of the line of best fit, organize the data into a table and compute the sums required for the formula.

Age, x	Income, y	x^2	xy
33	33	1089	1089
25	31	625	775
19	18	361	342
44	52	1936	2288
50	56	2500	2800
54	60	2916	3240
38	44	1444	1672
29	35	841	1015
$\sum x = 292$	$\sum y = 329$	$\sum x^2 = 11\,712$	$\sum xy = 13\,221$

Substitute these totals into the formula for a.

$$a = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

$$= \frac{8(13\,221) - (292)(329)}{8(11\,712) - (292)^2}$$

$$= \frac{9700}{8432}$$

$$\doteq 1.15$$

To determine b, you also need the means of x and y.

$$\bar{x} = \frac{\Sigma x}{n} \qquad \bar{y} = \frac{\Sigma y}{n} \qquad b = \bar{y} - a\bar{x}$$
$$= \frac{292}{8} \qquad = \frac{329}{8} \qquad = 41.125 - 1.15(36.5)$$
$$= -0.85$$
$$= 36.5 \qquad = 41.125$$

Now, substitute the values of a and b into the equation for the line of best fit.

$y = ax + b$
$= 1.15x - 0.85$

Therefore, the equation of the line of best fit is $y = 1.15x - 0.85$.

c) Use the equation of the line of best fit as a model.

For a 21-year-old employee,
$y = ax + b$
$= 1.15(21) - 0.85$
$= 23.3$

For a 65-year-old employee,
$y = ax + b$
$= 1.15(65) - 0.85$
$= 73.9$

Therefore, you would expect the new employee to have an income of about \$23 300 and the retiring employee to have an income of about \$73 900. Note that the second estimate is an extrapolation beyond the range of the data, so it could be less accurate than the first estimate, which is interpolated between two data points.

Note that the slope a indicates only how y varies with x on the line of best fit. The slope does not tell you anything about the strength of the correlation between the two variables. It is quite possible to have a weak correlation with a large slope or a strong correlation with a small slope.

Example 2 Linear Regression Using Technology

Researchers monitoring the numbers of wolves and rabbits in a wildlife reserve think that the wolf population depends on the rabbit population since wolves prey on rabbits. Over the years, the researchers collected the following data.

Year	1994	1995	1996	1997	1998	1999	2000	2001
Rabbit Population	61	72	78	76	65	54	39	43
Wolf Population	26	33	42	49	37	30	24	19

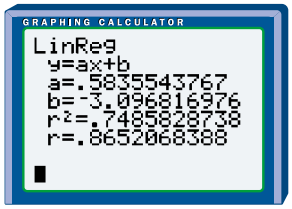
a) Determine the line of best fit and the correlation coefficient for these data.

b) Graph the data and the line of best fit. Do these data support the researchers' theory?

Solution 1 Using a Graphing Calculator

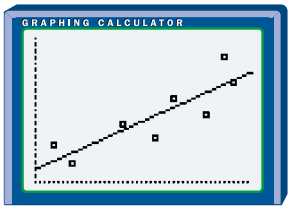
a) You can use the calculator's **linear regression instruction** to find both the line of best fit and the correlation coefficient. Since the theory is that the wolf population depends on the rabbit population, the rabbit population is the independent variable and the wolf population is the dependent variable.

Use the STAT EDIT menu to enter the rabbit data into list L1 and the wolf data into L2. Set **DiagnosticOn**, and then use the STAT CALC menu to select **LinReg(ax+b)**.



The equation of the line of best fit is $y = 0.58x - 3.1$ and the correlation coefficient is 0.87.

b) Store the equation for the line of best fit as a function, Y1. Then, use the STAT PLOT menu to set up the scatter plot. By displaying both Y1 and the scatter plot, you can see how closely the data plots are distributed around the line of best fit.



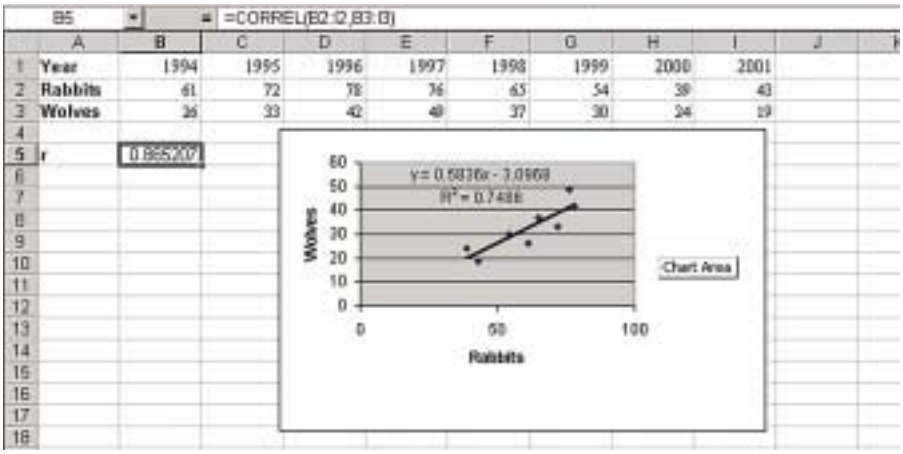
The correlation coefficient and the scatter plot show a strong positive linear correlation between the variables. This correlation supports the researchers' theory, but does not prove that changes in the rabbit population are the cause of the changes in the wolf population.

Solution 2 Using a Spreadsheet

Set up a table with the data for the rabbit and wolf populations. You can calculate the correlation coefficient with the **CORREL function**. Use the **Chart feature** to create a scatter plot.

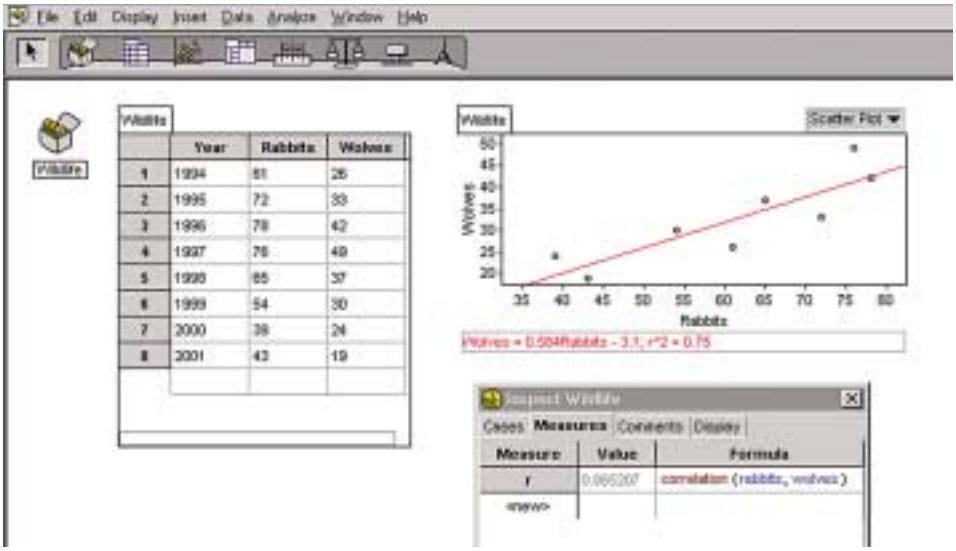
In Corel® Quattro® Pro, you can find the equation of the **line of best fit** by selecting Tools/Numeric Tools/Regression. Enter the cell ranges for the data, and the program will display regression calculations including the constant (b), the x-coefficient (or slope, a), and r^2.

In Microsoft® Excel, you can find the equation of the **line of best fit** by selecting Chart/Add Trendline. Check that the default setting is Linear. Select the straight line that appears on your chart, then click Format/Selected Trendline/Options. Check the Display equation on chart box. You can also display r^2.



Solution 3 Using Fathom™

Drag a new **case table** to the workspace, create attributes for Year, Rabbits, and Wolves, and enter the data. Drag a new **graph** to the workspace, then drag the Rabbits attribute to the x-axis and the Wolves attribute to the y-axis. From the Graph menu, select Least Squares Line. Fathom™ will display r^2 and the equation for the **line of best fit**. To calculate the correlation coefficient directly, select Inspect Collection, click the Measures tab, then create a new measure by selecting Functions/Statistical/Two Attributes/correlation and entering Rabbits and Wolves as the attributes.



Project Prep

When analysing two-variable data for your statistics project, you may wish to develop a linear model, particularly if a strong linear correlation is evident.

In Example 2, the sample size is small, so you should be cautious about making generalizations from it. Small samples have a greater chance of not being representative of the whole population. Also, outliers can seriously affect the results of a regression on a small sample.

Example 3 The Effect of Outliers

To evaluate the performance of one of its instructors, a driving school tabulates the number of hours of instruction and the driving-test scores for the instructor's students.

Instructional Hours	10	15	21	6	18	20	12
Student's Score	78	85	96	75	84	45	82

a) What assumption is the management of the driving school making? Is this assumption reasonable?

b) Analyse these data to determine whether they suggest that the instructor is an effective teacher.

c) Comment on any data that seem unusual.

d) Determine the effect of any outliers on your analysis.

Solution

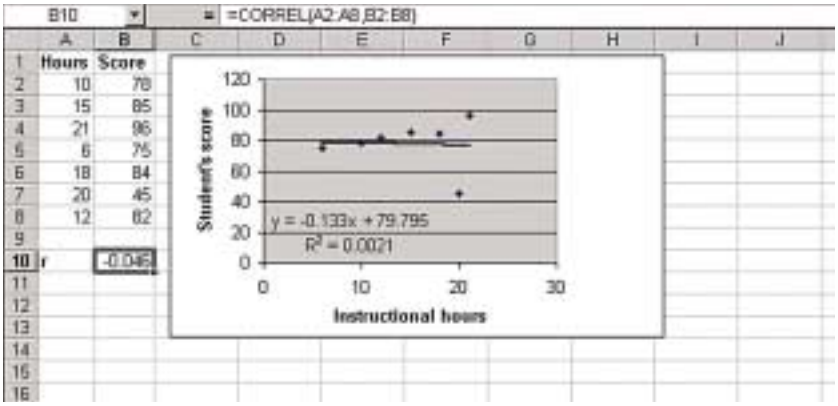
a) The management of the driving school is assuming that the correlation between instructional hours and test scores is an indication of the instructor's teaching skills. Such a relationship could be difficult to prove definitively. However, the assumption would be reasonable if the driving school has found that some instructors have consistently strong correlations between the time spent with their students and the students' test scores while other instructors have consistently weaker correlations.

b) The number of hours of instruction is the independent variable. You could analyse the data using any of the methods in the two previous examples. For simplicity, a spreadsheet solution is shown here.

Except for an obvious outlier at (20, 45), the scatter plot below indicates a strong positive linear correlation. At first glance, it appears that the number of instructional hours is positively correlated to the students' test scores. However, the linear regression analysis yields a line of best fit with the equation $y = -0.13x + 80$ and a correlation coefficient of -0.05.

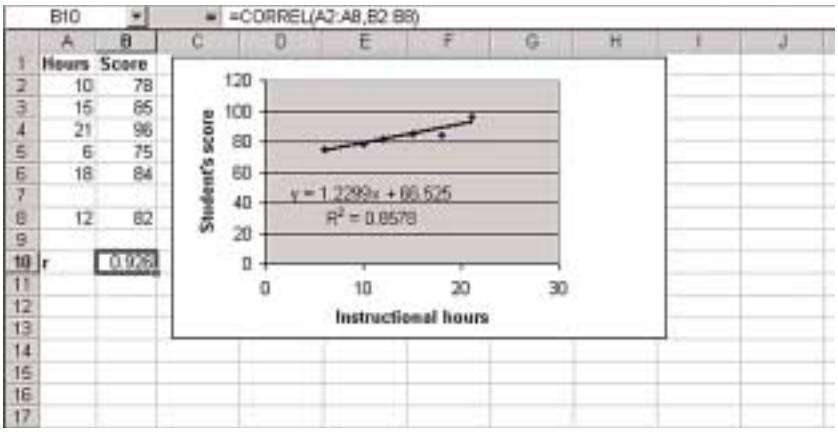
These results indicate that there is virtually a zero linear correlation, and the line of best fit even has a negative slope! The outlier has a dramatic impact on the regression results because it is distant from the other data points and the sample size is quite small. Although the scatter plot looked

favourable, the regression analysis suggests that the instructor's lessons had no positive effect on the students' test results.



c) The fact that the outlier is substantially below all the other data points suggests that some special circumstance may have caused an abnormal result. For instance, there might have been an illness or emotional upset that affected this one student's performance on the driving test. In that case, it would be reasonable to exclude this data point when evaluating the driving instructor.

d) Remove the outlier from your data table and repeat your analysis.



Notice that the line of best fit is now much closer to the data points and has a positive slope. The correlation coefficient, r, is 0.93, indicating a strong positive linear correlation between the number of instructional hours and the driver's test scores. This result suggests that the instructor may be an effective teacher after all. It is quite possible that the original analysis was not a fair evaluation. However, to do a proper evaluation, you would need a larger set of data, more information about the outlier, or, ideally, both.

As Example 3 demonstrates, outliers can skew a regression analysis, but they could also simply indicate that the data really do have large variations. A comprehensive analysis of a set of data should look for outliers, examine their possible causes and their effect on the analysis, and discuss whether they should be excluded from the calculations. As you observed in Chapter 2, outliers have less effect on larger samples.

Project Prep

If your statistics project involves a linear relationship that contains outliers, you will need to consider carefully their impact on your results, and how you will deal with them.

WEB CONNECTION

www.mcgrawhill.ca/links/MDM12

Visit the above web site and follow the links to learn more about linear regression. Describe an application of linear regression that interests you.

Key Concepts

- Linear regression provides a means for analytically determining a line of best fit. In the least-squares method, the line of best fit is the line which minimizes the sum of the squares of the residuals while having the sum of the residuals equal zero.
- You can use the equation of the line of best fit to predict the value of one of the two variables given the value of the other variable.
- The correlation coefficient is a measure of how well a regression line fits a set of data.
- Outliers and small sample sizes can reduce the accuracy of a linear model.

Communicate Your Understanding

1. What does the correlation coefficient reveal about the line of best fit generated by a linear regression?
2. Will the correlation coefficient always be negative when the slope of the line of best fit is negative? Explain your reasoning.
3. Describe the problem that outliers present for a regression analysis and outline what you could do to resolve this problem.

Practise

A

1. Identify any outliers in the following sets of data and explain your choices.

a)

X	25	34	43	55	92	105	16
Y	30	41	52	66	18	120	21

b)

X	5	7	6	6	4	8
Y	304	99	198	205	106	9

2. **a)** Perform a linear regression analysis to generate the line of best fit for each set of data in question 1.

b) Repeat the linear regressions in part a), leaving out any outliers.

c) Compare the lines of best fit in parts a) and b).

Apply, Solve, Communicate

B

3. Use the formula for the method of least squares to verify the slope and intercept values you found for the data in the investigation on page 171. Account for any discrepancies.

4. Use software or a graphing calculator to verify the regression results in Example 1.

5. Application The following table lists the heights and masses for a group of fire-department trainees.

Height (cm)	Mass (kg)
177	91
185	88
173	82
169	79
188	87
182	85
175	79

a) Create a scatter plot and classify the linear correlation.

b) Apply the method of least squares to generate the equation of the line of best fit.

c) Predict the mass of a trainee whose height is 165 cm.

d) Predict the height of a 79-kg trainee.

e) Explain any discrepancy between your answer to part d) and the actual height of the 79-kg trainee in the sample group.

6. A random survey of a small group of high-school students collected information on the students' ages and the number of books they had read in the past year.

Age (years)	Books Read
16	5
15	3
18	8
17	6
16	4
15	4
14	5
17	15

a) Create a scatter plot for this data. Classify the linear correlation.

b) Determine the correlation coefficient and the equation of the line of best fit.

c) Identify the outlier.

d) Repeat part b) with the outlier excluded.

e) Does removing the outlier improve the linear model? Explain.

f) Suggest other ways to improve the model.

g) Do your results suggest that the number of books a student reads depends on the student's age? Explain.

7. **Application** Market research has provided the following data on the monthly sales of a licensed T-shirt for a popular rock band.

Price ($)	Monthly Sales
10	2500
12	2200
15	1600
18	1200
20	800
24	250

a) Create a scatter plot for these data.

b) Use linear regression to model these data.

c) Predict the sales if the shirts are priced at $19.

d) The vendor has 1500 shirts in stock and the band is going to finish its concert tour in a month. What is the maximum price the vendor can charge and still avoid having shirts left over when the band stops touring?

8. **Communication** MDM Entertainment has produced a series of TV specials on the lives of great mathematicians. The executive producer wants to know if there is a linear correlation between production costs and revenue from the sales of broadcast rights. The costs and gross sales revenue for productions in 2001 and 2002 were as follows (amounts in millions of dollars).

2001		2002	
Cost ($M)	Sales ($M)	Cost ($M)	Sales ($M)
5.5	15.4	2.7	5.2
4.1	12.1	1.9	1.0
1.8	6.9	3.4	3.4
3.2	9.4	2.1	1.9
4.2	1.5	1.4	1.5

a) Create a scatter plot using the data for the productions in 2001. Do there appear to be any outliers? Explain.

b) Determine the correlation coefficient and the equation of the line of best fit.

c) Repeat the linear regression analysis with any outliers removed.

d) Repeat parts a) and b) using the data for the productions in 2002.

e) Repeat parts a) and b) using the combined data for productions in both 2001 and 2002. Do there still appear to be any outliers?

f) Which of the four linear equations do you think is the best model for the relationship between production costs and revenue? Explain your choice.

g) Explain why the executive producer might choose to use the equation from part d) to predict the income from MDM's 2003 productions.

9. At Gina's university, there are 250 business students who expect to graduate in 2006.

a) Model the relationship between the total number of graduates and the number hired by performing a linear regression on the data in the table on page 157. Determine the equation of the line of best fit and the correlation coefficient.

b) Use this linear model to predict how many graduates will be hired in 2006.

c) Identify any outliers in this scatter plot and suggest possible reasons for an outlier. Would any of these reasons justify excluding the outlier from the regression calculations?

d) Repeat part a) with the outlier removed.

e) Compare the results in parts a) and d). What assumptions do you have to make?

10. **Communication** Refer to Example 2, which describes population data for wolves and rabbits in a wildlife reserve. An alternate theory has it that the rabbit population depends on the wolf population since the wolves prey on the rabbits.

a) Create a scatter plot of rabbit population versus wolf population and classify the linear correlation. How are your data points related to those in Example 2?

b) Determine the correlation coefficient and the equation of the line of best fit. Graph this line on your scatter plot.

c) Is the equation of the line of best fit the inverse of that found in Example 2? Explain.

d) Plot both populations as a time series. Can you recognize a pattern or relationship between the two series? Explain.

e) Does the time series suggest which population is the dependent variable? Explain.

11. The following table lists the mathematics of data management marks and grade 12 averages for a small group of students.

Mathematics of Data Management Mark	Grade 12 Average
74	77
81	87
66	68
53	67
92	85
45	55
80	76

a) Using Fathom™ or *The Geometer's Sketchpad*®,

i) create a scatter plot for these data

ii) add a moveable line to the scatter plot and construct the geometric square for the deviation of each data point from the moveable line

iii) generate a dynamic sum of the areas of these squares

iv) manoeuvre the moveable line to the position that minimizes the sum of the areas of the squares.

v) record the equation of this line

b) Determine the equation of the line of best fit for this set of data.

c) Compare the equations you found in parts a) and b). Explain any differences or similarities.

12. **Application** Use E-STAT or other sources to obtain the annual consumer price index figures from 1914 to 2000.

a) Download this information into a spreadsheet or statistical software, or enter it into a graphing calculator. (If you use a graphing calculator, enter the data from every third year.) Find the line of best fit and comment on whether a straight line appears to be a good model for the data.

b) What does the slope of the line of best fit tell you about the rate of inflation?

c) Find the slope of the line of best fit for the data for just the last 20 years, and then repeat the calculation using only the data for the last 5 years.

d) What conclusions can you make by comparing the three slopes? Explain your reasoning.

ACHIEVEMENT CHECK

Knowledge / Understanding	Thinking / Inquiry / Problem Solving	Communication	Application

13. The Worldwatch Institute has collected the following data on concentrations of carbon dioxide (CO_2) in the atmosphere.

Year	CO_2 Level (ppm)
1975	331
1976	332
1977	333.7
1978	335.3
1979	336.7
1980	338.5
1981	339.8
1982	341
1983	342.6
1984	344.3
1985	345.7
1986	347
1987	348.8
1988	351.4
1989	352.7
1990	354
1991	355.5
1992	356.2
1993	357
1994	358.8
1995	360.7

a) Use technology to produce a scatter plot of these data and describe any correlation that exists.

b) Use a linear regression to find the line of best fit for the data. Discuss the reliability of this model.

c) Use the regression equation to predict the level of atmospheric CO_2 that you would expect today.

d) Research current CO_2 levels. Are the results close to the predicted level? What factors could have affected the trend?

C

14. Suppose that a set of data has a perfect linear correlation except for two outliers, one above the line of best fit and the other an equal distance below it. The residuals of these two outliers are equal in magnitude, but one is positive and the other negative. Would you agree that a perfect linear correlation exists because the effects of the two residuals cancel out? Support your opinion with mathematical reasoning and a diagram.

15. **Inquiry / Problem Solving** Recall the formulas for the line of best fit using the method of least squares that minimizes the squares of vertical deviations.

a) Modify these formulas to produce a line of best fit that minimizes the squares of *horizontal* deviations.

b) Do you think your modified formulas will produce the same equation as the regular least-squares formula?

c) Use your modified formula to calculate a line of best fit for one of the examples in this section. Does your line have the same equation as the line of best fit in the example? Is your equation the inverse of the equation in the example? Explain why or why not.

16. **a)** Calculate the residuals for all of the data points in Example 3 on page 177. Make a plot of these residuals versus the independent variable, X, and comment on any pattern you see.

b) Explain how you could use such residual plots to detect outliers.

3.3 Non-Linear Regression

Many relationships between two variables follow patterns that are not linear. For example, square-law, exponential, and logarithmic relationships often appear in the natural sciences. **Non-linear regression** is an analytical technique for finding a curve of best fit for data from such relationships. The equation for this curve can then be used to model the relationship between the two variables.

As you might expect, the calculations for curves are more complicated than those for straight lines. Graphing calculators have built-in regression functions for a variety of curves, as do some spreadsheets and statistical programs. Once you enter the data and specify the type of curve, these technologies can automatically find the best-fit curve of that type. They can also calculate the coefficient of determination, r^2, which is a useful measure of how closely a curve fits the data.

INVESTIGATE & INQUIRE: Bacterial Growth

A laboratory technician monitors the growth of a bacterial culture by scanning it every hour and estimating the number of bacteria. The initial population is unknown.

Time (h)	0	1	2	3	4	5	6	7
Population	?	10	21	43	82	168	320	475

1. a) Create a scatter plot and classify the linear correlation.
   b) Determine the correlation coefficient and the line of best fit.
   c) Add the line of best fit to your scatter plot. Do you think this line is a satisfactory model? Explain why or why not.
2. a) Use software or a graphing calculator to find a curve of best fit with a
      i) quadratic regression of the form $y = ax^2 + bx + c$
      ii) cubic regression of the form $y = ax^3 + bx^2 + cx + d$
   b) Graph these curves onto a scatter plot of the data.
   c) Record the equation and the coefficient of determination, r^2, for the curves.
   d) Use the equations to estimate the initial population of the bacterial culture. Do these estimates seem reasonable? Why or why not?

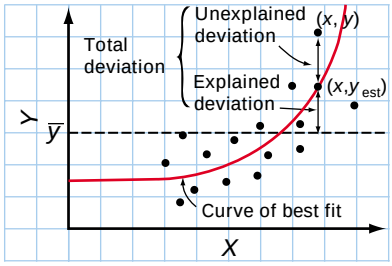
See Appendix B for details on using technology for non-linear regressions.

3. a) Perform an exponential regression on the data. Graph the curve of best fit and record its equation and coefficient of determination.
   b) Use this model to estimate the initial population.
   c) Do you think the exponential equation is a better model for the growth of the bacterial culture than the quadratic or cubic equations? Explain your reasoning.

Recall that Pearson's correlation coefficient, r, is a measure of the linearity of the data, so it can indicate only how closely a straight line fits the data. However, the **coefficient of determination, r^2**, is defined such that it applies to any type of regression curve.

$$r^2 = \frac{\text{variation in } y \text{ explained by variation in } x}{\text{total variation in } y}$$
$$= \frac{\Sigma(y_{est} - \bar{y})^2}{\Sigma(y - \bar{y})^2}$$

where $\bar{y}$ is the mean y value, y_{est} is the value estimated by the best-fit curve for a given value of x, and y is the actual observed value for a given value of x.



The total variation is the sum of the squares of the deviations for all of the individual data points.

The coefficient of determination can have values from 0 to 1. If the curve is a perfect fit, then y_{est} and y will be identical for each value of x. In this case, the variation in x accounts for all of the variation in y, so $r^2 = 1$. Conversely, if the curve is a poor fit, the total of $(y_{est} - \bar{y})^2$ will be much smaller than the total of $(y - \bar{y})^2$, since the variation in x will account for only a small part of the total variation in y. Therefore, r^2 will be close to 0. For any given type of regression, the curve of best fit will be the one that has the highest value for r^2.

For graphing calculators and Microsoft® Excel, the procedures for non-linear regression are almost identical to those for linear regression. At present, Corel® Quattro® Pro and Fathom™ do not have built-in functions for non-linear regression.

Exponential Regression

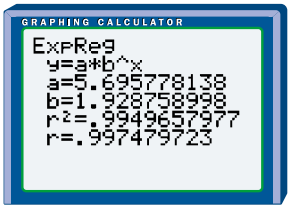
Exponential regressions produce equations with the form $y = ab^x$ or $y = ae^{kx}$, where $e = 2.718\,28\ldots$, an irrational number commonly used as the base for exponents and logarithms. These two forms are equivalent, and it is straightforward to convert from one to the other.

Example 1 Exponential Regression

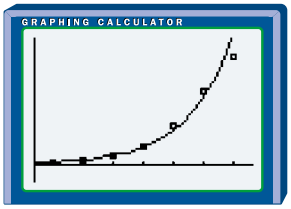
Generate an exponential regression for the bacterial culture in the investigation on page 184. Graph the curve of best fit and determine its equation and the coefficient of determination.

Solution 1 Using a Graphing Calculator

Use the ClrList command from the STAT EDIT menu to clear lists L1 and L2, and then enter the data. Set DiagnosticOn so that regression calculations will display the coefficient of determination. From the STAT CALC menu, select the non-linear regression function ExpReg. If you do not enter any list names, the calculator will use L1 and L2 by default.



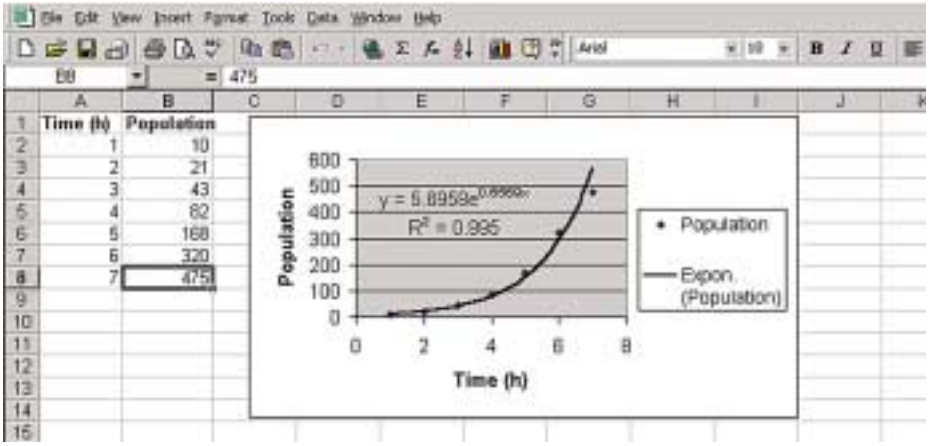
The equation for the curve of best fit is $y = 5.70(1.93)^x$, and the coefficient of determination is $r^2 = 0.995$. Store the equation as Y1. Use STAT PLOT to display a scatter plot of the data along with Y1. From the ZOOM menu, select 9:ZoomStat to adjust the window settings automatically.



Solution 2 Using a Spreadsheet

Enter the data into two columns. Next, highlight these columns and use the Chart feature to create an x-y scatter plot.

Select Chart/Add Trendline and then choose Expontenial regression. Then, select the curve that appears on your chart, and click Format/Selected Trendline/Options. Check the option boxes to display the equation and r^2.



The equation of the best-fit curve is $y = 5.7e^{0.66x}$ and the coefficient of determination is $r^2 = 0.995$. This equation appears different from the one found with the graphing calculator. In fact, the two forms are equivalent, since $e^{0.66} \doteq 1.93$.

Power and Polynomial Regression

In **power regressions**, the curve of best fit has an equation with the form $y = ax^b$.

Example 2 Power Regression

For a physics project, a group of students videotape a ball dropped from the top of a 4-m high ladder, which they have marked every 10 cm. During playback, they stop the videotape every tenth of a second and compile the following table for the distance the ball travelled.

Time (s)	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Distance (m)	0.05	0.2	0.4	0.8	1.2	1.7	2.4	3.1	3.9	4.9

a) Does a linear model fit the data well?

b) Use a power regression to find a curve of best fit for the data. Does the power-regression curve fit the data more closely than the linear model does?

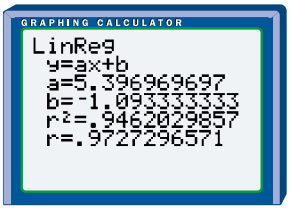
c) Use the equation for the regression curve to predict

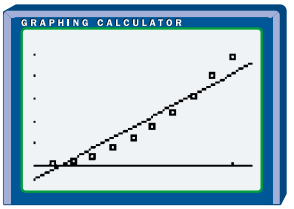
i) how long the ball would take to fall 10 m

ii) how far the ball would fall in 5 s

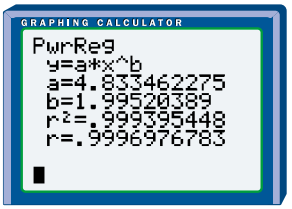
Solution 1 Using a Graphing Calculator

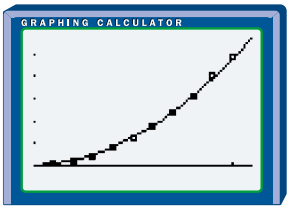
a) Although the linear correlation coefficient is 0.97, a scatter plot of the data shows a definite curved pattern. Since $b = -1.09$, the linear model predicts an initial position of about –1.1 m and clearly does not fit the first part of the data well. Also, the pattern in the scatter plot suggests the linear model could give inaccurate predictions for times beyond 1 s.





b) From the STAT CALC menu, select the **non-linear regression** function PwrReg and then follow the same steps as in Example 1.





The equation for the curve of best fit is $y = 4.83x^2$. The coefficient of determination and a graph on the scatter plot show that the quadratic curve is almost a perfect fit to the data.

c) Substitute the known values into the equation for the quadratic curve of best fit:

i) $10 = 4.83x^2$

$$x^2 = \frac{10}{4.83}$$

$$x = \sqrt{\frac{10}{4.83}}$$

$$= 1.4$$

ii) $y = 4.83(5)^2$

$$= 4.83(25)$$

$$= 121$$

The quadratic model predicts that

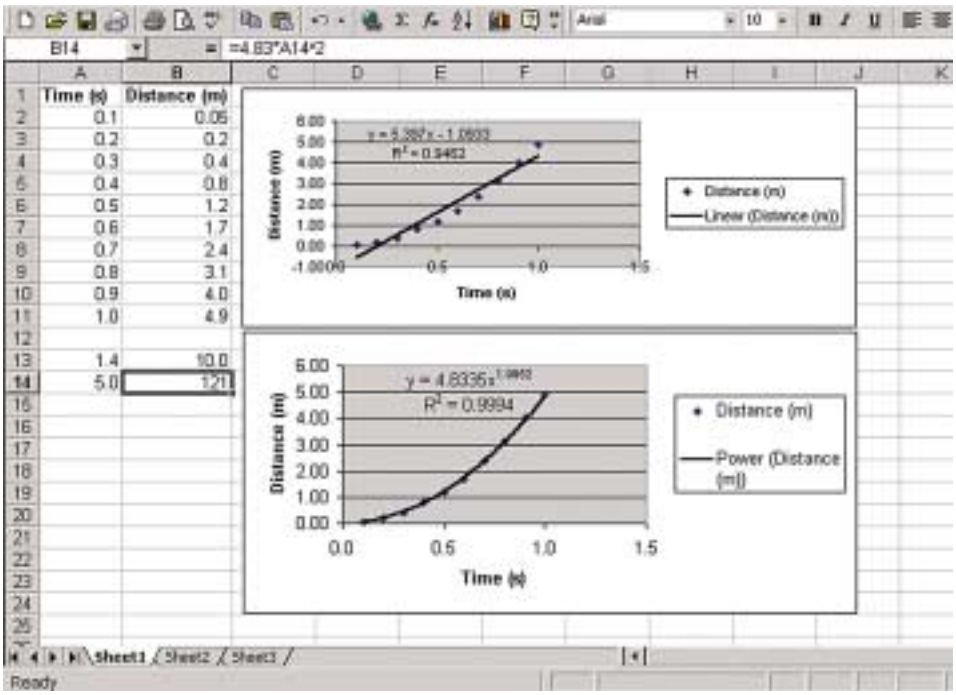
i) the ball would take approximately 1.4 s to fall 10 m

ii) the ball would fall 121 m in 5 s

Solution 2 Using a Spreadsheet

a) As in Solution 1, the scatter plot shows that a curve might be a better model.

b) Use the **Chart feature** as in Example 1, but select Power when adding the trend line.



The equation for the curve of best fit is $y = 4.83x^2$. The graph and the value for r^2 show that the quadratic curve is almost a perfect fit to the data.

c) Use the equation for the curve of best fit to enter formulas for the two values you want to predict, as shown in cells A13 and B14 in the screen above.

Example 3 Polynomial Regression

Suppose that the laboratory technician takes further measurements of the bacterial culture in Example 1.

Time (h)	8	9	10	11	12	13	14
Population	630	775	830	980	1105	1215	1410

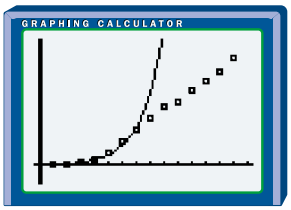
a) Discuss the effectiveness of the exponential model from Example 1 for the new data.

b) Find a new exponential curve of best fit.

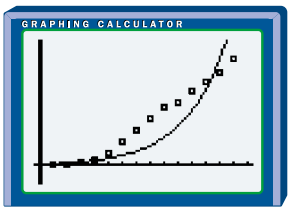
c) Find a better curve of best fit. Comment on the effectiveness of the new model.

Solution

a) If you add the new data to the scatter plot, you will see that the exponential curve determined earlier, $y = 5.7(1.9)^x$, is no longer a good fit.

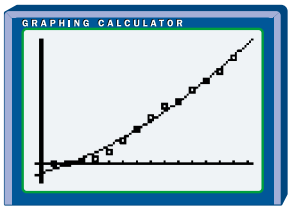


b) If you perform a new exponential regression on all 14 data points, you obtain the equation $y = 18(1.4)^x$ with a coefficient of determination of $r^2 = 0.88$. From the graph, you can see that this curve is not a particularly good fit either.

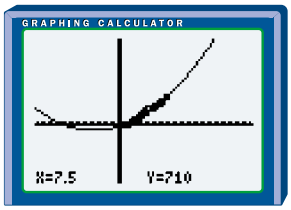


Because of the wide range of non-linear regression options, you can insist on a fairly high value of r^2 when searching for a curve of best fit to model the data.

c) If you perform a quadratic regression, you get a much better fit with the equation $y = 4.0x^2 + 55x - 122$ and a coefficient of determination of $r^2 = 0.986$.



This quadratic model will probably serve well for interpolating between most of the data shown, but may not be accurate for times before 3 h and after 14 h. At some point between 2 h and 3 h, the curve intersects the x-axis, indicating a negative population prior to this time. Clearly the quadratic model is not accurate in this range.



Similarly, if you zoom out, you will notice a problem beyond 14 h. The rate of change of the quadratic curve continues to increase after 14 h, but the trend of the data does not suggest such an increase. In fact, from 7 h to 14 h the trend appears quite linear.

It is important to recognize the limitations of regression curves. One interesting property of polynomial regressions is that for a set of n data points, a polynomial function of degree $n - 1$ can be produced which perfectly fits the data, that is, with $r^2 = 1$.

For example, you can determine the equation for a line (a first-degree polynomial) with two points and the equation for a quadratic (a second-degree polynomial) with three points. However, these polynomials are not always the best models for the data. Often, these curves can give inaccurate predictions when extrapolated.

Sometimes, you can find that several different types of curves fit closely to a set of data. Extrapolating to an initial or final state may help determine which model is the most suitable. Also, the mathematical model should show a logical relationship between the variables.

Project Prep

Non-linear models may be useful when you are analysing two-variable data in your statistics project.

Key Concepts

- Some relationships between two variables can be modelled using non-linear regressions such as quadratic, cubic, power, polynomial, and exponential curves.
- The coefficient of determination, r^2, is a measure of how well a regression curve fits a set of data.
- Sometimes more than one type of regression curve can provide a good fit for data. To be an effective model, however, the curve must be useful for extrapolating beyond the data.

Communicate Your Understanding

1. A data set for two variables has a linear correlation coefficient of 0.23. Does this value preclude a strong correlation between the variables? Explain why or why not.

2. A best-fit curve for a set of data has a coefficient of determination of $r^2 = 0.76$. Describe some techniques you can use to improve the model.

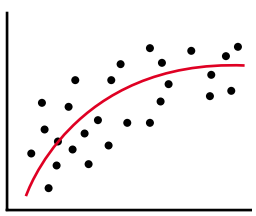
Practise

A

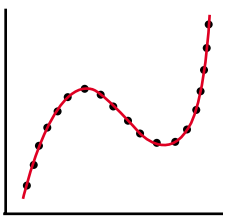
1. Match each of the following coefficients of determination with one of the diagrams below.

a) 0 **b)** 0.5 **c)** 0.9 **d)** 1

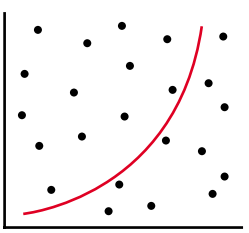
i)



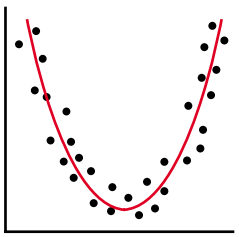
ii)



iii)



iv)



2. For each set of data use software or a graphing calculator to find the equation and coefficient of determination for a curve of best fit.

a)

x	y
−2.8	0.6
−3.5	−5.8
−2	3
−1	6
0.2	4
1	1
−1.5	5
1.4	−3.1
0.7	3
−0.3	6.1
−3.3	−3.1
−4	−7
2	−5.7

b)

x	y
−2.7	1.6
−3.5	−3
−2.2	3
−0.5	−0.5
0	1.3
0.6	4.7
−1.8	1.7
−3.8	−7
−1.3	0.6
0.8	7
0.5	2.7
−1	1.5
−3	−1.1

c)

x	y
1.1	2.5
3.5	11
2.8	8.6
2.3	7
0	1
3.8	14
1.4	4.2
−4	0.2
−1.3	0.6
3	12
4.1	17
2.2	5
−2.7	0.4

Apply, Solve, Communicate

B

3. The heights of a stand of pine trees were measured along with the area under the cone formed by their branches.

Height (m)	Area (m^2)
2.0	5.9
1.5	3.4
1.8	4.8
2.4	8.6
2.2	7.3
1.2	2.1
1.8	4.9
3.1	14.4

a) Create a scatter plot of these data.

b) Determine the correlation coefficient and the equation of the line of best fit.

c) Use a power regression to calculate a coefficient of determination and an equation for a curve of best fit.

d) Which model do you think is more accurate? Explain why.

e) Use the more accurate model to predict

i) the area under a tree whose height is 2.7 m

ii) the height of a tree whose area is 30 m^2

f) Suggest a reason why the height and circumference of a tree might be related in the way that the model in part d) suggests.

4. **Application** The biologist Max Kleiber (1893–1976) pioneered research on the metabolisms of animals. In 1932, he determined the relationship between an animal's mass and its energy requirements or basal metabolic rate (BMR). Here are data for eight animals.

Animal	Mass (kg)	BMR (kJ/day)
Frog	0.018	0.050
Squirrel	0.90	1.0
Cat	3.0	2.6
Monkey	7.0	4.0
Baboon	30	14
Human	60	25
Dolphin	160	44
Camel	530	116

a) Create a scatter plot and explain why Kleiber thought a power-regression curve would fit the data.

b) Use a power regression to find the equation of the curve of best fit. Can you rewrite the equation so that it has exponents that are whole numbers? Do so, if possible, or explain why not.

c) Is this power equation a good mathematical model for the relationship between an animal's mass and its basal metabolic rate? Explain why or why not.

d) Use the equation of the curve of best fit to predict the basal metabolic rate of

i) a 15-kg dog

ii) a 2-tonne whale

5. **Application** As a sample of a radioactive element decays into more stable elements, the amount of radiation it gives off decreases. The level of radiation can be used to estimate how much of the original element remains. Here are measurements for a sample of radium-227.

Time (h)	Radiation Level (%)
0	100
1	37
2	14
3	5.0
4	1.8
5	0.7
6	0.3

a) Create a scatter plot for these data.

b) Use an exponential regression to find the equation for the curve of best fit.

c) Is this equation a good model for the radioactive decay of this element? Explain why or why not.

d) A half-life is the time it takes for half of the sample to decay. Use the regression equation to estimate the half-life of radium-227.

6. Chapter Problem

a) Create a time-series graph for the mean starting salary of the graduates who find jobs. Describe the pattern that you see.

b) Use non-linear regression to construct a curve of best fit for the data. Record the equation of the curve and the coefficient of determination.

c) Comment on whether this equation is a good model for the graduates' starting salaries.

7. An engineer testing the transmitter for a new radio station measures the radiated power at various distances from the transmitter. The engineer's readings are in microwatts per square metre.

Distance (km)	Power Level ($\mu W/m^2$)
2.0	510
5.0	78
8.0	32
10.0	19
12.0	14
15.0	9
20.0	5

a) Find an equation for a curve of best fit for these data that has a coefficient of determination of at least 0.98.

b) Use the equation for this curve of best fit to estimate the power level at a distance of

i) 1.0 km from the transmitter

ii) 4.0 km from the transmitter

iii) 50.0 km from the transmitter

8. **Communication** *Logistic* curves are often a good model for population growth. These curves have equations with the form $y = \frac{c}{1 + ae^{-bx}}$, where a, b, and c are constants.

Consider the following data for the bacterial culture in Example 1:

Time (h)	0	1	2	3	4	5
Population	?	10	21	43	82	168

Time (h)	6	7	8	9	10	11
Population	320	475	630	775	830	980

Time (h)	12	13	14	15	16	17
Population	1105	1215	1410	1490	1550	1575

Time (h)	18	19	20
Population	1590	1600	1600

a) Use software or a graphing calculator to find the equation and coefficient of determination for the logistic curve that best fits the data for the bacteria population from 1 to 20 h.

b) Graph this curve on a scatter plot of the data.

c) How well does this curve appear to fit the entire data set? Describe the shape of the curve.

d) Write a brief paragraph to explain why you think a bacterial population may exhibit this type of growth pattern.

9. **Inquiry/Problem Solving** The following table shows the estimated population of a crop-destroying insect.

Year	Population (billions)
1995	100
1996	130
1997	170
1998	220
1999	285
2000	375
2001	490

a) Determine an exponential curve of best fit for the population data.

b) Suppose that 100 million of an arachnid that preys on the insect are imported from overseas in 1995. Assuming the arachnid population doubles every year, estimate when it would equal 10% of the insect population.

c) What further information would you need in order to estimate the population of the crop-destroying insect once the arachnids have been introduced?

d) Write an expression for the size of this population.

C

10. Use technology to calculate the coefficient of determination for two of the linear regression examples in section 3.2. Is there any relationship between these coefficients of determination and the linear correlation coefficients for these examples?

11. **Inquiry/Problem Solving** Use a software program, such as Microsoft® Excel, to analyse these two sets of data:

Data Set A		Data Set B	
x	y	x	y
2	5	2	6
4	7	4	5
6	2	7	−4
8	5	9	1
		12	2

a) For each set of data,

i) determine the degree of polynomial regression that will generate a perfectly fit regression curve

ii) perform the polynomial regression and record the value of r^2 and the equation of the regression curve

b) Assess the effectiveness of the best-fit polynomial curve as a model for the trend of the set of data.

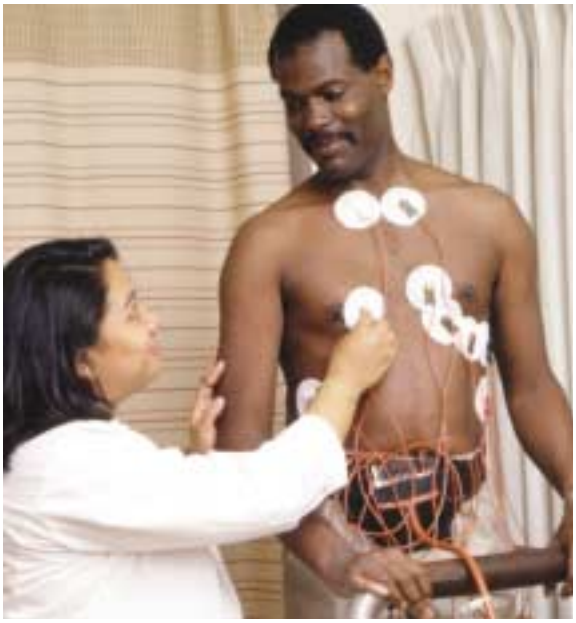
c) For data set B,

i) explain why the best-fit polynomial curve is an unsatisfactory model

ii) generate a better model and record the value of r^2 and the equation of your new best-fit curve

iii) explain why this curve is a better model than the polynomial curve found in part a)

3.4 Cause and Effect



Usually, the main reason for a correlational study is to find evidence of a cause-and-effect relationship. A health researcher may wish to prove that even mild exercise reduces the risk of heart disease. A chemical company developing an oil additive would like to demonstrate that it improves engine performance. A school board may want to know whether calculators help students learn mathematics. In each of these cases, establishing a strong correlation between the variables is just the first step in determining whether one affects the other.

INVESTIGATE & INQUIRE: Correlation Versus Cause and Effect

1. List the type of correlation that you would expect to observe between the following pairs of variables. Also list whether you think the correlation is due to a cause-and-effect relationship or some other factor.
   a) hours spent practising at a golf driving range, golf drive distance
   b) hours spent practising at a golf driving range, golf score
   c) size of corn harvest, size of apple harvest
   d) score on a geometry test, score on an algebra test
   e) income, number of CDs purchased
2. Compare your list with those of your classmates and discuss any differences. Would you change your list because of factors suggested by your classmates?
3. Suggest how you could verify whether there is a cause-and-effect relationship between each pair of variables.

A strong correlation does not prove that the changes in one variable cause changes in the other. There are various types and degrees of causal relationships between variables.

Cause-and-Effect Relationship: A change in X produces a change in Y. Such relationships are sometimes clearly evident, especially in physical processes. For example, increasing the height from which you drop an object increases its impact velocity. Similarly, increasing the speed of a production line increases the number of items produced each day (and, perhaps, the rate of defects).

Common-Cause Factor: An external variable causes two variables to change in the same way. For example, suppose that a town finds that its revenue from parking fees at the public beach each summer correlates with the local tomato harvest. It is extremely unlikely that cars parked at the beach have any effect on the tomato crop. Instead good weather is a common-cause factor that increases both the tomato crop and the number of people who park at the beach.

Reverse Cause-and-Effect Relationship: The dependent and independent variables are reversed in the process of establishing causality. For example, suppose that a researcher observes a positive linear correlation between the amount of coffee consumed by a group of medical students and their levels of anxiety. The researcher theorizes that drinking coffee causes nervousness, but instead finds that nervous people are more likely to drink coffee.

Accidental Relationship: A correlation exists without any causal relationship between variables. For example, the number of females enrolled in undergraduate engineering programs and the number of "reality" shows on television both increased for several years. These two variables have a positive linear correlation, but it is likely entirely coincidental.

Presumed Relationship: A correlation does not seem to be accidental even though no cause-and-effect relationship or common-cause factor is apparent. For example, suppose you found a correlation between people's level of fitness and the number of adventure movies they watched. It seems logical that a physically fit person might prefer adventure movies, but it would be difficult to find a common cause or to prove that the one variable affects the other.

Example 1 Causal Relationships

Classify the relationships in the following situations.

a) The rate of a chemical reaction increases with temperature.

b) Leadership ability has a positive correlation with academic achievement.

c) The prices of butter and motorcycles have a strong positive correlation over many years.

d) Sales of cellular telephones had a strong negative correlation with ozone levels in the atmosphere over the last decade.

e) Traffic congestion has a strong correlation with the number of urban expressways.

Solution

a) Cause-and-effect relationship: Higher temperatures cause faster reaction rates.

b) Presumed relationship: A positive correlation between leadership ability and academic achievement seems logical, yet there is no apparent common-cause factor or cause-and-effect relationship.

c) Common-cause factor: Inflation has caused parallel increases in the prices of butter and motorcycles over the years.

d) Accidental relationship: The correlation between sales of cellular telephones and ozone levels is largely coincidental. However, it is possible that the chemicals used to manufacture cellular telephones cause a small portion of the depletion of the ozone layer.

e) Cause-and-effect relationship and reverse cause-and-effect relationship: Originally expressways were built to relieve traffic congestion, so traffic congestion did lead to the construction of expressways in major cites throughout North America. However, numerous studies over the last 20 years have shown that urban expressways cause traffic congestion by encouraging more people to use cars.

As Example 1 demonstrates, several types of causal relationships can be involved in the same situation. Determining the nature of causal relationships can be further complicated by the presence of **extraneous variables** that affect either the dependent or the independent variable. Here, *extraneous* means external rather than irrelevant.

For example, you might expect to see a strong positive correlation between term marks and final examination results for students in your class since both these variables are affected by each student's aptitude and study habits. However, there are extraneous factors that could affect the examination results, including the time each student had for studying before the examination, the individual examination schedules, and varying abilities to work well under pressure.

In order to reduce the effect of extraneous variables, researchers often compare an **experimental group** to a **control group**. These two groups should be as similar as possible, so that extraneous variables will have about the same effect on both groups. The researchers vary the independent variable for the experimental group but not for the control group. Any *difference* in the dependent variables for the two groups can then be attributed to the changes in the independent variable.

Example 2 Using a Control Group

A medical researcher wants to test a new drug believed to help smokers overcome the addictive effects of nicotine. Fifty people who want to quit smoking volunteer for the study. The researcher carefully divides the volunteers into two groups, each with an equal number of moderate and heavy smokers. One group is given nicotine patches with the new drug, while the second group uses ordinary nicotine patches. Fourteen people in the first group quit smoking completely, as do nine people in the second group.

a) Identify the experimental group, the control group, the independent variable, and the dependent variable.

b) Can the researcher conclude that the new drug is effective?

c) What further study should the researcher do?

Solution

a) The experimental group consists of the volunteers being given nicotine patches with the new drug, while the control group consists of the volunteers being given the ordinary patches. The independent variable is the presence of the new drug, and the dependent variable is the number of volunteers who quit smoking.

b) The results of the study are promising, but the researcher has not proven that the new drug is effective. The sample size is relatively small, which is prudent for an early trial of a new drug that could have unknown side-effects. However, the sample is small enough that the results could be affected by random statistical fluctuations or extraneous variables, such as the volunteers' work environments, previous attempts to quit, and the influence of their families and friends.

c) Assuming that the new drug does not have any serious side-effects, the researcher should conduct further studies with larger groups and try to select the experimental and control groups to minimize the effect of all extraneous variables. The researcher might also conduct a study with several experimental groups that receive different dosages of the new drug.

When designing a study or interpreting a correlation, you often need background knowledge and insight to recognize the causal relationships present. Here are some techniques that can help determine whether a correlation is the result of a cause-and-effect relationship.

- Use sampling methods that hold the extraneous variables constant.
- Conduct similar investigations with different samples and check for consistency in the results.
- Remove, or account for, possible common-cause factors.

The later chapters in this book introduce probability theory and some statistical methods for a more quantitative approach to determining cause-and-effect relationships.

Project Prep

In your statistics project, you may wish to consider cause-and-effect relationships and extraneous variables that could affect your study.

Key Concepts

- Correlation does not necessarily imply a cause-and-effect relationship. Correlations can also result from common-cause factors, reverse cause-and-effect relationships, accidental relationships, and presumed relationships.
- Extraneous variables can invalidate conclusions based on correlational evidence.
- Comparison with a control group can help remove the effect of extraneous variables in a study.

Communicate Your Understanding

1. Why does a strong linear correlation not imply cause and effect?
2. What is the key characteristic of a reverse cause-and-effect relationship?
3. Explain the difference between a common-cause factor and an extraneous variable.
4. Why are control groups used in statistical studies?

Practise

A

1. Identify the most likely type of causal relationship between each of the following pairs of variables. Assume that a strong positive correlation has been observed with the first variable as the independent variable.
   - **a)** alcohol consumption, incidence of automobile accidents
   - **b)** score on physics examination, score on calculus examination
   - **c)** increase in pay, job performance
   - **d)** population of rabbits, consumer price index
   - **e)** number of scholarships received, number of job offers upon graduation
   - **f)** coffee consumption, insomnia
   - **e)** funding for athletic programs, number of medals won at Olympic games

2. For each of the following common-cause relationships, identify the common-cause factor. Assume a positive correlation between each pair of variables.
   a) number of push-ups performed in one minute, number of sit-ups performed in one minute
   b) number of speeding tickets, number of accidents
   c) amount of money invested, amount of money spent

Apply, Solve, Communicate

3. A civil engineer examining traffic flow problems in a large city observes that the number of traffic accidents is positively correlated with traffic density and concludes that traffic density is likely to be a major cause of accidents. What alternative conclusion should the engineer consider?

B

4. **Communication** An elementary school is testing a new method for teaching grammar. Two similar classes are taught the same material, one with the established method and the other with the new method. When both classes take the same test, the class taught with the established method has somewhat higher marks.
   a) What extraneous variables could influence the results of this study?
   b) Explain whether the study gives the school enough evidence to reject the new method.
   c) What further studies would you recommend for comparing the two teaching methods?

5. **Communication** An investor observes a positive correlation between the stock price of two competing computer companies. Explain what type of causal relationship is likely to account for this correlation.

6. **Application** A random survey of students at Statsville High School found that their interest in computer games is positively correlated with their marks in mathematics.
   a) How would you classify this causal relationship?
   b) Suppose that a follow-up study found that students who had increased the time they spent playing computer games tended to improve their mathematics marks. Assuming that this study held all extraneous variables constant, would you change your assessment of the nature of the causal relationship? Explain why or why not.

7. a) The net assets of Custom Industrial Renovations Inc., an industrial construction contractor, has a strong negative linear correlation with those of MuchMega-Fun, a toy distributor. How would you classify the causal relationship between these two variables?
   b) Suppose that the two companies are both subsidiaries of Diversified Holdings Ltd., which often shifts investment capital between them. Explain how this additional information could change your interpretation of the correlation in part a).

8. **Communication** Aunt Gisele simply cannot sleep unless she has her evening herbal tea. However, the package for the tea does not list any ingredients known to induce sleep. Outline how you would conduct a study to determine whether the tea really does help people sleep.

9. Find out what a *double-blind* study is and briefly explain the advantages of using this technique in studies with a control group.

10. a) The data on page 157 show a positive correlation between the size of the graduating class and the number of

Chapter Problem

graduates hired. Does this correlation mean that increasing the number of graduates causes a higher demand for them? Explain your answer.

b) A recession during the first half of the 1990s reduced the demand for business graduates. Review the data on page 157 and describe any trends that may be caused by this recession.

ACHIEVEMENT CHECK

Knowledge / Understanding	Thinking / Inquiry / Problem Solving	Communication	Application

11. The table below lists numbers of divorces and personal bankruptcies in Canada for the years 1976 through 1985.

Year	Divorces	Bankruptcies
1976	54 207	10 049
1977	55 370	12 772
1978	57 155	15 938
1979	59 474	17 876
1980	62 019	21 025
1981	67 671	23 036
1982	70 436	30 643
1983	68 567	26 822
1984	65 172	22 022
1985	61 976	19 752

a) Create a scatter plot and classify the linear correlation between the number of divorces and the number of bankruptcies.

b) Perform a regression analysis. Record the equation of the line of best fit and the correlation coefficient.

c) Identify an external variable that could be a common-cause factor.

d) Describe what further investigation you could do to analyse the possible relationship between divorces and bankruptcies.

12. Search the E-STAT, CANSIM II, or other databases for a set of data on two variables with a positive linear correlation that you believe to be accidental. Explain your findings and reasoning.

13. Use a library, the Internet, or other resources to find information on the Hawthorne effect and the placebo effect. Briefly explain what these effects are, how they can affect a study, and how researchers can avoid having their results skewed by these effects.

14. **Inquiry / Problem Solving** In a behavioural study of responses to violence, an experimental group was shown violent images, while a control group was shown neutral images. From the initial results, the researchers suspect that the gender of the people in the groups may be an extraneous variable. Suggest how the study could be redesigned to

a) remove the extraneous variable

b) determine whether gender is part of the cause-and-effect relationship

15. Look for material in the media or on the Internet that incorrectly uses correlational evidence to claim that a cause-and-effect relationship exists between the two variables. Briefly describe

a) the nature of the correlational study

b) the cause and effect claimed or inferred

c) the reasons why cause and effect was not properly proven, including any extraneous variables that were not accounted for

d) how the study could be improved

3.5 Critical Analysis

Newspapers and radio and television news programs often run stories involving statistics. Indeed, the news media often commission election polls or surveys on major issues. Although the networks and major newspapers are reasonably careful about how they present statistics, their reporters and editors often face tight deadlines and lack the time and mathematical knowledge to thoroughly critique statistical material. You should be particularly careful about accepting statistical evidence from sources that could be biased. Lobby groups and advertisers like to use statistics because they appear scientific and objective. Unfortunately, statistics from such sources are sometimes flawed by unintentional or, occasionally, entirely deliberate bias. To judge the conclusions of a study properly, you need information about its sampling and analytical methods.

INVESTIGATE & INQUIRE: Statistics in the Media

1. Find as many instances as you can of statistical claims made in the media or on the Internet, including news stories, features, and advertisements. Collect newspaper and magazine clippings, point-form notes of radio and television stories, and printouts of web pages.
2. Compare the items you have collected with those found by your classmates. What proportion of the items provide enough information to show that they used valid statistical methods?
3. Select several of the items. For each one, discuss
   a) the motivation for the statistical study
   b) whether the statistical evidence justifies the claim being made

The examples in this section illustrate how you can apply analytical tools to assess the results of statistical studies.

WEB CONNECTION

www.mcgrawhill.ca/links/MDM12

Visit the above web site and follow the links to learn more about how statistics can be misused. Describe two examples of the misuse of statistics.

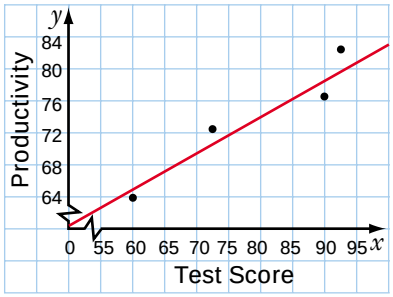
Example 1 Sample Size and Technique

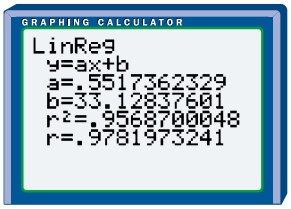
A manager wants to know if a new aptitude test accurately predicts employee productivity. The manager has all 30 current employees write the test and then compares their scores to their productivities as measured in the most recent performance reviews. The data is ordered alphabetically by employee surname. In order to simplify the calculations, the manager selects a systematic sample using every seventh employee. Based on this sample, the manager concludes that the company should hire only applicants who do well on the aptitude test. Determine whether the manager's analysis is valid.

Test Score	Productivity
98	78
57	81
82	83
76	44
65	62
72	89
91	85
87	71
81	76
39	71
50	66
75	90
71	48
89	80
82	83
95	72
56	72
71	90
68	74
77	51
59	65
83	47
75	91
66	77
48	63
61	58
78	55
70	73
68	75
64	69

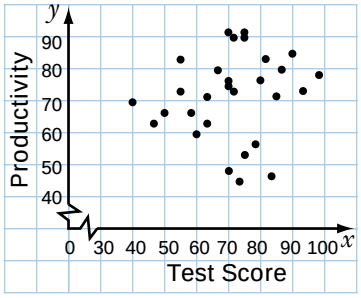
Solution

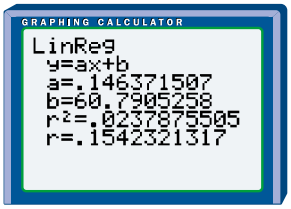
A linear regression of the systematic sample produces a line of best fit with the equation $y = 0.55x + 33$ and a correlation coefficient of $r = 0.98$, showing a strong linear correlation between productivity and scores on the aptitude test. Thus, these calculations seem to support the manager's conclusion. However, the manager has made the questionable assumption that a systematic sample will be representative of the population. The sample is so small that statistical fluctuations could seriously affect the results.





Examine the raw data. A scatter plot with all 30 data points does not show any clear correlation at all. A linear regression yields a line of best fit with the equation $y = 0.15x + 60$ and a correlation coefficient of only 0.15.





Thus, the new aptitude test will probably be useless for predicting employee productivity. Clearly, the sample was far from representative. The manager's choice of an inappropriate sampling technique has resulted in a sample size too small to make any valid conclusions.

In Example 1, the manager should have done an analysis using all of the data available. Even then the data set is still somewhat small to use as a basis for a major decision such as changing the company's hiring procedures. Remember that small samples are also particularly vulnerable to the effects of outliers.

Example 2 Extraneous Variables and Sample Bias

An advertising blitz by SuperFast Computer Training Inc. features profiles of some of its young graduates. The number of months of training that these graduates took, their job titles, and their incomes appear prominently in the advertisements.

Graduate	Months of Training	Income ($000)
Sarah, software developer	9	85
Zack, programmer	6	63
Eli, systems analyst	8	72
Yvette, computer technician	5	52
Kulwinder, web-site designer	6	66
Lynn, network administrator	4	60

a) Analyse the company's data to determine the strength of the linear correlation between the amount of training the graduates took and their incomes. Classify the linear correlation and find the equation of the linear model for the data.

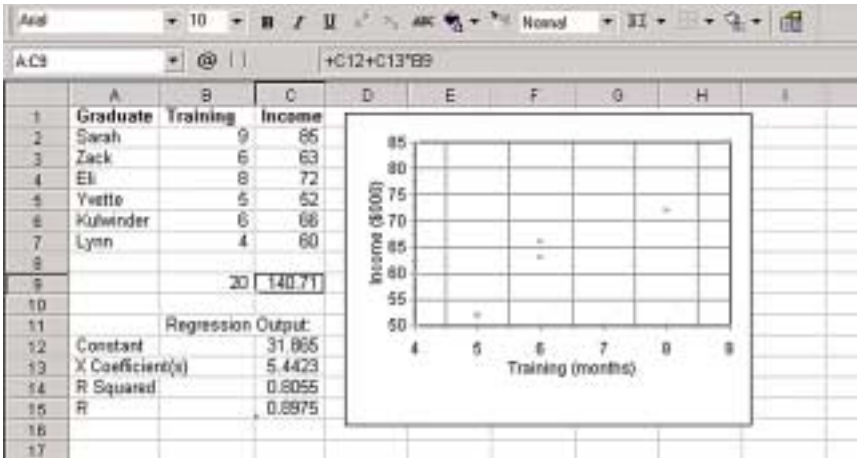
b) Use this model to predict the income of a student who graduates from the company's two-year diploma program after 20 months of training. Does this prediction seem reasonable?

c) Does the linear correlation show that SuperFast's training accounts for the graduates' high incomes? Identify possible extraneous variables.

d) Discuss any problems with the sampling technique and the data.

Solution

a) The scatter plot for income versus months of training shows a definite positive linear correlation. The regression line is $y = 5.44x + 31.9$, and the correlation coefficient is 0.90. There appears to be a strong positive correlation between the amount of training and income.



b) As shown in cell C9 in the screen above, substituting 20 months into the linear regression equation predicts an income of approximately

$y = 5.44(20) + 31.9$
$= 141$

Therefore, the linear model predicts that a graduate who has taken 20 months of training will make about $141 000 a year. This amount is extremely high for a person with a two-year diploma and little or no job experience. The prediction suggests that the linear model may not be accurate, especially when applied to the company's longer programs.

c) Although the correlation between SuperFast's training and the graduates' incomes appears to be quite strong, the correlation by itself does not prove that the training causes the graduates' high incomes. A number of extraneous variables could contribute to the graduates' success, including experience prior to taking the training, aptitude for working with computers, access to a high-end computer at home, family or social connections in the industry, and the physical stamina to work very long hours.

d) The sample is small and could have intentional bias. There is no indication that the individuals in the advertisements were randomly chosen from the population of SuperFast's students. Quite likely, the company carefully selected the best success stories in order to give potential customers inflated expectations of future earnings. Also, the company shows youthful graduates, but does not actually state that the graduates earned their high incomes immediately after graduation. It may well have taken the graduates years of hard work to reach the income levels listed in the advertisements. Further, the amounts given are incomes, not salaries. The income of a graduate working for a small start-up company might include stock options that could turn out to be worthless. In short, the advertisements do not give you enough information to properly evaluate the data.

Example 2 had several fairly obvious extraneous variables. However, extraneous variables are sometimes difficult to recognize. Such **hidden** or **lurking variables** can also invalidate conclusions drawn from statistical results.

Example 3 Detecting a Hidden Variable

An arts council is considering whether to fund the start-up of a local youth orchestra. The council has a limited budget and knows that the number of youth orchestras in the province has been increasing. The council needs to know whether starting another youth orchestra will help the development of young musicians. One measure of the success of such programs is the number of youth-orchestra players who go on to professional orchestras. The council has collected the following data.

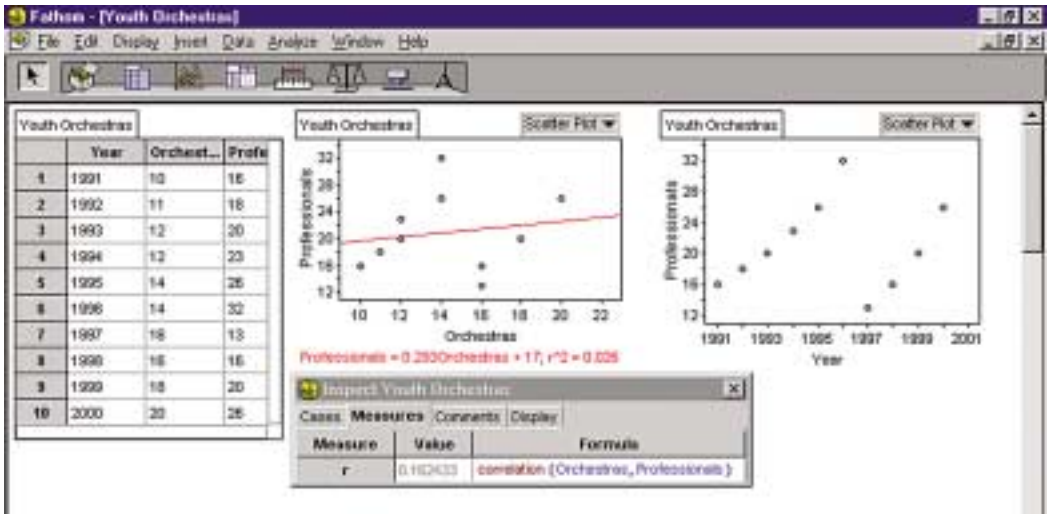
Year	Number of Youth Orchestras	Number of Players Becoming Professionals
1991	10	16
1992	11	18
1993	12	20
1994	12	23
1995	14	26
1996	14	32
1997	16	13
1998	16	16
1999	18	20
2000	20	26

a) Does a linear regression allow you to determine whether the council should fund a new youth orchestra? Can you draw any conclusions from other analysis?

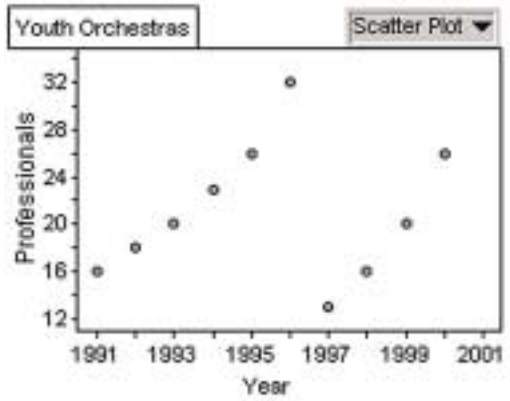
b) Suppose you discover that one of the country's professional orchestras went bankrupt in 1997. How does this information affect your analysis?

Solution

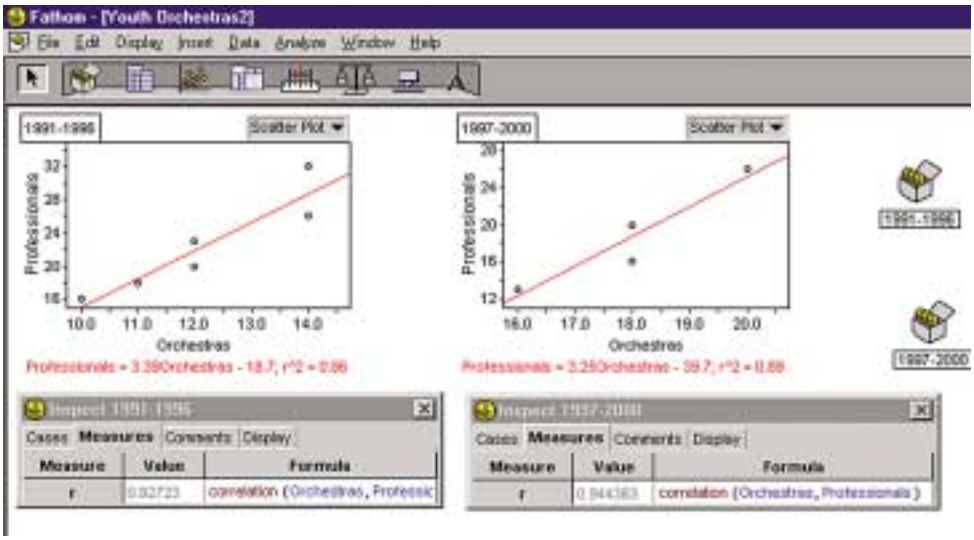
a) A scatter plot of the number of youth-orchestra members who go on to play professionally versus the number of youth orchestras shows that there may be a weak positive linear correlation. The correlation coefficient is 0.16, indicating that the linear correlation is very weak. Therefore, you might conclude that starting another youth orchestra will not help the development of young musicians. However, notice that the data points seem to form two clusters in the scatter plot, one on the left side and the other on the right. This unusual pattern suggests the presence of a hidden variable, which could affect your analysis. You will need more information to determine the nature and effect of the possible hidden variable.



You have enough data to produce a time-series graph of the numbers of young musicians who go on to professional orchestras. This graph also has two clusters of data points. The numbers rise from 1991 to 1996, drop substantially in 1997, and then rise again. This pattern suggests that something unusual happened in 1997.



b) The collapse of a major orchestra means both that there is one less orchestra hiring young musicians and that about a hundred experienced players are suddenly available for work with the remaining professional orchestras. The resulting drop in the number of young musicians hired by professional orchestras could account for the clustering of data points you observed in part a). Because of the change in the number of jobs available for young musicians, it makes sense to analyse the clusters separately.



Observe that the two sets of data both exhibit a strong linear correlation. The correlation coefficients are 0.93 for the data prior to 1997 and 0.94 for the data from 1997 on. The number of players who go on to professional orchestras is strongly correlated to the number of youth orchestras. So, funding the new orchestra may be a worthwhile project for the arts council.

The presence of a hidden variable, the collapse of a major orchestra, distorted the data and masked the underlying pattern. However, splitting the data into two sets results in smaller sample sizes, so you still have to be cautious about drawing conclusions.

When evaluating claims based on statistical studies, you must assess the methods used for collecting and analysing the data. Some critical questions are:

- Is the sampling process free from intentional and unintentional bias?
- Could any outliers or extraneous variables influence the results?
- Are there any unusual patterns that suggest the presence of a hidden variable?
- Has causality been inferred with only correlational evidence?

Project Prep

When collecting and analysing data for your statistics project, you can apply the concepts in this section to ensure that your conclusions are valid.

Key Concepts

- Although the major media are usually responsible in how they present statistics, you should be cautious about accepting any claim that does not include information about the sampling technique and analytical methods used.
- Intentional or unintentional bias can invalidate statistical claims.
- Small sample sizes and inappropriate sampling techniques can distort the data and lead to erroneous conclusions.
- Extraneous variables must be eliminated or accounted for.
- A hidden variable can skew statistical results and yet still be hard to detect.

Communicate Your Understanding

1. Explain how a small sample size can lead to invalid conclusions.

2. A city councillor states that there are problems with the management of the police department because the number of reported crimes in the city has risen despite increased spending on law enforcement. Comment on the validity of this argument.

3. Give an example of a hidden variable not mentioned in this section, and explain why this variable would be hard to detect.

Apply, Solve, Communicate

A

1. An educational researcher discovers that levels of mathematics anxiety are negatively correlated with attendance in mathematics class. The researcher theorizes that poor attendance causes mathematics anxiety. Suggest an alternate interpretation of the evidence.

2. A survey finds a correlation between the proportion of high school students who own a car and the students' ages. What hidden variable could affect this study?

B

3. A student compares height and grade average with four friends and collects the following data.

Height (cm)	Grade Average (%)
171	73
145	91
162	70
159	81
178	68

From this table, the student concludes that taller students tend to get lower marks.

a) Does a regression analysis support the student's conclusion?

b) Why are the results of this analysis invalid?

c) How can the student get more accurate results?

4. **Inquiry/Problem Solving** A restaurant chain randomly surveys its customers several times a year. Since the surveys show that the level of customer satisfaction is rising over time, the company concludes that its customer service is improving. Discuss the validity of the surveys and the conclusion based on these surveys.

5. **Application** A teacher offers the following data to show that good attendance is important.

Days Absent	Final Grade
8	72
2	75
0	82
11	68
15	66
20	30

A student with a graphing calculator points out that the data indicate that anyone who misses 17 days or more is in danger of failing the course.

a) Show how the student arrived at this conclusion.

b) Identify and explain the problems that make this conclusion invalid.

c) Outline statistical methods to avoid these problems.

6. Using a graphing calculator, Gina found the cubic curve of best fit for the salary data in the table on page 157. This curve has a coefficient of determination of 0.98, indicating an almost perfect fit to the data. The equation of the cubic curve is

$$\text{starting salary} = 0.0518y^3 - 310y^2 + 618\,412y - 411\,344\,091$$

where the salary is given in thousands of dollars and y is the year of graduation.

a) What mean starting salary does this model predict for Gina's class when they graduate in 2005?

b) Is this prediction realistic? Explain.

c) Explain why this model generated such an inaccurate prediction despite having a high value for the coefficient of determination.

d) Suggest methods Gina could use to make a more accurate prediction.

7. **Communication** Find a newspaper or magazine article, television commercial, or web page that misuses statistics of two variables. Perform a critical analysis using the techniques in this chapter. Present your findings in a brief report.

8. **Application** A manufacturing company keeps records of its overall annual production and its number of employees. Data for a ten-year period are shown below.

Year	Number of Employees	Production (000)
1992	158	75
1993	165	81
1994	172	84
1995	148	68
1996	130	58
1997	120	51
1998	98	50
1999	105	57
2000	110	62
2001	120	70

a) Create a scatter plot to see if there is a linear correlation between annual production and number of employees. Classify the correlation.

b) At some point, the company began to lay off workers. When did these layoffs begin?

c) Does the scatter plot suggest the presence of a hidden variable? Could the layoffs account for the pattern you see? Explain why or why not.

d) The company's productivity is its annual production divided by the number of

employees. Create a time-series graph for the company's productivity.

e) Find the line of best fit for the graph in part d).

f) The company has adopted a better management system. When do you think the new system was implemented? Explain your reasoning.

C

9. Search E-STAT, CANSIM II, or other sources for time-series data for the price of a commodity such as gasoline, coffee, or computer memory. Analyse the data and comment on any evidence of a hidden variable. Conduct further research to determine if there are any hidden variables. Write a brief report outlining your analysis and conclusions.

10. **Inquiry/Problem Solving** A study conducted by Stanford University found that behavioural counselling for people who had suffered a heart attack reduced the risk of a further heart attack by 45%. Outline how you would design such a study. List the independent and dependent variables you would use and describe how you would account for any extraneous variables.

Career Connection

Economist

Economists apply statistical methods to develop mathematical models of the production and distribution of wealth. Governments, large businesses, and consulting firms are employers of economists. Some of the functions performed by an economist include

- recognizing and interpreting domestic and international market trends
- using supply and demand analysis to assess market potential and set prices
- identifying factors that affect economic growth, such as inflation and unemployment
- advising governments on fiscal and monetary policies
- optimizing the economic activity of financial institutions and large businesses

Typically, a bachelor's degree in economics is necessary to enter this field. However, many positions require a master's or doctorate degree or specialized training. Since economists often deal with large amounts of data, a strong background in statistics and an ability to work with computers are definite assets.

An economist can expect to earn a comfortable living. Most employment opportunities for economists are in large cities. The current demand for economists is reasonably strong and likely to remain so for the foreseeable future, as governments and large businesses will continue to need the information and analysis that economists provide.

WEB CONNECTION

www.mcgrawhill.ca/links/MDM12

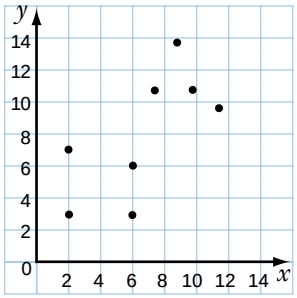
Visit the above web site and follow the links to learn more about a career as an economist and other related careers.

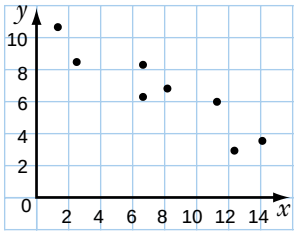
Review of Key Concepts

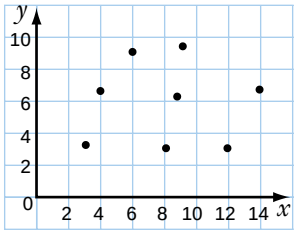
3.1 Scatter Plots and Linear Correlation

Refer to the Key Concepts on page 167.

1. **a)** Classify the linear correlation in each scatter plot shown below.







b) Determine the correlation coefficient for data points in the scatter plots in part a).

c) Do these correlation coefficients agree with your answers in part a)?

2. A survey of a group of randomly selected students compared the number of hours of television they watched per week with their grade averages.

Hours Per Week	12	10	5	3	15	16	8
Grade Average (%)	70	85	82	88	65	75	68

a) Create a scatter plot for these data. Classify the linear correlation.

b) Determine the correlation coefficient.

c) Can you make any conclusions about the effect that watching television has on academic achievement? Explain.

3.2 Linear Regression

Refer to the Key Concepts on page 179.

3. Use the method of least squares to find the equation for the line of best fit for the data in question 2.

4. The scores for players' first and second games at a bowling tournament are shown below.

First Game	169	150	202	230	187	177	164
Second Game	175	162	195	241	185	235	171

a) Create a scatter plot for these data.

b) Determine the correlation coefficient and the line of best fit.

c) Identify any outliers.

d) Repeat part b) with the outliers removed.

e) A player scores 250 in the first game. Use both linear models to predict this player's score for the second game. How far apart are the two predictions?

3.3 Non-Linear Regression

Refer to the Key Concepts on page 191.

5. An object is thrown straight up into the air. The table below shows the height of the object as it ascends.

Time (s)	0	0.1	0.2	0.3	0.4	0.5	0.6
Height (m)	0	1	1.8	2.6	3.2	3.8	4.2

a) Create a scatter plot for these data.

b) Perform a non-linear regression for these data. Record the equation of the curve of best fit and the coefficient of determination.

c) Use your model to predict the maximum height of the object.

d) Use your model to predict how long the object will be in the air.

e) Do you think that your model is accurate? Explain.

6. The table shows the distance travelled by a car as a function of time.

Time (s)	Distance (m)
0	0
2	6
4	22
6	50
8	90
10	140
12	190
14	240
16	290
18	340
20	380
22	410
24	430
26	440
28	440

a) Determine a curve of best fit to model the data.

b) Do you think the equation for this curve of best fit is a good model for the situation? Explain your reasoning.

c) Describe what the driver did between 0 and 28 s.

3.4 Cause and Effect

Refer to the Key Concepts on page 199.

7. Define or explain the following terms and provide an example of each one.

a) common-cause factor

b) reverse cause-and-effect relationship

c) extraneous variable

8. a) Explain the relationship between experimental and control groups.

b) Why is a control group needed in some statistical studies?

9. a) Explain the difference between an accidental relationship and a presumed relationship.

b) Provide an example of each.

10. The price of eggs is positively correlated with wages. Explain why you cannot conclude that raising the price of eggs should produce a raise in pay.

11. An educational researcher compiles data on Internet use and scholastic achievement for a random selection of students, and observes a strong positive linear correlation. She concludes that Internet use improves student grades. Comment on the validity of this conclusion.

3.5 Critical Analysis

Refer to the Key Cconcepts on page 209.

12. A teacher is trying to determine whether a new spelling game enhances learning. In his gifted class, he finds a strong positive correlation between use of the game and spelling-test scores. Should the teacher recommend the use of the game in all English classes at his school? Explain your answer.

13. a) Explain what is meant by the term *hidden variable*.

b) Explain how you might detect the presence of a hidden variable in a set of data.

Chapter Test

ACHIEVEMENT CHART				
Category	Knowledge / Understanding	Thinking / Inquiry / Problem Solving	Communication	Application
Questions	All	5, 7, 10	1, 5, 6, 8, 10	3, 4, 7, 10

1. Explain or define each of the following terms.

a) perfect negative linear correlation

b) experimental research

c) outlier

d) extraneous variable

e) hidden variable

2. Match the following.

	Correlation Type	Coefficient, r
a)	strong negative linear	1
b)	direct	0.6
c)	weak positive linear	0.3
d)	moderate positive linear	−0.8
e)	perfect negative linear	−1

3. The following set of data relates mean word length and recommended age level for a set of children's books.

Recommended Age	Mean Word Length
4	3.5
6	5.5
5	4.6
6	5.0
7	5.2
9	6.5
8	6.1
5	4.9

a) Create a scatter plot and classify the linear correlation.

b) Determine the correlation coefficient.

c) Determine the line of best fit.

d) Use this model to predict the average word length in a book recommended for 12-year olds.

Use the following information in order to answer questions 4–6.

Jerome has kept track of the hours he spent studying and his marks on examinations.

Subject	Hours Studied	Mark
Mathematics, grade 9	5	70
English, grade 9	3	65
Science, grade 9	4	68
Geography, grade 9	4	72
French, grade 9	2	38
Mathematics, grade 10	7	74
English, grade 10	5	69
Science, grade 10	6	71
History, grade 10	5	75
Mathematics, grade 11	12	76
English, grade 11	9	74
Physics, grade 11	14	78

4. a) Create a scatter plot for Jerome's data and classify the linear correlation.

b) Perform a regression analysis. Identify the equation of the line of best fit as y_1, and record the correlation coefficient.

c) Identify any outliers.

d) Repeat part b) with the outlier removed. Identify this line as y_2.

5. Which of the two linear models found in question 4 gives a more optimistic prediction for Jerome's upcoming biology examination? Explain.

6. **a)** Identify at least three extraneous variables in Jerome's study.

   b) Suggest some ways that Jerome might improve the validity of his study.

7. A phosphorescent material can glow in the dark by absorbing energy from light and then gradually re-emitting it. The following table shows the light levels for a phosphorescent plastic.

Time (h)	Light Level (lumens)
0	0.860
1	0.695
2	0.562
3	0.455
4	0.367
5	0.305
6	0.247

   a) Create a scatter plot for the data.

   b) Perform a quadratic regression. Record the equation of the curve of best fit and the coefficient of determination.

   c) Repeat part b) for an exponential regression.

   d) Compare how well these two models fit the data.

   e) According to each model, what will be the light level after 10 h?

   f) Which of these two models is superior for extrapolating beyond 6 h? Explain.

8. Explain how you could minimize the effects of extraneous variables in a correlation study.

9. Provide an example of a reverse cause-and-effect relationship.

ACHIEVEMENT CHECK

Knowledge / Understanding | Thinking / Inquiry / Problem Solving | Communication | Application

10. The table shown on the right contains data from the Ontario Road Safety Annual Report for 1999.

    a) Organize the data so that the age intervals are consistent. Create a scatter plot of the proportion of drivers involved in collisions versus age.

    b) Perform a regression analysis. Record the equations of the curves of best fit for each regression you try as well as the coefficient of determination.

    c) In Ontario, drivers over 80 must take vision and knowledge tests every two years to renew their licences. However, these drivers no longer have to take road tests as part of the review. Advocacy groups for seniors had lobbied the Ontario government for this change. How could such groups have used your data analysis to support their position?

Age	Licensed Drivers	Number of Collisions	% of Drivers in Age Group in Collisions
16	85 050	1 725	2.0
17	105 076	7 641	7.3
18	114 056	9 359	8.2
19	122 461	9 524	7.8
20	123 677	9 320	7.5
21–24	519 131	36 024	6.9
25–34	1 576 673	90 101	5.7
35–44	1 895 323	90 813	4.8
45–54	1 475 588	60 576	4.1
55–64	907 235	31 660	3.5
65–74	639 463	17 598	2.8
75 and older	354 581	9 732	2.7
Total	7 918 314	374 073	4.7

Wrap-Up

Implementing Your Action Plan

1. Look up the most recent census data from Statistics Canada. Pick a geographical region and study the data on age of all respondents by gender. Conjecture a relationship between age and the relative numbers of males and females. Use a table and a graph to organize and present the data. Does the set of data support your conjecture?

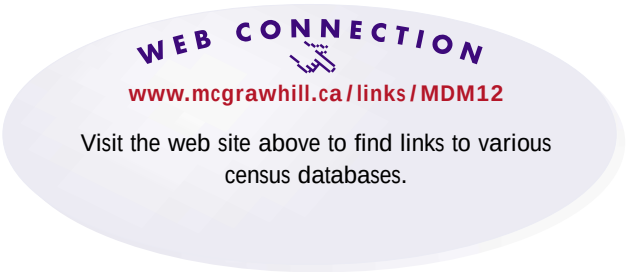
2. You may want to compare the data you analysed in step 1 to the corresponding data for other regions of Canada or for other countries. Identify any significant similarities or differences between the data sets. Suggest reasons for any differences you notice.

3. Access data on life expectancies in Canada for males and females from the 1920s to the present. Do life expectancies appear to be changing over time? Is there a correlation between these two variables? If so, use regression analysis to predict future life expectancies for males and females in Canada.

4. Access census data on life expectancies in the various regions of Canada. Select another attribute from the census data and conjecture whether there is a correlation between this variable and life expectancies. Analyse data from different regions to see if the data support your conjecture.

Suggested Resources

- Statistics Canada web sites and publications
- Embassies and consulates
- United Nations web sites and publications such as UNICEF's CyberSchoolbus and World Health Organization reports
- Statistical software (the Fathom™ sample documents include census data for Beverly Hills, California)
- Spreadsheets
- Graphing calculators



Evaluating Your Project

To help assess your own project, consider the following questions.

1. Are the data you selected appropriate?

2. Are your representations of the data effective?

3. Are the mathematical models that you used reliable?

4. Who would be interested in your findings? Is there a potential market for this information?

5. Are there questions that arose from your research that warrant further investigation? How would you go about addressing these issues in a future project?

6. If you were to do this project again, what would you do differently? Why?

Section 9.4 describes methods for evaluating your own work.

Presentation

Present the findings of your investigation in one or more of the following forms:

- written report
- oral presentation
- computer presentation (using software such as Corel® Presentations™ or Microsoft® PowerPoint®)
- web page
- display board

Remember to include a bibliography. See section 9.5 and Appendix D for information on how to prepare a presentation.

Preparing for the Culminating Project

Applying Project Skills

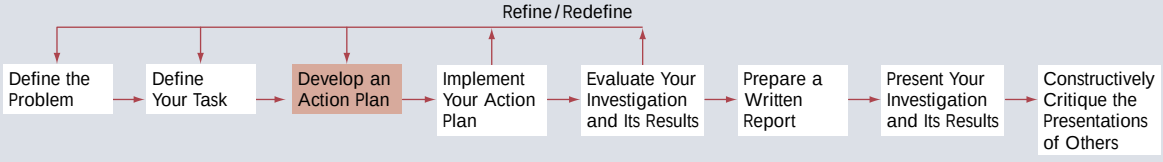
Throughout this statistics project, you have developed skills in statistical research and analysis that may be helpful in preparing your culminating project:

- making a conjecture or hypothesis
- using technology to access, organize, and analyse data
- applying a variety of statistical tools
- comparing two sets of data
- presenting your findings

Keeping on Track

At this point, you should have a good idea of the basic nature of your culminating project. You should have identified the issue that will be the focus of your project and begun to gather relevant data. Section 9.2 provides suggestions to help you clearly define your task. Your next steps are to develop and implement an action plan.

Make sure there are enough data to support your work. Decide on the best way to organize and present the data. Then, determine what analysis you need to do. As you begin to work with the data, you may find that they are not suitable or that further research is necessary. Your analysis may lead to a new approach or topic that you would like to pursue. You may find it necessary to refine or alter the focus of your project. Such changes are a normal part of the development and implementation process.



1. Let $A = \begin{bmatrix} 7 & 3 \\ 0 & -2 \\ -5 & 4 \end{bmatrix}$, $B = \begin{bmatrix} 8 & 1 \\ -5 & 4 \end{bmatrix}$, and $C = \begin{bmatrix} -8 & 0 \\ 5 & 6 \\ 9 & -3 \end{bmatrix}$. Calculate, if possible,

a) $-2(A + C)$ **b)** AC

c) $(BA)^t$ **d)** B^2

e) C^2 **f)** B^{-1}

2. a) Describe the iterative process used to generate the table below.

b) Continue the process until all the cells are filled.

	17	16	15	14	13	
	18	5	4	3	12	
		6	1	2	11	
		7	8	9	10	

3. Which of the following would you consider to be databases? Explain your reasoning.

a) a novel

b) school attendance records

c) the home page of a web site

d) an advertising flyer from a department store

4. What sampling techniques are most likely to be used for the following surveys? Explain each of your choices.

a) a radio call-in show

b) a political poll

c) a scientific study

5. Classify the type of linear correlation that you would expect for each pair of variables.

a) air temperature, altitude

b) income, athletic ability

c) people's ages from 1 to 20 years, their masses

d) people's ages from 21 to 40 years, their masses

6. Identify the most likely causal relationship between each of the following pairs of variables.

a) grade point average, starting salary upon graduation

b) grade in chemistry, grade in physics

c) sales of symphony tickets, carrot harvest

d) monthly rainfall, monthly umbrella sales

7. a) Sketch a map that can be coloured using only three colours.

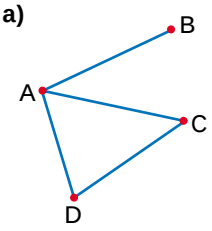
b) Reconfigure your map as a network.

8. State whether each of the following networks is

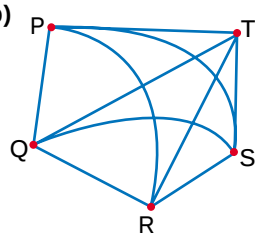
i) connected **ii)** traceable **iii)** planar

Provide evidence for your decisions.

a)



b)

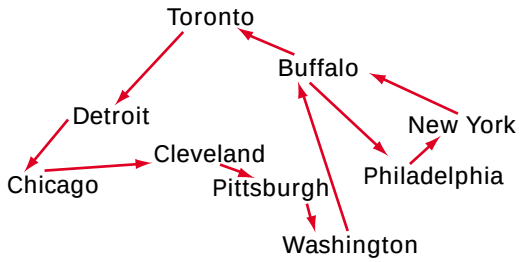


9. Use a tree diagram to represent the administrative structure of a school that has a principal, vice-principals, department heads, assistant heads, and teachers.

10. A renowned jazz pianist living in Toronto often goes on tours in the United States. For the tour shown below, which city has the most routes

a) with exactly one stopover?

b) with no more than two stopovers?



11. The following are responses to a survey that asked: "On average, how many hours per week do you read for pleasure?"

1	3	0	0	7	2	0	1	10	5	2	2	2	0	1	4	0	8	3	1	3
0	0	2	15	4	9	1	6	7	0	3	3	14	5	7	0	1	1	0	10	0

Use a spreadsheet to

a) sort the data from smallest to largest value

b) determine the mean hours of pleasure reading

c) organize the data into a frequency table with appropriate intervals

d) make a histogram of the information in part c)

12. The annual incomes of 40 families surveyed at random are shown in the table.

Income ($000)						
28.5	38	61	109	42	56	19
27	44.5	81	36	39	51	40.5
67	28	60	87	58	120	111
73	65	34	54	16.5	135	70.5
59	47	92	38	55	84.5	107
71	59	26.5	76	50		

a) Group these data into 8 to 12 intervals and create a frequency table.

b) Create a histogram and a cumulative-frequency diagram for the data.

c) What proportion of the families surveyed earn an annual income of $60 000 or less?

13. Classify the bias in each of the following situations. Explain your reasoning in each case.

a) At a financial planning seminar, the audience were asked to raise their hands if they had ever considered declaring bankruptcy.

b) A supervisor asked an employee if he would mind working late for a couple of hours on Friday evening.

c) A survey asked neighbourhood dog-owners if dogs should be allowed to run free in the local park.

d) An irascible talk show host listed the mayor's blunders over the last year and invited listeners to call in and express their opinions on whether the mayor should resign.

14. The scores in a recent bowling tournament are shown in the following table.

150	260	213	192	176	204	138	214	298	188
168	195	225	170	260	254	195	177	149	224
260	222	167	182	207	221	185	163	112	189

a) Calculate the mean, median, and mode for this distribution. Which measure would be the most useful? Which would be the least useful? Explain your choices.

b) Determine the standard deviation, first quartile, third quartile, and interquartile range.

c) Explain what each of the quantities in part b) tells you about the distribution of scores.

d) What score is the 50th percentile for this distribution?

e) Is the player who scored 222 above the 80th percentile? Explain why or why not.

15. The players on a school baseball team compared their batting averages and the hours they spent at the batting practice.

Batting Average	Practice Hours
0.220	20
0.215	18
0.185	15
0.170	14
0.200	18
0.245	22
0.230	19
0.165	15
0.205	17

a) Identify the independent variable and dependent variable. Explain your choices.

b) Produce a scatter plot for the data and classify the linear correlation.

c) Determine the correlation coefficient and the equation of the line of best fit.

d) Use this linear model to predict the batting average for players who had batting practice for

i) 16 h **ii)** 13 h **iii)** 35 h

e) Discuss how accurate you think each of these predictions will be.

16. Describe a method you could use to detect outliers in a sample.

17. A bright, young car salesperson has made the following gross sales with her first employer.

Year	Gross Sales ($ millions)
1997	0.8
1998	1.1
1999	1.6
2000	2.3
2001	3.5
2002	4.7

a) Create a time-series graph for these data.

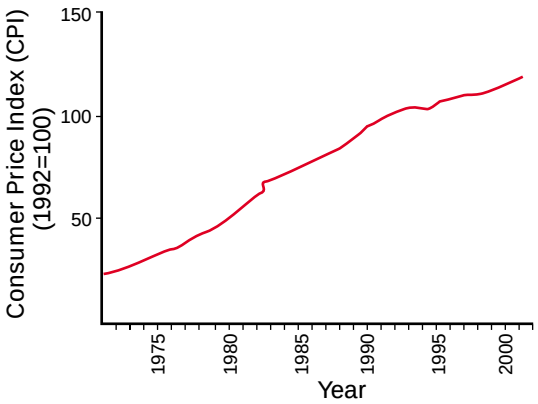
b) Based on this graph, what level of sales would you predict for 2003?

c) List three factors that could affect the accuracy of your prediction.

d) Compute an index value for the sales each year using the 1997 sales as a base. What information do the index values provide?

e) Suppose that this salesperson is thinking of changing jobs. Outline how she could use the sales index to convince other employers to hire her.

18. The following time-series graph shows the Consumer Price Index (CPI) for the period 1971 to 2001.



a) What is the base for this index? When did the CPI equal half of this base value?

b) Approximately how many times did the average price of goods double from 1971 to 1992?

c) Which decade on this graph had the highest rate of inflation? Explain your answer.

d) Estimate the overall rate of inflation for the period from 1971 to 2001.

Designing a Game

Background

Many games introduce elements of chance with random processes. For example, card games use shuffled cards, board games often use dice, and bingo uses randomly selected numbers.

Your Task

Design and then analyse a game for two or more players, involving some form of random process. One of the players may assume the role of dealer or game master.

Developing an Action Plan

You will need to decide on one or more instruments of chance, such as dice, cards, coins, coloured balls, a random-number generator, a spinner, or a nail maze. Recommend a method of tracking progress or keeping score, such as a game board or tally sheet. Create the rules of the game. Submit a proposal to your teacher outlining the concept and purpose of your game.