6 Further statistical analysis

Syllabus topic — S3 Further statistical analysis

This topic will introduce a variety of methods for identifying, analysing and describing associations between pairs of variables.

Outcomes

- Construct a bivariate scatterplot in identify patterns in data.
- Use bivariate scatterplot to describe the patterns, features and associations of bivariate datasets.
- Identify the dependent and independent variables within bivariate datasets.
- Model a linear association by fitting an appropriate line of best fit to a scatterplot and using it to describe patterns and associations.
- Use an appropriate line of best fit to make predictions by either interpolation and extrapolation.
- Implement the statistical investigation process that involves two numerical variables.

Digital Resources for this chapter

In the Interactive Textbook:

- Videos
- Literacy worksheet
- Quick Quiz
- Solutions (enabled by teacher)
- Desmos widgets
- Spreadsheets
- Study guide

In the Online Teaching Suite:

- Teaching Program
- Tests
- Review Quiz
- Teaching Notes

Knowledge check

The Interactive Textbook provides a test of prior knowledge for this chapter, and may direct you to revision from the previous years' work.

ISBN 978-1-108-44806-2

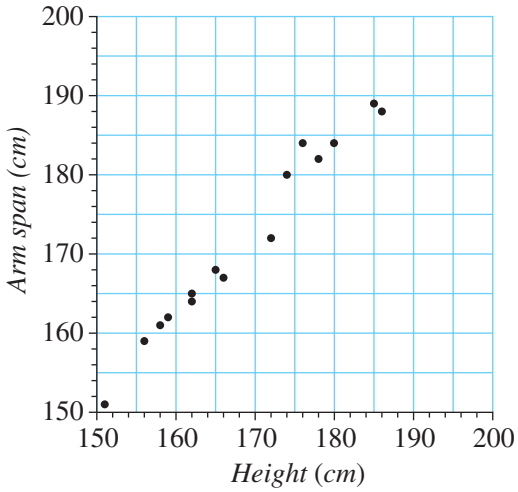
6A Constructing a bivariate scatterplot

Bivariate data is data that has two variables.

A scatterplot is used to determine if there is a relationship between two numerical variables. Data is collected on the two variables and often displayed in a table of ordered pairs. A scatterplot is a graph of the ordered pairs of numbers. Each ordered pair is a dot on the graph. To illustrate this process, a scatterplot has been constructed to determine the relationship between the height and arm span. The data collected on these variables is shown below in the table of ordered pairs.

Height (in cm)	172	159	178	162	156	174	151	162	165	185	186	176	166	180	158
Arm span (in cm)	172	162	182	164	159	180	151	165	168	189	188	184	167	184	161

Each person has two numerical variables, height and arm span. To construct a scatterplot, draw a number plane with the height on the horizontal axis and arm span on the vertical axis. Plot each ordered pair as a dot. The scatterplot shows there is a relationship between these variables.



CONSTRUCTING A SCATTERPLOT

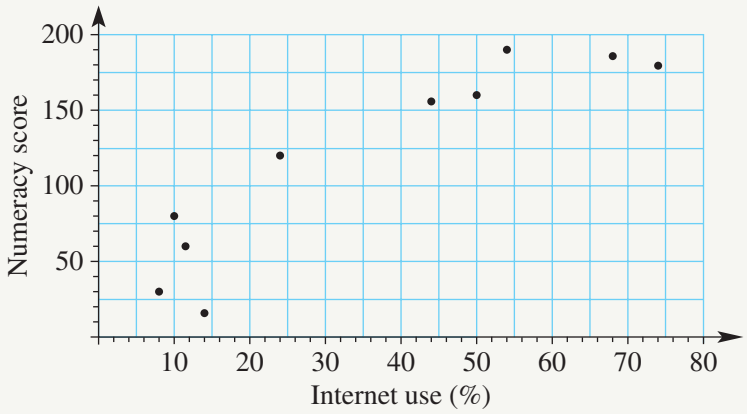
1. Draw a number plane.
2. Determine a scale and a title for the horizontal or x-axis.
3. Determine a scale and a title for the vertical or y-axis.
4. Plot each ordered pair of numbers with a dot.

Mathematics Standard 1
Cambridge Maths Stage 6
ISBN 978-1-108-44806-2

Cambridge University Press

Example 1: Reading a scatterplot 6A

The average numeracy score for year 6 students and their general rate of internet use (%) for 10 countries are displayed in the scatterplot below.



a What is the scale for the vertical axis?
b What is the average numeracy score for the country which has an internet use rate of 24%?
c What is the internet use (%) for the country which has an average numeracy score 160?
d How many countries have internet use of less than 50%?
e How many countries have a numeracy score greater than 100?
f Is there a relationship between these two variables?

SOLUTION:

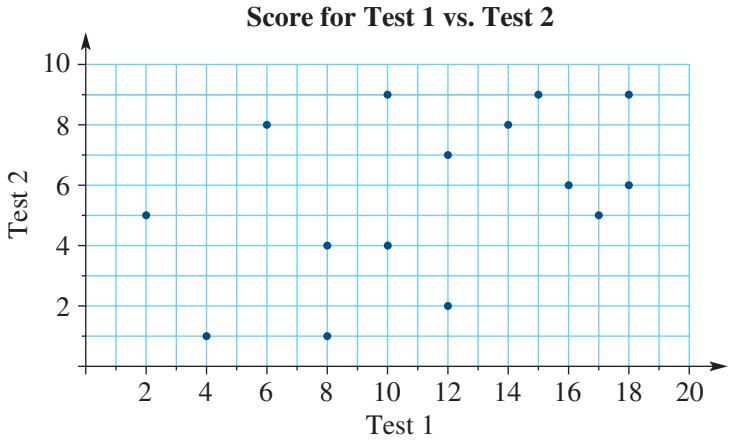
1 Count the number of divisions between 0 and 50 (5). Therefore 1 unit is 50 divided by 5 (10).	**a** 1 unit = 10
2 Read from the scatterplot (when internet use is 24% the numeracy score is 120).	**b** 120
3 Read from the scatterplot (when the numeracy score is 160 the internet use is 50%).	**c** 50%
4 Count the number of dots less than 50% (left-hand side).	**d** 6 countries
5 Count the number of dots greater than 100 (top-half).	**e** 6 countries
6 Look for any pattern in the dots. In this scatterplot when the internet use is greater than 20%, there is a clear increase in the numeracy score. However, this relationship does not exist when the internet use is less than 20%.	**f** Yes, there is a relationship. When the internet use is greater than 20%, both the variables are increasing.

ISBN 978-1-108-44806-2

Exercise 6A

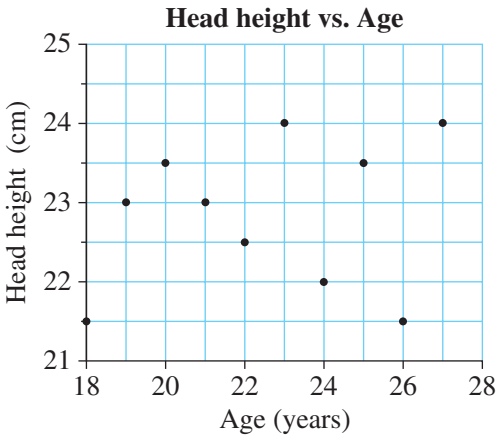
Example 1

1 The scatterplot shows the results for 15 students in two tests.



- **a** What is the highest mark in test 2?
- **b** What is the lowest mark in test 1?
- **c** What is the range for test 1?
- **d** What is the mode for test 2?
- **e** How many students scored greater than 6 in test 1?

2 The scatterplot shows the head height to age for 10 people.

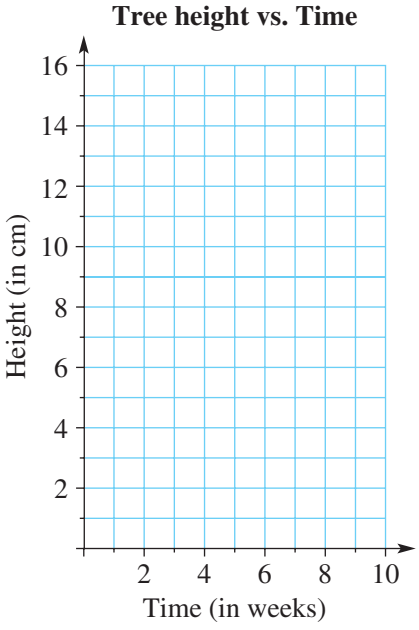


- **a** What is the head height for a person who is 21 years old?
- **b** What is the age of the person who has a head height of 22 cm?
- **c** What is the largest head height?
- **d** What is the age of the youngest person?
- **e** How many people have a head height greater than 23 cm?
- **f** Is there a clear relationship between these two variables?

3 The table below shows the height (in cm) of a eucalyptus tree seedling as it grows.

Time (in weeks)	0	1	2	3	4	5	6	7	8	9	10
Height (in cm)	0	6.6	8.8	9.0	10.5	12.0	13.5	15.2	15.4	15.8	15.9

- **a** Copy the number plane opposite to construct a scatterplot using the above table.
- **b** What is the increase in the height of the seedling during the first week?
- **c** What is the increase in the height of the seedling during the last week?
- **d** How many weeks does it take for the seedling to increase in height from 9 cm to 12 cm?
- **e** Estimate the height of the seedling after 4.5 weeks.
- **f** Estimate the time taken for the seedling to grow to a height of 14 cm.



Mathematics Standard 1
Cambridge Maths Stage 6
ISBN 978-1-108-44806-2

Cambridge University Press

4 Adrian is a political commentator who has been studying the effects of television exposure time on the approval ratings of nine politicians. The data is shown below.

Time (in minutes)	5	15	15	75	25	70	40	55	20
Approval rating (%)	60	30	50	90	25	55	55	45	40

a Construct a scatterplot of the data given in the table.
b Are there any conclusions to be drawn from the scatterplot?

5 The table shows the number of runs scored and the number of balls faced by batsmen in a one-day cricket match.

Balls faced	49	29	26	16	19	13	16	10	28	40	6
Runs scored	47	27	10	8	21	3	13	6	15	30	2

a Construct a scatterplot of the data given in the table.
b Are there any conclusions to be drawn from the scatterplot?

6 The maximum wind speed and maximum temperature were recorded for 2 weeks. The data is displayed in the table below.

Wind speed (in km/h)	2	6	12	15	19	20	22	25	17	14	5	11	24	13
Temperature (in °C)	28	26	23	22	21	22	19	16	20	24	25	24	19	26

a Construct a scatterplot of the data given in the table.
b Are there any conclusions to be drawn from the scatterplot?

7 The table below shows the age of the car (in months) and the minimum stopping distance (in metres) when the car is travelling at 60 km/h.

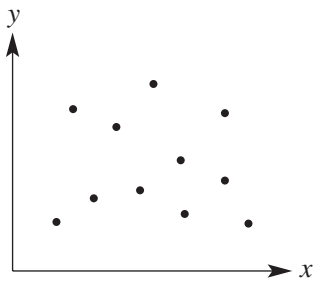
Age of car (in months)	48	12	65	42	98	34
Stopping distance (in metres)	29	28	38	35	36	37

a Prepare a scatterplot using the above data.
b Are there any conclusions to be drawn from the scatterplot?

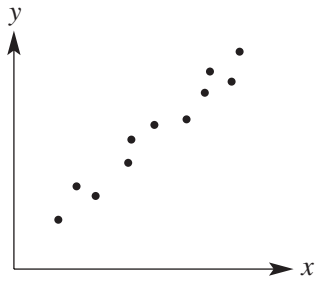
ISBN 978-1-108-44806-2

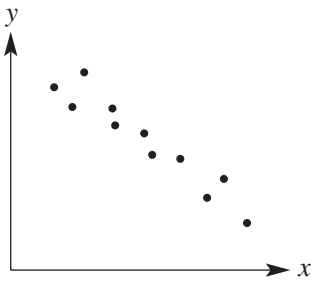
6B Using a bivariate scatterplot

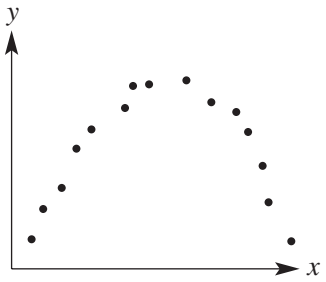
What are the features in a scatterplot that will identify and describe any relationship? First look for a clear pattern. In the scatterplot opposite, there is no clear pattern in the points: they are just randomly spread on the scatterplot. There is no relationship or association between the variables.



For the three examples below, there is a clear (but different) pattern in each set of points, so we conclude that there is an association in each case.







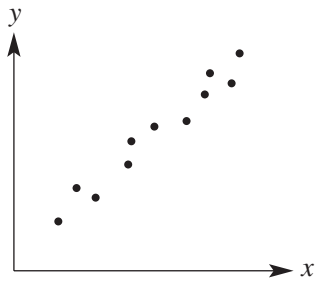
Having found a clear pattern, we need to be able to describe these associations clearly, as they are obviously quite different. There are three things we look for in the pattern of points: form, direction and strength.

Form of an association

If an association exists between the variables then the points in a scatterplot tend to follow a linear pattern or a curved pattern. This is called the form of an association.

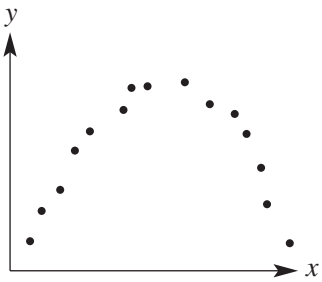
Linear form

If the points seem to approximate a straight line, the association is a linear form.



Non-linear form

If the points seem to approximate a curve, the association is a non-linear form.



FORM OF AN ASSOCIATION

Linear form – when the points tend to follow a straight line.
Non-linear form – when the points tend to follow a curved line.

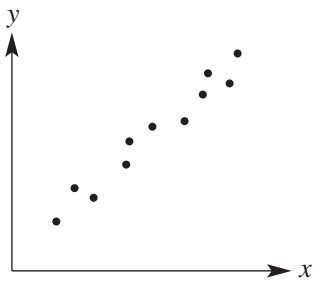
ISBN 978-1-108-44806-2

Direction of an association

There are two types of direction if the association is in linear form.

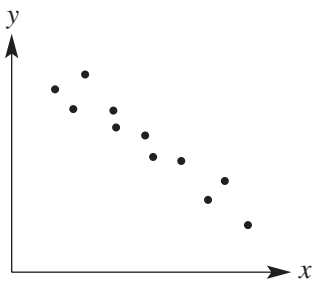
Positive

Positive association exists between the variables if the gradient of the line is positive. That is, the dots on the scatterplot tend to go up as we go from left to right.



Negative

Negative association exists between the variables if the gradient of the line is negative. That is, the dots on the scatterplot tend to go down as we go from left to right.



DIRECTION OF AN ASSOCIATION

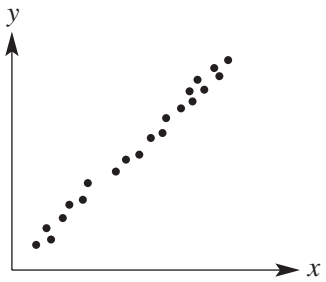
Positive – gradient of the line is positive.
Negative – gradient of the line is negative.

Strength of an association

The strength of an association is a measure of how much scatter there is in the scatterplot.

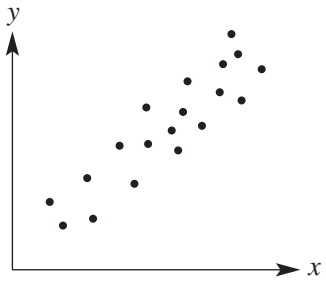
Strong

In strong association the dots will tend to follow a single stream. A pattern is clearly seen. There is only a small amount of scatter in the plot.



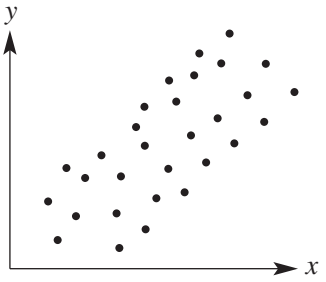
Moderate

In moderate association the amount of scatter in the plot increases and the pattern becomes less clear. This indicates that the association is less strong.



Weak

In weak association the amount of scatter increases further and the pattern becomes even less clear. Linear form is less evident.



STRENGTH OF AN ASSOCIATION

Strong – small amount of scatter in the plot.
Moderate – modest amount of scatter in the plot.
Weak – large amount of scatter in the plot

ISBN 978-1-108-44806-2

Example 2: Describing bivariate datasets 6B

The table below shows the height (cm) and weight (kg) of nine people.

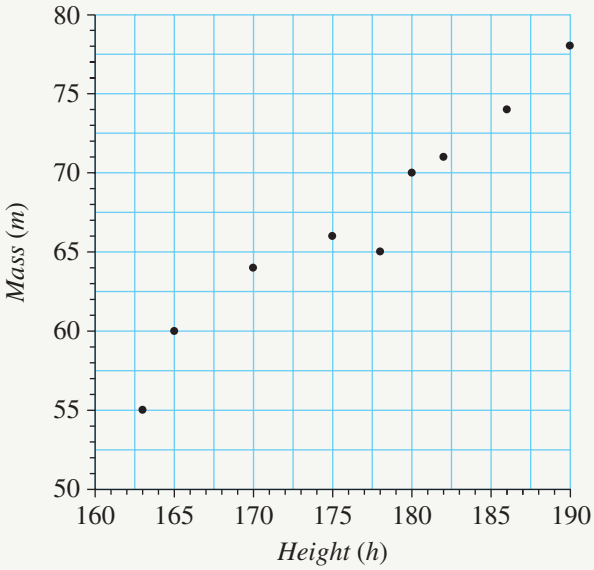
Height (h)	163	165	170	175	178	180	182	186	190
Mass (m)	55	60	64	66	65	70	71	74	78

a Construct a scatterplot using the above table.
b Describe the form of the association.
c Describe the direction of the association.
d Describe the strength of the association.
e Predict the mass of a person who is 173 cm tall using the scatterplot.
f Predict the height of a person who has a mass of 75 kg using the scatterplot.

SOLUTION:

1 Draw a number plane with h as the horizontal axis and m as the vertical axis.
2 Determine a scale for the horizontal axis. Let each unit represent 1 cm.
3 Determine a scale for the vertical axis. Let each unit represent 1 kg.
4 Write titles for the horizontal and vertical axes.
5 Plot the points (163, 55), (165, 60), (170, 64), (175, 66), (178, 65), (180, 70), (182, 71), (186, 74) and (190, 78).

a



6 Look for a pattern. The points approximate a straight line.
7 Gradient of the line is positive. The dots tend to go up as you move from left to right.
8 There is a small amount of scatter in the scatterplot.
9 Draw an imaginary vertical line from 173 cm.
10 Try to maintain the linear relationship and guess the weight.
11 Draw an imaginary horizontal line from 75 kg.
12 Try to maintain the linear relationship and guess the height.

b Linear form
c Positive
d Strong
e The person weighs approximately 65 kg.
f The person's height is approximately 187 cm.

ISBN 978-1-108-44806-2

Independent and dependent variables

Bivariate data has two variables that are often identified as the independent and dependent variables. The independent variable is the input. It is not affected by the other variable and is represented on the horizontal axis of the scatterplot. The dependent variable is the output and is 'dependent' on the independent variable. It is represented on the vertical axis of a scatterplot.

Example 3: Identifying independent and dependent variables 6B

The table below shows the time taken (hours) relative to the distance travelled (km).

Distance (d)	0	10	20	30	40	50
Time (t)	0	0.25	0.38	0.59	0.82	1.00

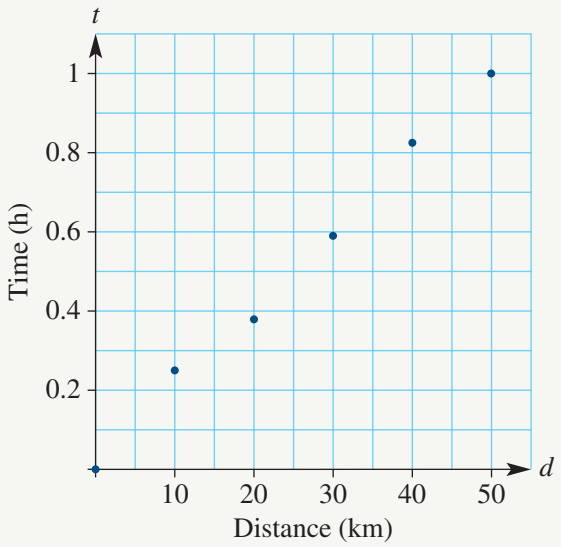
a Draw a scatterplot using the above table.

b Which are the independent and dependent variables?

SOLUTION:

1 Draw a number plane with d as the horizontal axis and t as the vertical axis.
2 Determine a scale for the horizontal axis. Let each unit represent 10 km.
3 Determine a scale for the vertical axis. Let each unit represent 0.2 hours.
4 Write titles for the horizontal and vertical axes.
5 Plot the points (0, 0) (10, 0.25) (20, 0.38) (30, 0.59) (40, 0.82) (50, 1).

a



6 The independent variable is the input and represented on the horizontal axis of the scatterplot.
7 The dependent variable is the output and represented on the vertical axis of the scatterplot.

b Independent variable is distance (d).

Dependent variable is time (t).

ISBN 978-1-108-44806-2

Exercise 6B

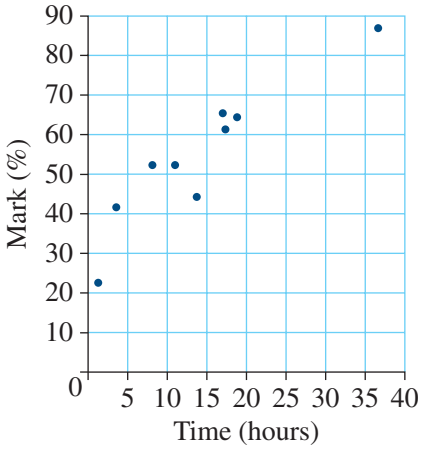
Example 2

1 Describe the association in the following scatterplots as:

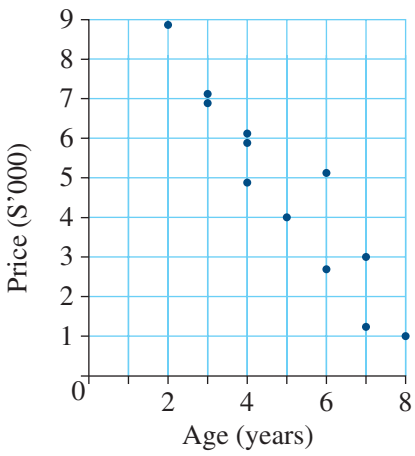
i linear or non-linear **ii** positive or negative

iii strong, moderate or weak.

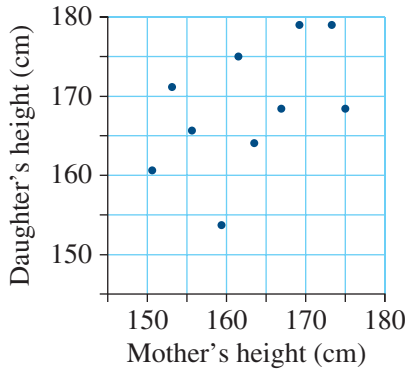
a



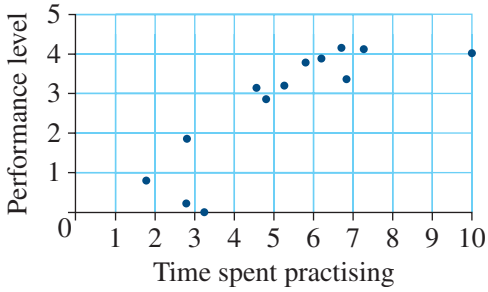
b



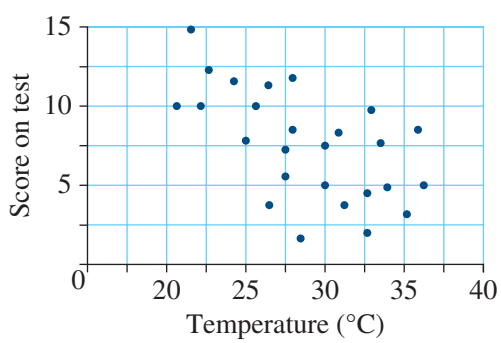
c



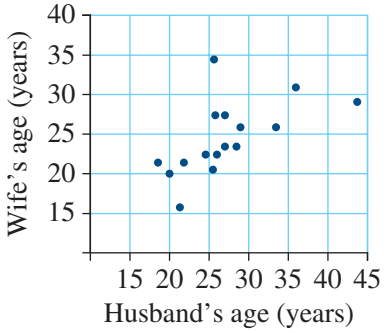
d



e



f



2 For each of the following pairs of variables, indicate whether you expect an association to exist and, if so, whether you would expect the association to be positive or negative.

a Independent variable: Distance travelled Dependent variable: Time taken

b Independent variable: Amount of daily exercise Dependent variable: Fitness level

c Independent variable: Foot length of an adult Dependent variable: Intelligence

d Independent variable: Number of pages Dependent variable: Book price

e Independent variable: Temperature above 30°C Dependent variable: Comfort level

ISBN 978-1-108-44806-2

Example 3

3 Chocolates are sold for $12 per kg. The table below shows weight against cost.

Weight (w)	1	2	3	4	5
Cost (c)	12	24	36	48	60

a Which is the independent variable?
b Which is the dependent variable?
c Draw a scatterplot of weight against cost.
d Is the form linear or non-linear?
e Is the direction positive or negative?
f Is the strength strong, moderate or weak?

4 The table below shows the drug dosage against reaction time.

Drug dosage (d)	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	6.0
Reaction time (t)	66	48	35	19	18	17	11	15	10	10	11

a Which is the independent variable?
b Which is the dependent variable?
c Draw a scatterplot of time against cost.
d Is the form linear or non-linear?
e Is the direction positive or negative?
f Is the strength strong, moderate or weak?

5 Kayla conducted a science experiment and presented the results in a table.

Mass (m)	3	6	9	12	15
Time (t)	8.2	6.7	5.2	3.7	2.2

a Which is the independent variable?
b Which is the dependent variable?
c Draw a scatterplot of mass against time.
d Is the form linear or non-linear?
e Is the direction positive or negative?
f Is the strength strong, moderate or weak?

6 The table below shows leg length compared with height.

Leg length (in cm)	83	83	85	87	89	89	92	93	94
Height (in cm)	166	167	170	174	179	178	183	185	188

a Which is the independent variable?
b Which is the dependent variable?
c Draw a scatterplot of leg length against height.
d Is the form linear or non-linear?
e Is the direction positive or negative?
f Is the strength strong, moderate or weak?

ISBN 978-1-108-44806-2

6C Line of best fit

If the points on the scatterplot tend to lie on a straight line, then we can fit a line on the scatterplot. The process of fitting a straight line to the data is known as linear regression. Linear regression is completed in many different ways. The simplest method is to draw a line that seems to be a balance of the points above and below the line. The aim of a linear regression is to model the association between two numerical variables by using the equation of a straight line. This equation of the straight line is found using the gradient–intercept formula: $y = mx + c$ where m is the gradient and c is the y-intercept.

LINE OF BEST FIT

A line of best fit is a straight line that approximates a linear association between points.
The equation of the line of best fit is found using the gradient–intercept formula: $y = mx + c$.

The line of best fit is used to make a prediction about one of the variables. When it is used to make a prediction within the data range it is called interpolation. Extrapolation is a prediction outside the data range and must be used carefully, as the line of best fit may not apply, for example, predicting an adult's height based on their increasing height as a child. Interpolation and extrapolation will be examined in detail in section 6D.

Example 4: Drawing a line of best fit by eye 6C

The table below shows the height (cm) and weight (kg) of nine people.

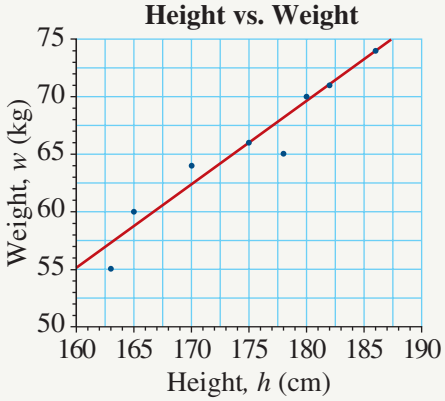
Height (h)	163	165	170	175	178	180	182	186
Weight (w)	55	60	64	66	65	70	71	74

a Construct a scatterplot and draw a line of best fit.
b Describe the association between height and weight.

SOLUTION:

1 Draw a number plane with h as the horizontal axis and w as the vertical axis.
2 Plot the points (163, 55), (165, 60), (170, 64), (175, 66), (178, 65), (180, 70), (182, 71) and (186, 74).
3 Draw a straight line as close as possible to every point. There should be some points above, below and on the line.
4 The line of best fit has a positive gradient and is close to the points.

a



b Strong positive linear association.

ISBN 978-1-108-44806-2

Exercise 6C

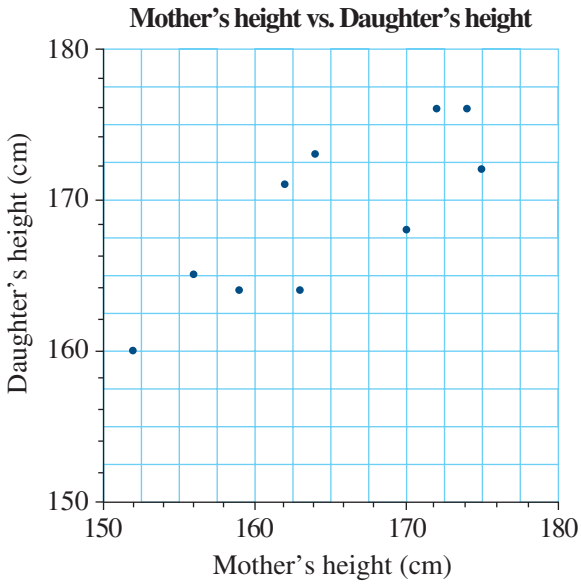
Example 4

1 Draw a scatterplot and a line of best fit by eye for the following points.

- **a** (0, 0) (10, 30) (20, 67) (30, 93) (40, 126) (50, 158) (60, 178)
- **b** (5, 20) (10, 42) (15, 73) (20, 94) (25, 122) (30, 150) (35, 165)
- **c** (0, 6) (2, 24) (3, 39) (4, 44) (5, 59) (6, 64) (7, 79) (8, 84)
- **d** (10, 55) (12, 45) (14, 20) (16, 40) (18, 30) (20, 28) (22, 25)

2 The scatterplot shows the mother's height (in cm) and her daughter's height (in cm).

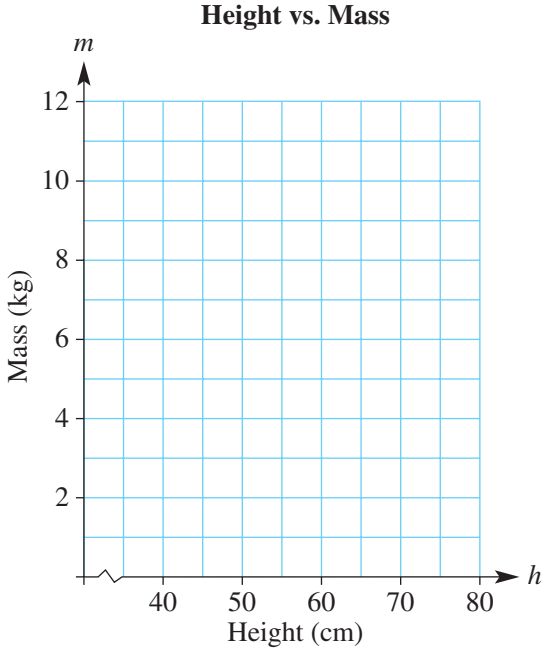
- **a** Copy the scatterplot and draw a line of best fit by eye.
- **b** Describe the strength of the relationship as strong, moderate or weak.
- **c** Estimate the daughter's height if the mother's height is 170 cm.
- **d** Estimate the mother's height if the daughter's height is 162 cm.



3 The height and masses of young children are measured and recorded below.

Height h(cm)	40	45	50	55	60	65	70	75	80	85
Mass m(kg)	1.5	3.1	3.6	5.5	6.0	6.9	7.6	8.6	10.0	11.2

- **a** Complete the scatterplot opposite and draw a line of best fit by eye.
- **b** What is the expected mass of a child given their height is 73 cm?
- **c** What is the expected height of a child given their mass is 4.8 kg?
- **d** What is the expected mass of a child given their height is 48 cm?
- **e** What is the expected height of a child given their mass is 9.0 kg?

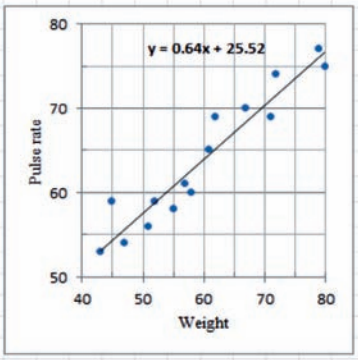


ISBN 978-1-108-44806-2

4 Create the spreadsheet and scatterplot below.

XLS 6CQ4

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1																
2	Worksheet to compare weight to pulse rate.															
3																
4	*Weight (x)*	58	51	43	55	62	79	61	52	47	71	80	57	45	67	72
5	*Pulse rate (y)*	60	56	53	58	69	77	65	59	54	69	75	61	59	70	74
6																



a Use the trendline tool to insert the least-squares line of best fit onto the scatterplot.
b What is the pulse rate when the weight is 65 kg?
c What is the weight when the pulse rate is 60 beats per minute?

5 The table below shows the amount of energy (in megajoules, MJ) used per day for 12 people of various mass (in kg).

Energy (MJ)	1.5	1.6	1.7	1.8	1.9	2.0	2.0	2.1	2.2	2.3	2.4	2.5
Mass (kg)	50	54	70	71	78	88	98	101	110	115	119	125

a Draw a scatterplot using energy for the horizontal axis and mass for the vertical axis.
b Draw a line of best fit by eye.
c What is the mass when 1.55 MJ of energy is used?
d What is the mass when 2.15 MJ of energy is used?
e What is the energy used when the mass is 100 kg?
f What is the energy used when the mass is 80 kg?

ISBN 978-1-108-44806-2

6D Interpolation and extrapolation

The equation of the line of best fit or regression line can provide important information and be used to make predictions. The gradient (m) indicates the change in dependent variable as the independent variable increases by 1 unit. The vertical intercept (b) indicates the value of the dependent variable when the independent variable is zero. In addition to this information, the equation of best fit is used for interpolation and extrapolation.

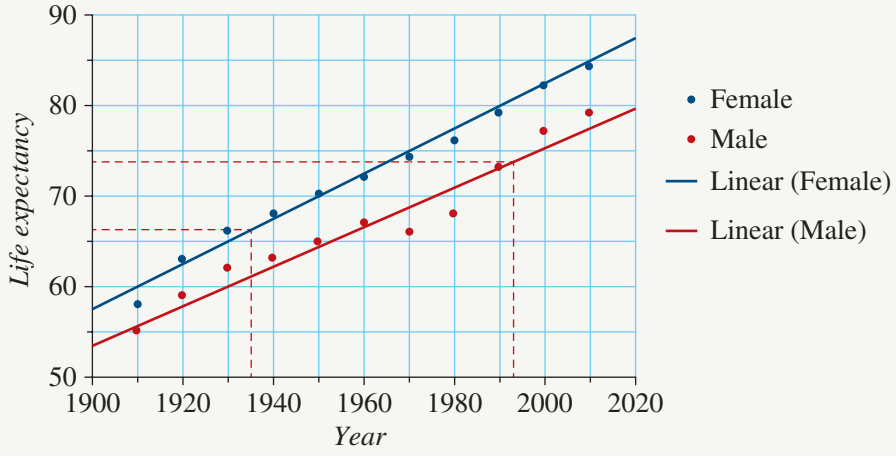
Interpolation

Interpolation is the use of the linear regression line to predict values within the range of the dataset. If the data has a strong linear association then we can be confident our predictions are accurate. However, if the data has a weak linear association, we are less confident with our predictions.

Example 5: Making predictions using interpolation 6D

Life expectancy at birth for females and males is shown below.

Year	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000	2010
Female	58	63	66	68	70	72	74	76	79	82	84
Male	55	59	62	63	65	67	66	68	73	77	79



a What was the life expectancy in 1935 for females?
b What was the life expectancy in 1995 for males?

SOLUTION:

1 Draw a vertical line from 1935 until it intersects the blue line. At this point draw a horizontal line until it reaches the vertical axis. Read the value.
2 Draw a vertical line from 1995 until it intersects the red line. At this point draw a horizontal line until it reaches the vertical axis. Read the value.

a Life expectancy for females in 1935 is approximately 67 years.

b Life expectancy for males in 1995 is approximately 74 years.

ISBN 978-1-108-44806-2

Extrapolation

Extrapolation is the use of the linear regression line to predict values outside the range of the dataset. Predicted values are either smaller or larger than the dataset. The accuracy of predictions using extrapolation depends on the strength of the linear association similar to interpolation. It may not be reasonable to extrapolate too far as this example shows.

Example 6: Making predications using extrapolation 6D

The table below shows the age of a student and their height in centimetres.

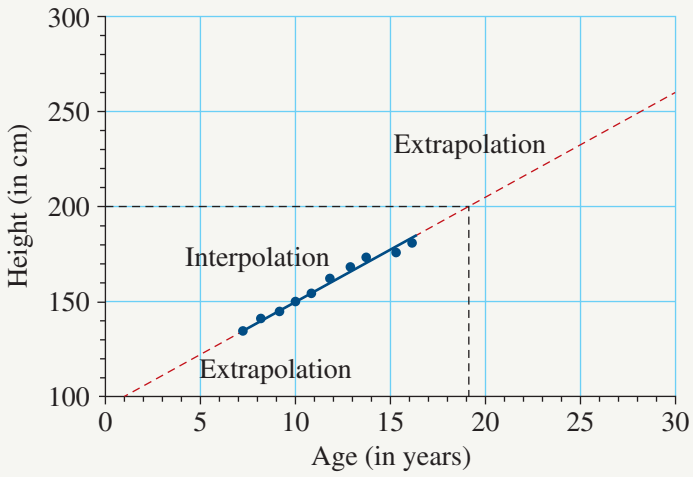
Age (in years)	7	8	9	10	11	12	13	14	15	16
Height (in cm)	133	139	144	149	156	163	170	174	177	181

a Construct a scatterplot from the table using age from 0 to 30 and height from 100 to 300.
b Draw the line of best fit and describe the association between age and height.
c Predict the height of the student when they are aged 19 years.
d What are the limitations of this linear model?

SOLUTION:

1 Draw a number plane with age as the horizontal axis and height as the vertical axis.
2 Determine a scale for the horizontal axis. Let each unit represent 1 year.
3 Determine a scale for the vertical axis. Let each unit represent 10 cm.
4 Write a title for the horizontal and vertical axes.
5 Plot the points (7, 133) (8, 139) (9, 144),…
6 There is a small amount of scatter in the scatterplot.
7 Read the height from the scatterplot when age is 19.
8 Extrapolation too far from the dataset needs to be done carefully.

a



b Strong positive linear association.

c Height of the student is 200 cm when they are 19 years old.

d Adult height does not grow at the same rate as a child. Using the model to extrapolate is flawed, e.g., the prediction is the height will be 260 cm at age 30.

INTERPOLATION	EXTRAPOLATION
Predicting values within the dataset range	Predicting values outside the dataset range

ISBN 978-1-108-44806-2

Exercise 6D

Example 5, 6

1 The scatterplot opposite shows the rainfall (in mm) and the percentage of clear days for each month in 2018.

a How many months had 10% of clear days?

b What was the percentage of clear days when the rainfall was 70 mm?

c Predict the rainfall in the month given the following percentage of clear days:

i 4%

ii 22%

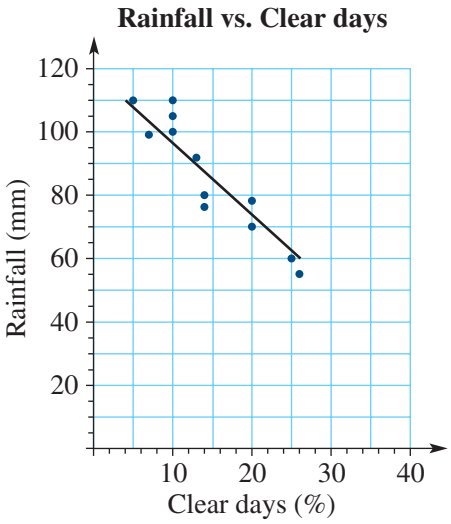
iii 26%.

d Predict the percentage of clear days in the month given the following rainfall:

i 80 mm

ii 90 mm

iii 100 mm.



2 The scatterplot shows the average annual female income plotted against average annual male income for 15 countries.

a What was the female income for a country whose average annual male income was $45 000?

b How many countries had an average annual female income of $25 000?

c Predict the female income given the following male income:

i $20 000

ii $40 000

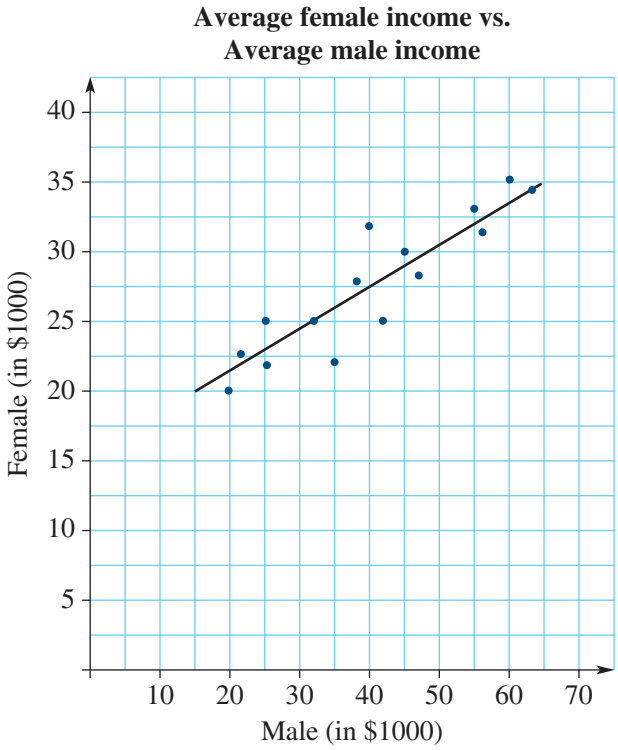
iii $60 000.

d Predict the male income given the following female income:

i $25 000

ii $30 000

iii $35 000.



3 For time ranging from 5 to 25 seconds, the equation relating the number of errors to time is:

$errors = -0.53 \times time + 15$

Use this equation to predict the number of errors (to nearest whole number) with the following times. Are you interpolating or extrapolating?

a 10 seconds **b** 20 seconds **c** 30 seconds

ISBN 978-1-108-44806-2

4 For minimum temperatures from 5°C to 20°C, the equation relating the maximum and minimum temperature (in °C) at a weather station is shown below:

$$maximum = 0.67 \times minimum + 13$$

Use this equation to predict the maximum temperature given the following minimum temperatures. Are you interpolating or extrapolating?

a 10°C
b 20°C
c 30°C

5 The equation relating life expectancy at birth from 1900 to the current year for a particular country is given below:

$$life\ expectancy = 0.21 \times year - 353.78$$

Use this equation to predict a life expectancy in the following years. Are you interpolating or extrapolating?

a 1900
b 1950
c 1870
d 2000
e 2030
f 1970

6 When a person's height is between 160 cm and 190 cm, the equation relating weight (in kg) to the height (in cm) is shown below:

$$weight = 0.75 \times height - 65.63$$

Use this equation to predict a person's weight with the following heights. Are you interpolating or extrapolating?

a 150 cm **b** 175 cm **c** 200 cm

7 When a worker's average pay rate is between \$5 and \$25, the equation relating a country's development index (%) to the average pay rate (in dollars per hour) is shown below:

$$development\ index = 0.272 \times pay\ rate + 81.3$$

Use this equation to predict a country's development index with the following average pay rates. Are you interpolating or extrapolating?

a \$40 per hour **b** \$20 per hour **c** \$10 per hour

8 When the area in a large city is between 1 km^2 and 8 km^2, the equation relating a population to area (in square kilometres) is shown below:

$$population = 2680 \times area + 5330$$

Use this equation to predict the population with the following areas. Are you interpolating or extrapolating?

a 2.5 km^2 **b** 5.0 km^2 **c** 7.5 km^2

ISBN 978-1-108-44806-2

6E Statistical investigation

Statistical investigation is the process of gathering statistics. The information gained from a statistical investigation is a vital part of our society. A statistical investigation involves four steps.

Four steps in a statistical investigation

1. **Collect the data**

   Collecting data involves deciding what to collect, locating it and collecting it. The gathering of statistical data may take the form of a:
   - census – data is collected from the whole population
   - survey – data is collected from a smaller group of the population.

   It is important that procedures are in place to ensure the collection of data is accurate, up-to-date, relevant and secure. If the data collected comes from unreliable sources or is inaccurate, the information gained from it will be incorrect. When taking a sample, the data gathered must be representative of the entire population otherwise the information collected may be biased towards a particular outcome.
2. **Organise the data**

   Organising data is the process of arranging, representing and formatting data. It is carried out after the data is collected. The organisation of the data depends on the purpose of the statistical investigation. For example, to store and search a large amount of data, the data needs to be categorised. Organising gives structure to the data.
3. **Summarise and display the data**

   Displaying data is the presentation of the data and information. Information must be well organised, readable, attractively presented and easy to understand. Information is often displayed using graphs such as scatterplots, dot plots, histograms, line graphs, stem-and-leaf plots and box plots. Data is summarised using statistics such as the mean, median, mode and standard deviation.

4. **Analyse the data**

   Analysing data is the process of interpreting data and transforming it into information. It involves examining the data and giving meaning to it. When analysing bivariate data, the form, direction and strength of the association is determined. Scatterplots and lines of best fit are commonly used to analyse the data. They make it easy to interpret data by making instant comparisons and revealing trends. Predictions and conclusions are completed by interpolating and extrapolating the data.

STATISTICAL INVESTIGATION

A statistical investigation involves four steps: collecting data, organising data, summarising and displaying data and analysing data.

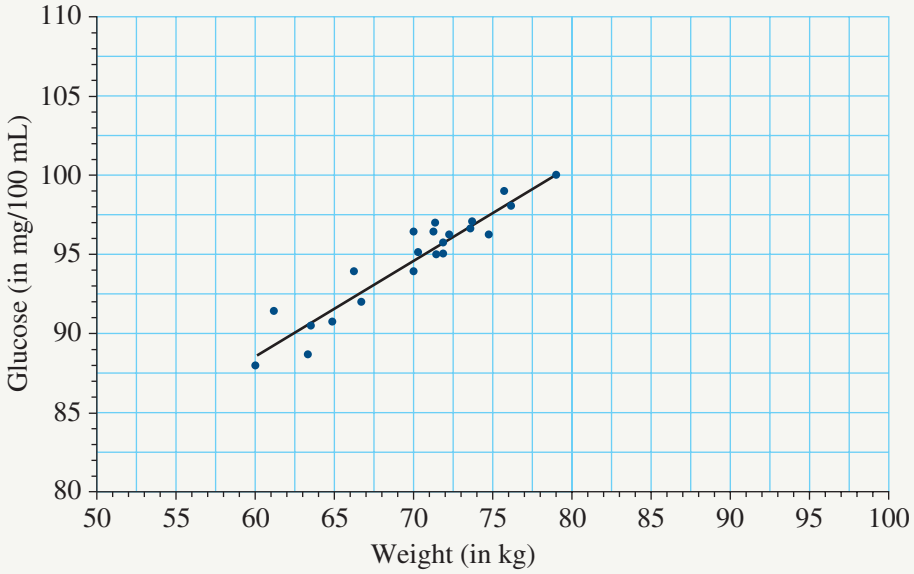
ISBN 978-1-108-44806-2

Example 7: Case study of a statistical investigation 6E

James has been asked to complete a statistical investigation on whether the blood glucose level (in mg/100 mL) of an adult can be predicted from their weight (in kg).
James performed the following steps.

1 Collecting the data – James accessed the medical data on 20 adults.
2 Organising the data – James categorised the data into blood glucose levels and weight.
3 Summarising and displaying the data – James presented the bivariate data into the table shown opposite and the scatterplot shown below.

Weight	Glucose	Weight	Glucose
60.2	88.1	71.6	94.9
61.3	91.5	72.0	95.7
63.5	88.7	72.0	95.1
63.7	90.6	72.5	96.4
65.0	90.9	73.7	96.6
66.4	94.0	73.8	97.0
66.9	92.1	74.8	96.3
70.1	96.5	75.9	99.1
70.2	93.9	76.3	98.2
70.5	95.2	78.9	99.9



4 Analysing the data –

a James calculated Pearson's coefficient to measure the strength of a linear association.

$r = 0.9507479097 \ldots$
This indicates a strong positive linear association between weight and blood glucose levels.

b James calculated the equation and graphed the least-squares regression line on the scatterplot.

$A = 52.78718161 \ldots \quad B = 0.5966957534 \ldots$
$y = mx + b$
$= Bx + A$
$= 0.60x + 52.79$
$glucose = 0.60 \times weight + 53$

c James applied the results of his statistical investigation to predict the glucose level of a person who weighs 75 kg.

$glucose = 0.60 \times 75 + 53$
$= 98$
$\therefore$ A person weighing 75 kg has a blood glucose level of 98 mg/100 mL.

ISBN 978-1-108-44806-2

Issues in a statistical investigation

A statistical investigation raises a number of ethical issues such as bias, accuracy, copyright and privacy.

- Data needs to be free from bias. Bias means that the data is unfairly skewed or gives too much weight to a particular result. For example, if a survey about favourite music was only completed by teenagers, and the results were generalised to the entire population, it would have a bias. Several checks should be made to limit the impact of bias.
- The accuracy of the collected data is a vital ingredient of a statistical investigation. It depends on the source of the data and whether the data has been recorded correctly. The accuracy of the data is often difficult to check in a reasonable time. It is often necessary to compare data from a number of different sources and determine which data is accurate.
- Copyright is the right to use, copy or control the work of authors and artists. It is against the law to infringe copyright. You are not allowed to use or copy the work of another person without their permission. If data is collected from the internet, it should be assumed to be protected by copyright.
- Privacy is the ability of an individual to control personal data. Data collected on individuals is not always accurate. Inaccuracies can be caused by mistakes in gathering or entering the data, by mismatch of the data and the person or by information being out-of-date. Most people give information about themselves to selected parts of the outside world. Often people are quite willing to tell A something but would be shocked if B knew. But what prevents A telling B?

ISSUES IN A STATISTICAL INVESTIGATION

A statistical investigation raises a number of ethical issues such as bias, accuracy, copyright and privacy.

Causation

Causation indicates that one event is the result of the occurrence of another event (or variable). This is often referred to as the cause and effect. That is, one event is the cause of another event happening. For example, the bell at the end of the period is an event that causes students to leave for the next period. When completing a statistical investigation it is important to be aware that two events (or variables) may have a high correlation but be unrelated. That is, high correlation does not imply causation. For example, the increase in the use of mobile phones has a strong correlation to the increase in life expectancy. However, the use of mobile phones does not cause the increase in life expectancy.

CAUSATION

Causation indicates that one event is the result of the occurrence of another event (or variable).

ISBN 978-1-108-44806-2

Exercise 6E

1 Copy and complete the following sentences.

a A statistical _____ involves four steps: collecting data, organising data, summarising and displaying data, and analysing data.

b Census data is collected from the whole _____.

c When taking a _____ the data gathered must be representative of the entire population.

d Displaying data is the _____ of the data and information.

e When analysing _____ data the form, direction and strength of the association is determined.

f Bias means that the data is unfairly _____ or gives too much weight to a particular result.

2 True or false?

a A survey is when data is collected from a smaller group of the population.

b Data collected from unreliable sources results in incorrect information.

c Data is often displayed using graphs such as scatterplots, dot plots, histograms, line graphs, stem-and-leaf plots and box plots.

d Analysing data is the process that interprets data, transforms it into information.

e You are allowed to use or copy the work of another person without their permission.

f Data collected on individuals is always accurate.

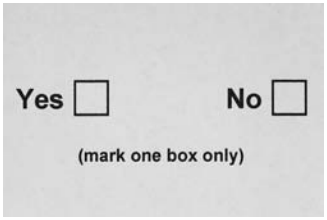
3 The Australian Bureau of Statistics collects data for our society. Collecting data is one step in a statistical investigation. List the four steps involved in a statistical investigation.

4 Explain the difference between a census and a survey.

5 How can you limit the impact of biased data?

6 There is a strong positive correlation between number of car accidents and the number of teachers in cities around the world. Can we conclude from this that teachers are causing car accidents? Give a possible explanation.

7 There is a strong positive correlation between the number of churches in a town and the amount of alcohol consumed by its inhabitants. Does this mean that religion is encouraging people to drink? What common cause might counter this conclusion?



ISBN 978-1-108-44806-2

PPT STUDY GUIDE

Key ideas and chapter summary

Scatterplot

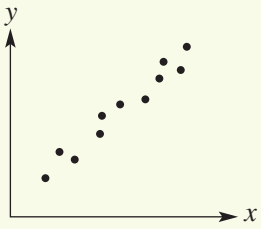
To construct a scatterplot:

1. Draw a number plane.
2. Determine a scale and a title for the horizontal or x-axis.
3. Determine a scale and a title for the vertical or y-axis.
4. Plot each ordered pair of numbers with a dot.

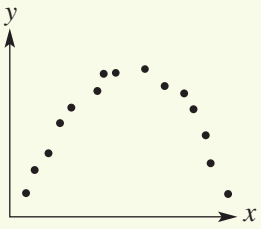
Using a bivariate scatterplot

Form of an association

Linear form – a straight line

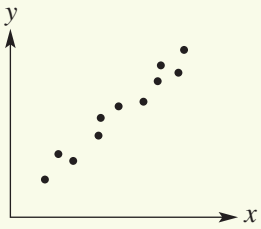


Non-linear form – a curved line

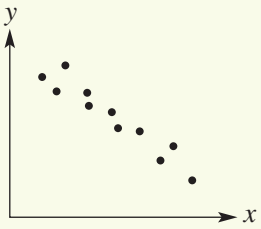


Direction of an association

Positive gradient

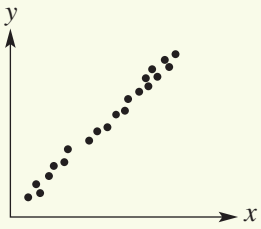


Negative gradient

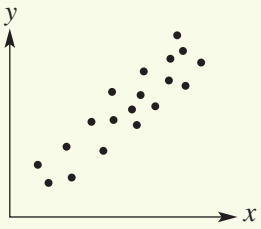


Strength of an association

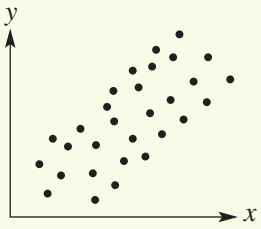
Strong – small amount of scatter



Moderate – modest amount of scatter



Weak – large amount of scatter



Line of best fit by eye

Line of best fit is a straight line that approximates a linear association between points.

Interpolation

Predicting values within the range of the dataset.

Extrapolation

Predicting values outside the range of the dataset.

Statistical investigation

Four steps: collecting data, organising data, summarising and displaying data, and analysing data.

A statistical investigation raises a number of ethical issues such as bias, accuracy, copyright and privacy.

ISBN 978-1-108-44806-2

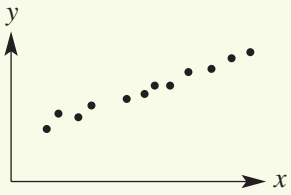
Review

Multiple-choice

1 Blood pressure levels for women increase as they get older. What is the best description for the association between blood pressure levels and a woman's age?

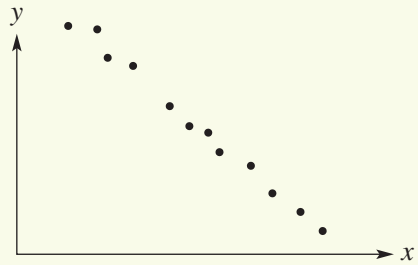
A Positive correlation
B Zero correlation
C Negative correlation
D Constant correlation

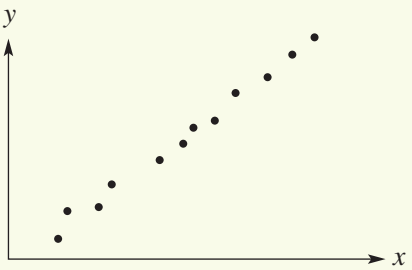
2 What is the correlation between the variables in the scatterplot?

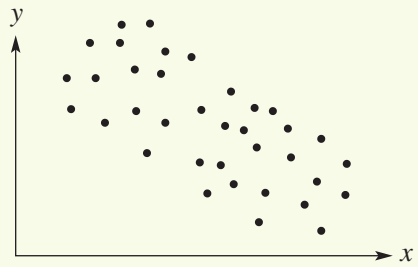


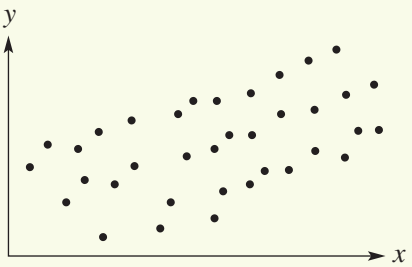
A Strong positive
B Weak positive
C Strong negative
D Weak negative

3 Which of the following scatterplots shows weak negative correlation?

A


B


C


D


4 The birth weight and weight at age 21 of eight women are given in the table below.

Birth weight	1.9	2.4	2.6	2.7	2.9	3.2	3.4	3.6
Weight at 21	47.6	53.1	52.2	56.2	57.6	59.9	55.3	56.7

Which of the following conclusions is correct from the above table?

A There is no clear pattern in the bivariate data.
B The form of the association is linear.
C The direction of the association is negative.
D Strength of the association is weak.

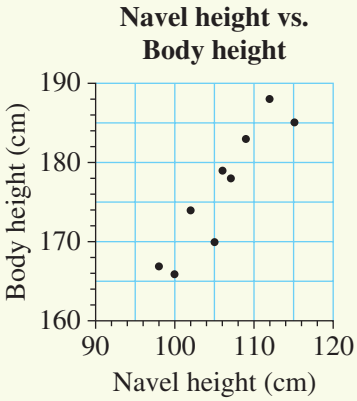
5 The linear equation that enables weekly amount spent on entertainment (in dollars) to be predicted from weekly income is given by: $amount = 0.10 \times income + 40$. What is the predicted amount spent on entertainment with a weekly income of \$600?

A \$40 **B** \$46 **C** \$100 **D** \$240

ISBN 978-1-108-44806-2

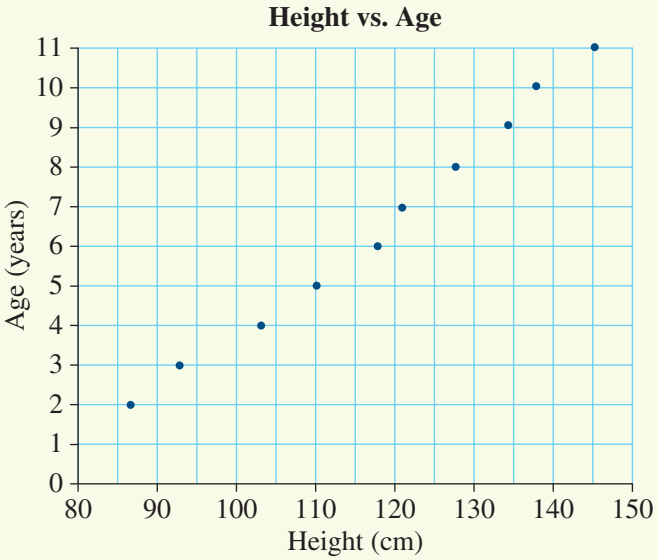
Short-answer

1 The scatterplot shows the navel height and the body height for 9 students.



- **a** Which has been plotted as the independent variable?
- **b** Which has been plotted as the dependent variable?
- **c** Is the association between these two variables linear or non-linear?
- **d** Describe the association as strong, moderate or weak.
- **e** What is the body height for the student with a navel height of 112 cm?
- **f** What is the navel height for the student with a body height of 166 cm?
- **g** Use the scatterplot to predict the body height of a student with a navel height of 110 cm.

2 The scatterplot shows a student's height (in cm) and their age (in years).



- **a** What is the age for the student when their height is 120 cm?
- **b** What is the height for the student whose age is 11 years?
- **c** State whether the association is positive or negative.
- **d** Describe the strength of the association as strong, moderate or weak.

3 The table below shows the length of the right foot (in cm) and body height (in cm).

Length of right foot	27.5	24	22.6	23.7	26.4	27.1	25.5	26.1
Body height	174.4	156	155.3	160.5	170.7	169.3	163.3	164.9

- **a** Draw a scatterplot using the above table.
- **b** State whether the association is positive or negative.
- **c** Describe the strength of the association as strong, moderate or weak.

4 A strong positive linear association exists between the hours spent studying for an exam and the mark achieved. The equation for this association is $mark = 4.5 \times study\ hours + 2$.

- **a** Predict the exam mark if the student studied for 12 hours a week.
- **b** Predict the exam mark if the student studied for 20 hours a week.

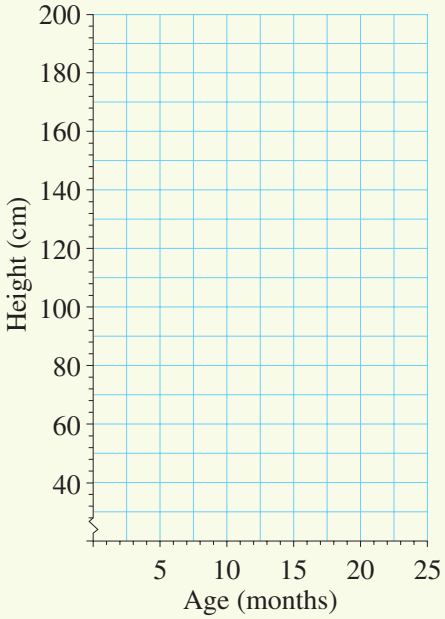
ISBN 978-1-108-44806-2

Review

5 The table below shows the age (in months) and the height (in cm) of a young plant.

Age (in months)	1	4	5	8	12	14	15	19	22	24
Height (in cm)	48	65	78	87	114	128	131	159	169	188

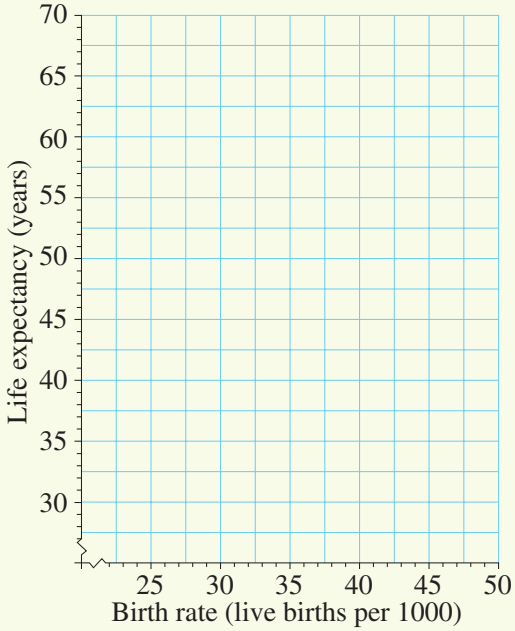
a Complete the scatterplot and draw a line of best fit by eye.
b What is the predicted height of the plant after 11 months? Answer to the nearest centimetre.
c What is the predicted height of the plant after 6 months? Answer to the nearest centimetre.
d Do these questions involve interpolation or extrapolation?



6 The table shows the birth rate (live births per 1000) and the life expectancy (in years).

Birth rate	30	31	34	38	40	42	43
Life expectancy	66	64	46	54	48	45	42

a Complete the scatterplot and draw a line of best fit by eye.
b State whether the association is positive or negative.
c Describe the strength of the association as strong, moderate or weak.
d What is the life expectancy when the birth rate is 35?
e What is the birth rate when the life expectancy is 60?



PDF
LITERACY

ISBN 978-1-108-44806-2