EXERCISE 19.1 TYPES OF DATA AND THEIR INTERPRETATION

2 Total number of Year 7 students is $25+40+24+12+4+2=107$

Total number of Year 12 students is $2+8+25+55+24+6=120$

For the category $0 \dashv < 10$ the calculations would be:

Percentage of Year 7 students: $\frac{25}{107}\times 100\% \approx 23.4\%$

Percentage of Year 12 students: $\frac{2}{120}\times 100\% \approx 1.7\%$

For the category $10 \dashv < 20$ the calculations would be:

Percentage of Year 7 students: $\frac{40}{107}\times 100\% \approx 37.4\%$

Percentage of Year 12 students: $\frac{8}{120}\times 100\% \approx 6.7\%$

For the category $20 \dashv < 30$ the calculations would be:

Percentage of Year 7 students: $\frac{24}{107}\times 100\% \approx 22.4\%$

Percentage of Year 12 students: $\frac{25}{120}\times 100\% \approx 20.8\%$

For the category $30 \dashv < 40$ the calculations would be:

Percentage of Year 7 students: $\frac{12}{107}\times 100\% \approx 11.2\%$

Percentage of Year 12 students: $\frac{55}{120}\times 100\% \approx 45.8\%$

For the category $40 \dashv < 50$ the calculations would be:

Percentage of Year 7 students: $\frac{4}{107}\times 100\% \approx 3.7\%$

Percentage of Year 12 students: $\frac{24}{120}\times 100\% \approx 20\%$

For the category $50 \dashv < 60$ the calculations would be:

Percentage of Year 7 students: $\frac{2}{107}\times 100\% \approx 1.9\%$

Percentage of Year 12 students: $\frac{6}{120}\times 100\% \approx 5\%$

 ISBN 978 1 4886 1829 1

Percentage of Year 7 students	Travel time (minutes)	Percentage of Year 12 students
23.4	$0 \dashv < 10$	1.7
37.4	$10 \dashv < 20$	6.7
22.4	$20 \dashv < 30$	20.8
11.2	$30 \dashv < 40$	45.8
3.7	$40 \dashv < 50$	20
1.9	$50 \dashv < 60$	5

Travel time (minutes)	Percentage of Year 7 students	Percentage of Year 12 students
$0 \dashv < 10$	23.4	1.7
$10 \dashv < 20$	37.4	6.7
$20 \dashv < 30$	22.4	20.8
$30 \dashv < 40$	11.2	45.8
$40 \dashv < 50$	3.7	20
$50 \dashv < 60$	1.9	5

4 (a) For sugar tablet:

Percentage of Cold: $\frac{29}{125} \times 100\% \approx 23\%$

Percentage of No cold: $\frac{96}{125} \times 100\% \approx 77\%$

For Vitamin C:

Percentage of Cold: $\frac{24}{125} \times 100\% \approx 19\%$

 ISBN 978 1 4886 1829 1

Percentage of No cold: $\frac{101}{125} \times 100\% \approx 81\%$

	Sugar tablet	Vitamin C tablet
Cold (%)	23	19
No cold (%)	77	81

(b) A slightly lower percentage of those taking the vitamin C tablets caught colds. The difference small and unlikely to be significant. A much larger study would be needed to research the effectiveness of such products, but it would appear that any effect would be minor..

6 (a) The empty spaces can be filled in a different order. This method goes largely left to right and top to bottom.

Total for 0 aces is $7 + 1 = 8$.

Number of slow servers who served 1 ace is $27 - 7 - 8 - 4 = 8$.

Total for 1 ace is $8 + 8 = 16$.

Number of fast servers who served 2 aces is $22 - 8 = 14$.

Number of fast servers who served 3 aces is $34 - 4 = 30$.

Total for 4 aces is $100 - 8 - 16 - 22 - 34 = 20$.

Number of fast servers who served 4 aces is $20 - 0 = 20$.

Total number of fast servers is $1 + 8 + 14 + 30 + 20 = 73$, or

Total number of fast servers is $100 - 27 = 73$.

	0 aces	1 ace	2 aces	3 aces	4 aces	Total
Slow serve	7	8	8	4	0	27
Fast serve	1	8	14	30	20	73
Total	8	16	22	34	20	100

(b) From the table, a total of 34 players served 3 aces.

(c) More than 2 aces:

There are 73 fast servers, and $30 + 20 = 50$ of these served more than two aces.

 ISBN 978 1 4886 1829 1

This is $\frac{50}{73} \times 100\% \approx 68.5\%$.

There are 27 slow servers, and $4 + 0 = 4$ of these served more than two aces.

This is $\frac{4}{27} \times 100\% \approx 14.8\%$

(d) Fast servers serve more aces than slow servers.

8 (a) Set up intervals $100 - < 150$, $150 - < 200$, … , ≥ 300, and count how many rivers fall into each category. Fill in a table as below.

Length of river (km)	$100 - < 150$	$150 - < 200$	$200 - < 250$	$250 - < 300$	≥ 300	Total
North Island	6	7	1	1	1	16
South Island	5	3	2	1	1	12
Total	11	10	3	2	2	28

(b) From the table, there are 10 rivers in New Zealand in the group of $150 - <200$ km.

(c) This is the three last categories. Add the totals section of the table.

$3 + 2 + 2 = 7$

There are 7 rivers in New Zealand which are longer than 200 km.

(d) Use the rightmost column in the table.

The North Island has 16 rivers which are at least 100 km in length. The South Island has 12 rivers which are at least 100 km in length.

(e) A larger area does not indicate a greater number of substantial rivers, in fact the opposite is true.

EXERCISE 19.2 SCATTERPLOTS AND ASSOCIATION

2 (a) Value of superannuation is dependent on the number of years of employment. So the independent variable is years of employment.

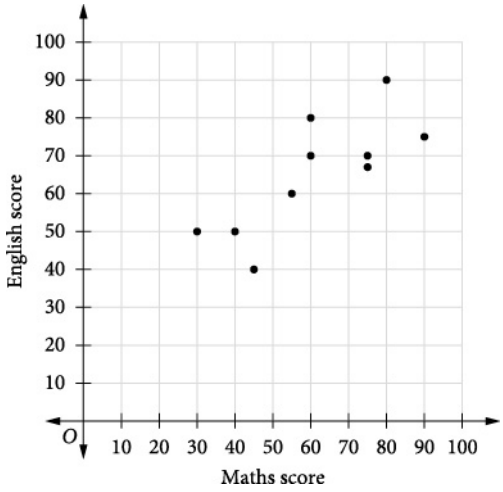
(b) The size of plants is dependent on the amount of rainfall. Rainfall is not affected by the size of plants. So the independent variable is rainfall.

 ISBN 978 1 4886 1829 1

(c) The number of ice-creams sold is dependent on the temperature. The temperature is not dependent on the number of ice-creams sold. So the independent variable is the temperature.

(d) Waist measurement is dependent on the cans of soft drink consumed. So the independent variable is the cans of soft drink consumed.

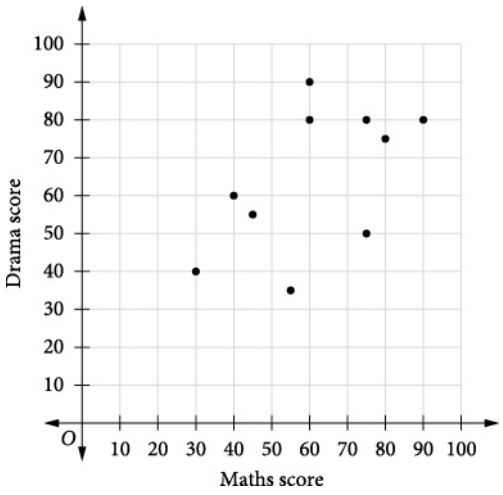
4 (a) Draw up axes and plot the pairs of scores as points.



There is a pattern, which is linear and the slope is positive. The pattern represents the data well.

Therefore, there is a moderate, positive, linear association.

(b) Draw up axes and plot the pairs of scores as points.

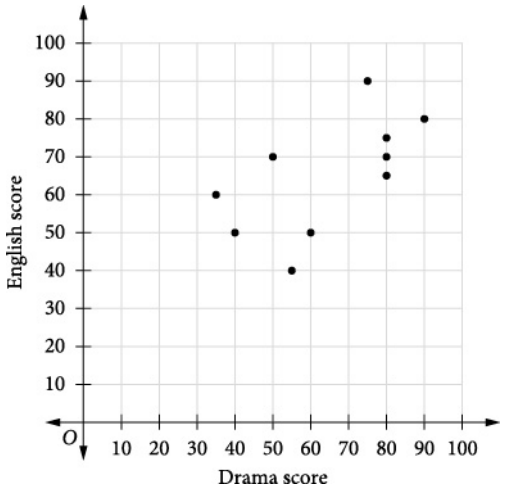


There is only a slight pattern, which is linear and the slope is positive.

Therefore, there is a weak, positive, linear association.

 ISBN 978 1 4886 1829 1

(c) Draw up axes and plot the pairs of scores as points.



There is only a slight pattern, which is linear and the slope is positive.

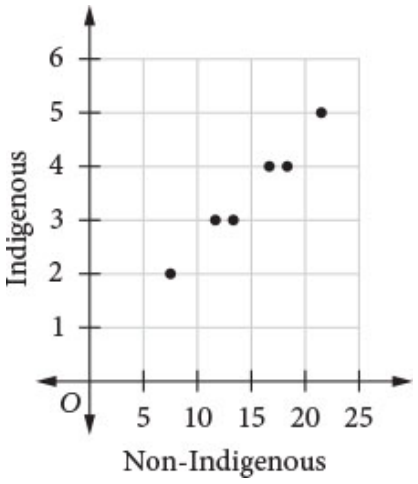
Therefore, there is a weak, positive, linear association.

(d) The Maths versus English points show the strongest association.

6 (a)

Year	1986	1991	1996	2001	2006	2011
Non-Indigenous	7	12	14	16	18	22
Indigenous	2	3	3	4	4	5

(b)



(c) There is a strong pattern, which is linear and the slope is positive. The pattern represents the data well.

Therefore, there is a strong, positive, linear association.

 ISBN 978 1 4886 1829 1

(d) Percentage change:

Non-Indigenous increase: $22-7=15$

Non-Indigenous percentage increase: $\frac{15}{7}\times 100\% \approx 214\%$

Indigenous increase: $5-2=3$

Indigenous percentage increase: $\frac{3}{2}\times 100\% = 150\%$

The non-Indigenous population has a greater percentage increase in higher-education engagement.

EXERCISE 19.3 CALCULATING THE CORRELATION COEFFICIENT

2 (a) D

There is a pattern, which is linear and the slope is positive. A trend line with a positive gradient would fit the data fairly well: moderate, positive, linear.

(b) A

Must be positive, and since there is a moderate trend, 0.2 would be too small. A moderate, positive, linear trend gives $r \approx 0.5$.

4 (a) From technology:

$r^2 = 0.2156$, trend line has a positive gradient.

$r = \sqrt{0.2156}$
≈ 0.46

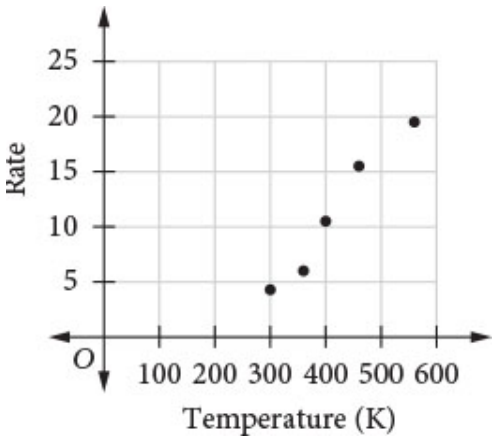
(b) The outlier is $(182,\ 500)$, since 182 is extremely different to the other reaction times. It is possible it should have been 18.2.

(c) A

Outliers are generally excluded from the data set, so calculations are an accurate representation of the study

 ISBN 978 1 4886 1829 1

(d)



From technology:

$r^2 = 0.9735$, the trend line has a positive gradient.

$r = \sqrt{0.9735}$

≈ 0.99

(e) This is evident since r is 0.46 when the outlier is included, and 0.99 when it is (correctly) excluded.

(f) The number 182 (mol/min) was most likely meant to be 18.2 (mol/min).

6 **(a)** $r^2 = 0.64 = 64\%$. 64% of the change in the quality of performance is due to the change in hours of rehearsal. 36% of the change is due to other factors.

(b) The number of fast food outlets is not dependent on the number of hospitals in a town. They probably both depend on another variable, the population of the town.

(c) The number of registered cars is not dependent on the number of registered motorcycles. They probably both depend on another variable, e.g., the population or the number of adults.

(d) The results in maths and physics exams depend upon other factors such as interest, hours of study and intelligence.

(e) As the number of episodes increase, the number of contestants who must leave the show increases, so there will be a (negative) causation. $r^2 = 0.96 = 96\%$. 96% of the change in the number of contestants remaining in a reality TV series is due to the change in the episode number. 4% of the change is due to other factors.

 ISBN 978 1 4886 1829 1

EXERCISE 19.4
MODELLING BY FINDING THE EQUATION OF THE LINE OF BEST FIT

2 (a) Using technology, we can see that the points line up perfectly giving a trend line with negative gradient. $r^2 = 1,\ r = -1$

Using technology, the least square regression equation is $y = -6.5x + 15$ where y is the temperature (°C) and x is the altitude (km).

(b) The altitude is the independent variable and the temperature is the response variable. There is a perfect, negative, linear association between the variables. For every increase in altitude of one kilometre, the temperature decreases by $6.5°C$.

4 (a) The fixed costs are the cost if there were no teams and hence no variable costs.

From the graph, this is the vertical intercept, $\$875$.

The fixed costs are $\$875$.

(b) The extra cost per team is the gradient.

Choose two points that can be read easily from the graph.

The marked points are $(0,\ 875)$ and $(20,\ 3900)$

Find the gradient.

$$\begin{aligned} m &= \frac{3500 - 2000}{17.5 - 7.5} \\ &= \frac{1500}{10} \\ &= 150 \end{aligned}$$

The extra cost for every team is \$150.

Your answers may vary slightly from these, since they depend on reading numbers off a graph.

6 (a) D

The gradient is positive and the s-intercept seems to be below the p-axis which suggests either option **A** or **D**.

Check for option **A** by substitution in $s = 0.9p - 2$:

$p = 5$

$$\begin{aligned} s &= 0.9 \times 5 - 2 \\ &= 2.5 \end{aligned}$$

$p = 35$

 ISBN 978 1 4886 1829 1

$s = 0.9 \times 35 - 2$
$= 29.5$

Check for **D** by substitution in $s = p - 1$:

$p = 5$

$s = 5 - 1$
$= 4$

$p = 35$

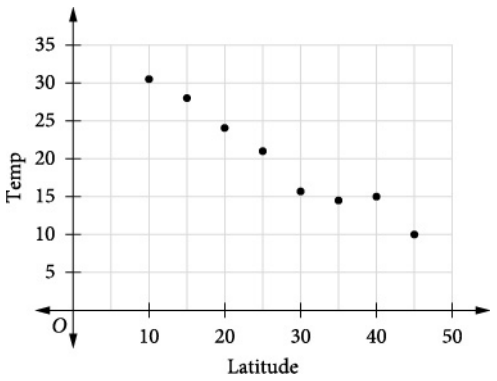
$s = 35 - 1$
$= 34$

(b) **B**

The lower part of the graph will move very little. The upper part will move to the left making the graph slightly steeper.

8 **(a)** The information states that the average temperature is affected by latitude, so latitude is the independent variable.

Plot the pairs of numbers as points.



(b) There is a strong, negative, linear association. As the latitude increases the average maximum temperature decreases.

(c) Using technology, $r^2 = 0.9584...,\ r = -0.9790... \approx -0.98.$

(d) Using technology, $y = -0.6x + 36.0$

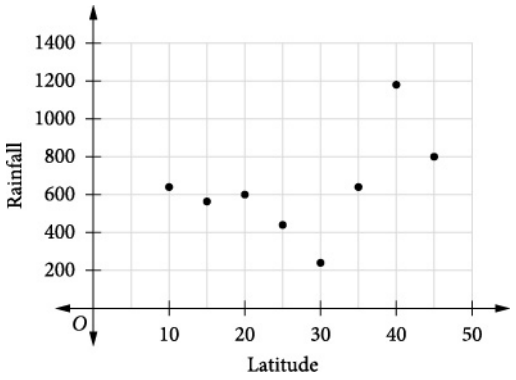
Average maximum temperature $(°C) = -0.6 \times$ latitude $(°S) + 36.0$

(e) Calculate the Pearson correlation for the other association and compare.

Using technology, for Latitude and average rainfall, $r^2 = 0.2131...,\ r \approx 0.46.$

 ISBN 978 1 4886 1829 1

You could also draw a scatterplot.



Comparing scatterplots, or correlation coefficients shows that the correlation with temperature is much stronger.

(f) Interpolation must lie within the data set, so for latitudes from $10°$ to $45°$.

(g) Using technology, $r^2 = 0.2131...,\ r = 0.461... \approx 0.46.$

EXERCISE 19.5 THE STATISTICAL PROCESS

2 (a) False

Most people would not attend a football match if they weren't interested in at least one sport (football) If surveying people at a football match, you will be surveying people who are most likely passionate about sport, therefore the survey would be biased. Your survey will also be biased against people who like sports other than football.

(b) True

Statistics can be manipulated to support data, therefore providing support where in fact there is none. When conducting statistical analysis, you need to be careful to ensure that the data is not biased and that the analysis is accurate and fair.

(c) False

The definition of a population is all of the people in a particular environment or situation, or every member of the relevant group.

(d) False

For a survey to produce useful results the questions need to be carefully considered. Open questions for example may not provide useful information, however closed questions may provide useful data which can be analysed. It is also possible that valid results may not be useful, for example, when there is a low correlation, and useful predictions are not possible.

(e) True

 ISBN 978 1 4886 1829 1

Bias must be avoided when creating a survey so that any results of the data analysis are fair and a true and accurate representation of the sample taken.

(f) True

Outliers can affect statistical calculations and should be removed before completing statistical calculations. They will skew the data unfairly and lead to misleading results.

4 The analysis of this data uses the data in the question. All lengths are in millimetres.

Person	Forearm (mm)	Right foot (mm)		Person	Forearm (mm)	Right foot (mm)
1	265	280		11	225	230
2	249	249		12	240	238
3	248	235		13	268	250
4	254	275		14	280	274
5	248	230		15	292	268
6	264	250		16	234	253
7	228	224		17	277	256
8	237	228		18	303	305
9	244	263		19	285	273
10	286	270		20	254	234

The length of the forearm is the explanatory variable, x, and the length of the right foot is the response variable, y.

Calculate the regression statistics (rounded to 2 decimal places):

$m = 0.76$

$c = 58.28$

$r^2 = 0.62$

$r = 0.79$

 ISBN 978 1 4886 1829 1

As the value of r is significant there is an association between the variables. Based on the value of r, the relationship is considered to be moderate, positive and linear.

There is a moderate, positive and linear relationship between the length of your forearm and the length of your right foot.

6 Possible investigations are comparisons between:

- smokers' age vs height and non-smokers' age vs height
- smokers' age vs lung capacity and non-smokers' age vs lung capacity
- smokers' age vs smokers' sex and non-smokers' age vs smokers' sex

In your report:

- Include scatterplots of the data and find regression lines for each set of data.
- Calculate the r value for each data set and comment on the association between the variables.
- Complete a residual analysis of the data.
- Complete any transformations on the data and find the best fit for each data set.
- Comment upon your findings.
- Use your statistical analysis to support your discussion.

Answers will vary and will need to be checked by your teacher.

 ISBN 978 1 4886 1829 1

CHAPTER REVIEW 19

2 B

There is a clear trend, but a curve, so the trend is non-linear.

4 C

There is no pattern. Therefore, there is no association between the variables. The correlation will be very low, so $-0.2 < r < 0.2$.

6 C

The negative value for the correlation coefficient means the gradient of the trend line is negative, so increasing the number of ants decreases the amount of food.

8 B

$r^2 = 0.64$

$r = \pm\sqrt{0.64} = \pm 0.8$

10 D

Because the new position of the point will put it closer to the line of best fit for the other points, the association will be stronger. The gradient remains negative and increases in magnitude so the value will be closer to -1.

12 (a) D

$r = 0.6$

$r^2 = 0.36$

so 36% of the variation in Helena's success is dependent on her first serve.

(b) A

$r = 0.8$

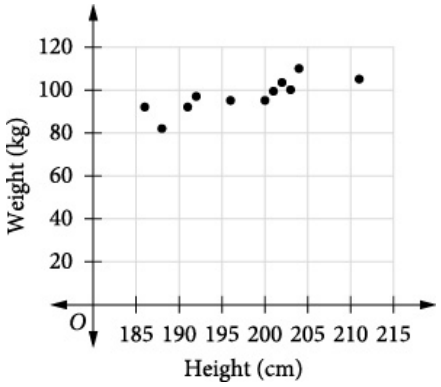
$r^2 = 0.64$

so 64% of the variation in the number of hospital beds is dependent on the population of a town.

 ISBN 978 1 4886 1829 1

14 (a) Weight depends on height, so weight is the dependent variable.

(b) This scatterplot shows a moderate to strong positive, linear association.



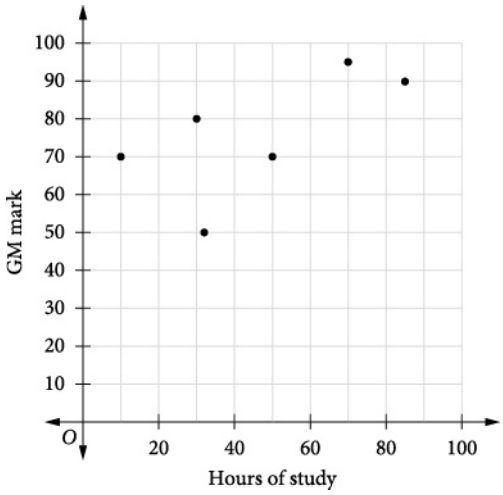
(c) Using technology, $r = 0.83$

(d) Using technology, $y = 0.7903x - 59.017$

So $w = 0.8h - 59$ where w is weight (kg) and h is height (cm).

16 (a) Exam marks are dependent on the total hours of study, so the total hours of study is the independent variable.

(b) Plot the pairs of corresponding values as points.

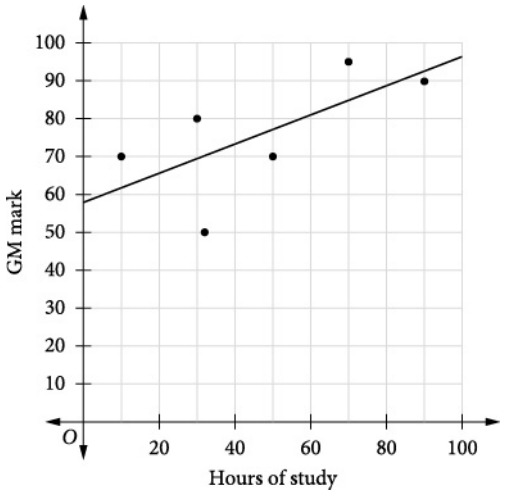


(c) The association is moderate, positive and probably linear.

(d) As the association is moderate, you could estimate around $r = 0.44$ to $r = 0.66$.

 ISBN 978 1 4886 1829 1

(e)



General maths mark $= 0.4\times$ hours of study $+57$. (Answers may vary)

(f) Because the correlation is only 0.66, the equation will only give a rough estimate of the expected mark.

(g) (i) English mark $= 0.58\times$ hours of study $+48$

(ii) Using technology, $r \approx 0.83$ (2 d.p.)

(h) 40 hours of study: English mark is $0.58\times 40 + 48 = 71.2$

60 hours of study: English mark is $0.58\times 60 + 48 = 82.8$

120 hours of study: English mark is $0.58\times 120 + 48 = 117.6$

(i) The prediction for 120 hours of study is outside the data set and is an extrapolation. In this case it is not possible to get a score of 117.6 out of 100.

(j) Both comparisons have a positive correlation; this indicates that more study is likely to improve your results.

(k) There is a clear association, but the association will have some confounding variables that are not controlled. For example, the effectiveness of the study time and the distribution of hours spent on each individual subject (only the total is given). For example, studying only a large number of hours just before the exam may be unproductive.

Explanations may vary.

 ISBN 978 1 4886 1829 1