Solutions to Exercise 15A

1a categorical (each day of the week forms a discrete category)

1b numeric and continuous. But 'height correct to the nearest mm' is numeric and discrete.

1c numeric and continuous. But 'age in years' is numeric and discrete.

1d categorical by party or political code. This would need to be defined carefully — if a person can be affiliated to two parties, it would not be a function.

1e categorical (the categories are red and blue)

1f categorical (the categories are male and female)

1g numeric and discrete (will be an integer between 2 and 12)

1h Shoe sizes are often arranged into categories.

1i These are frequently integers from 1–100, that is, numeric and discrete. If results are reported by a grade, for example, A, B, C, . . . , this might be considered categorical.

2a median 14, mode 14, range 8

2b median 10, every score is trivially a mode, range 12

2c median 8, mode 3, range 12

2d median 6.5, mode 4 & 6, range 6

2e median 4, mode 4, range 7

2f median 5.5, mode 2 & 3 & 9, range 8

3a

score x	1	2	3	4	5	6	7	8
frequency f	4	3	4	2	1	1	1	6
cumulative	4	7	11	13	14	15	16	22

3b There are 22 datapoints so the median will be an average of the 11th and 12th datapoints. This is $\frac{3+4}{2} = 3.5$

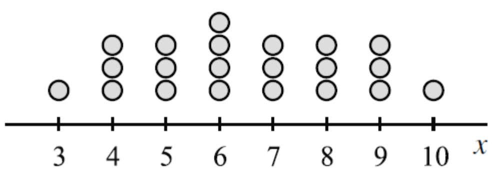
3c 8 (as there are 6 instances of this). Note in a frequency table like this you just need to look for the value with the tallest peak.

3d i This is a median, but it might be more useful to use the mode in this case. It may be easier to develop a square box for four cupcakes rather than three.

3d ii See the previous comments. It is also common for sales to package a larger box to encourage customers to overbuy.

3d iii This is the mode, but if a box of four is marketed, customers can just pick up two boxes of four.

4a



4b

score x	3	4	5	6	7	8	9	10
frequency f	1	3	3	4	3	3	3	1
cumulative	1	4	7	11	14	17	20	21

4c There are 21 datapoints so the median will be at the 11th datapoint. Looking at the cumulative frequency of the above table, this occurs when the score is 6 hoops.

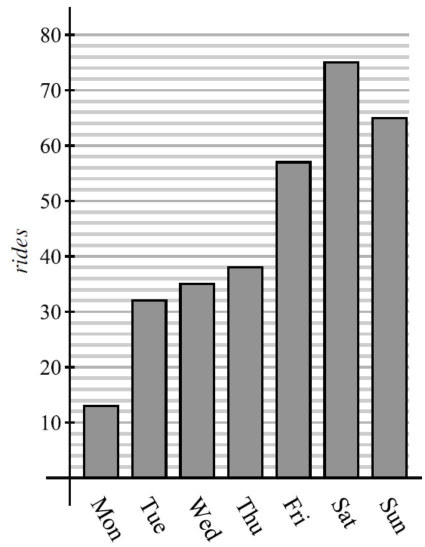
4d The new frequency table will be:

score x	3	4	5	6	7	8	9	10	11
frequency f	1	3	3	4	3	3	3	1	1
cumulative	1	4	7	11	14	17	20	21	22

There are 22 datapoints so the median will be at the average of the 11th and 12th datapoint. Looking at the cumulative frequency of the above table, this will be the average of 6 and 7 which is 6.5 hoops.

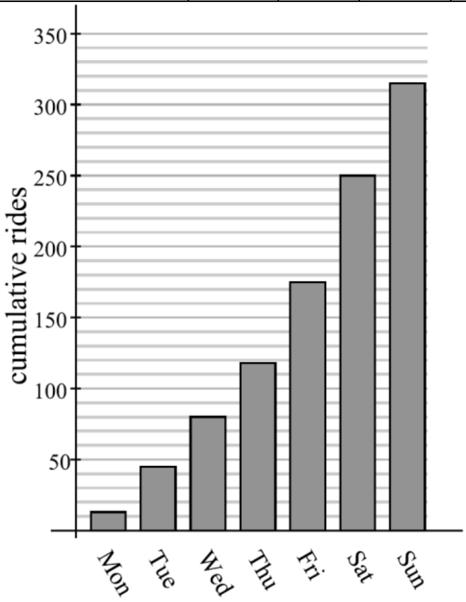
4e Not really. If the scores are ordered by time, his scores improve over the sessions. This information is lost in the table and plot.

5a



5b

Day	Mon	Tue	Wed	Thu	Fri	Sat	Sun
frequency f	13	32	35	38	57	75	65
cumulative	13	45	80	118	175	250	315



6a Blond hair and blue eyes (note that different results might be expected in a different part of the world due to differing genetic and environmental factors).

6b Red hair and green eyes (there are only 3 instances of this combination).

6c $P(\text{blue eyes}|\text{blond})$

$$= \frac{f(\text{blue eyes and blond hair})}{f(\text{blond hair})}$$

$$= \frac{324}{728}$$

$$\doteqdot 45\%$$

6d $P(\text{blue eyes}|\text{black hair})$

$$= \frac{f(\text{blue eyes and black hair})}{f(\text{black hair})}$$

$$= \frac{9}{54}$$

$$\doteqdot 17\%$$

6e $P(\text{brown eyes}|\text{black hair})$

$$= \frac{f(\text{brown eyes and black hair})}{f(\text{black hair})}$$

$$= \frac{25}{54}$$

$$\doteqdot 46\%$$

6f $P(\text{brown eyes}|\text{dark hair})$

$$= \frac{f(\text{brown eyes and black hair}) + f(\text{brown eyes and brown hair})}{f(\text{black hair}) + f(\text{brown hair})}$$

$$= \frac{65 + 25}{193 + 54}$$

$$= \frac{90}{247}$$

$$\doteqdot 36\%$$

6g Using the same reasoning as in part f,

$P(\text{blue eyes}|\text{light hair})$

$$= \frac{324 + 252 + 74 + 10 + 8 + 4}{728 + 25}$$

$$= \frac{671}{752}$$

$$\doteqdot 89\%$$

6h These two results would suggest so. Geneticists link this to various pigment genes that affect both characteristics.

6i The proportion of the various eye and hair colours will vary in different genetic populations and ethnic groups. Studies such as this may be done with a relatively non-diverse population to prevent the clouding effects of differing genetics.

7a $26 + 10 + 7 + 16 + 21 = 80$ orders

7b

Salad	Pie	Soup	Panini	Burger
$\frac{26}{80} \times 100\%$ $= 32.5\%$	12.5%	8.75%	20%	26.25%

7c

Menu item	Salad	Pie	Soup	Panini	Burger
Frequency	26	10	7	16	21
$ Markup	5	6	10	6	8
Profit (frequency $\times$ markup)	$130	$60	$70	$96	$168

7d Total profit $= \$130 + \$60 + \$70 + \$96 + \$168 = \$524$

7e It returns more money than the more popular pie option. It is probably also important for the café to include a vegetarian option on the menu to cater for such customers or for groups with such customers.

8a In 2002 the price was $400 thousand, and in 2017 it was $1 million.

8b Prices increased by

$$\frac{1\,000\,000 - 400\,000}{400\,000}$$

$$= \frac{600\,000}{400\,000}$$

$$= 1.5$$

$$= 150\%$$

8c Average increase

$$= \frac{600\,0000}{15}$$

$$= 40\,000$$

$$= \$40 \text{ thousand per year}$$

8d They will increase another $(13 \times \$40\,000 =)$ \$520 000 to around \$1.5 million.

8e In 2014, median Sydney House Prices were \$760 thousand and in 2015, median Sydney House Prices were \$880 thousand. Hence, from 2014 to 2015, median house prices increased \$120 thousand.

8f From 2010 to 2011, median house prices decreased \$40 thousand.

9a 35%, 140 dogs (both read directly from graph)

9b There are 5% rabbits and 6% Guinea pigs (reading from the graphs). Together, this gives 11%.

9c Dogs, cats and birds are the 3 most common pets.

They form $35\% + 25\% + 15\% = 75\%$ of the pets.

9d The three least common pets are guinea pigs, rabbits and reptiles. Together, they form 6% + 5% + 4% = 15% of the population.

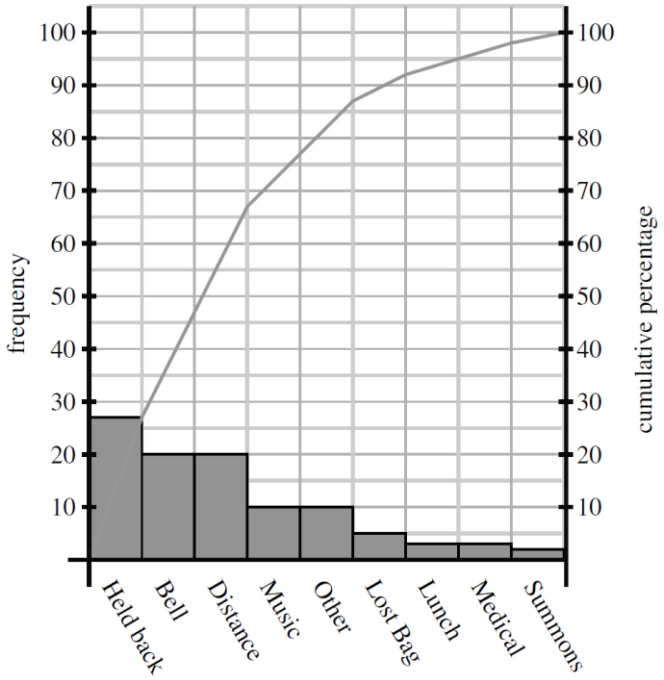
9e This is quite a large category, and it may be that more investigation should be done to see if there were any other popular types of pets lumped into this category.

9f Some pets may require more care and attention. For example, dogs may require frequent exercise and attention. This may give an opportunity for 'value adding' if owners are willing to pay for it. They should also consider what other pet boarding facilities are in the area, because it may be better to pick up a niche market, not covered by other pet boarding houses. Some pets may also be able to use the same types of accommodation, for example, rabbits and guinea pigs.

10a

Reason	frequency	cumulative
Held back	27	27
Bell	20	47
Distance	20	67
Music	10	77
Other	10	87
Lost bag	5	92
Lunch	3	95
Medical	3	98
Summons	2	100

10b



10c The categories are arranged in descending order, so the function will be increasing (if every frequency is greater than zero), but by less at nearly every stage, causing it to curve downwards.

10d 67% (this can be read directly off the cumulative percentage for distance)

10e Remind teachers to release students promptly, increase the volume of the bell or the number of locations where the bell sounds, timetable students in rooms closer together where possible.

11a 6% (read directly from the graph)

11b 64% (read off the cumulative frequency value at the 3rd most common colour – Grey)

11c Reading from the cumulative frequency graph, the seven most popular colours (up to blue) make up 95% of cars.

Hence $100\% - 95\% = 5\%$, i.e. five cars, are not one of the seven most popular colours.

11d Care is needed when the graph is read in a hurry. Compare this with the Pareto chart later in this exercise where both axes are the same scale.

12a The vertical origin is not at a 0% unemployment rate. This exaggerates the scale of the graph, which only shows a variation of 0.25%. This is still potentially significant, but it is only shown over a four-month period, so it is impossible to examine long-term trends. There are natural cycles — for example, there may be a rise when school pupils enter the employment market, and a drop when Christmas provides short-term retail employment. January may be a low point in economic indicators, before businesses return from holidays and begin to hire staff.

12b There has been a significant increase over this five-year period, but more questions need to be asked by someone viewing the graph. What does the vertical scale represent — is it spending per citizen or spending per household? If it is per household, have the household structures changed over the period, such as more larger households? Is this a small community, in which case the data won't be very robust to changes in population? Is the data collected from sales at local shops, and does it include tourists and people passing through — has there been an increase in tourism, and was the data collected at the same time of year (more takeaways may be sold at the height of the tourist season)? What is included in the category of 'takeaway food' — if this is a health study, takeaway salads may be considered healthier than takeaway burgers (which the graphic is trying to suggest). Finally, note that the eye interprets the increase by the size of the graphic, but in fact it is the height that holds information, suggesting a greater increase than was actually the case.

12c i People who do not have access to the internet, or do not feel as comfortable accessing and filling in an online survey, will not be represented. This may be more prevalent amongst older demographics.

12c ii The group should look at other hospitals, unless they particularly want to investigate the change in costs at their local hospital. Hospital costs could be influenced by government policy increasing the staffing numbers at the hospital, by purchase of new expensive diagnostic equipment, by opening and closing particular hospital wards (possibly relocating them to other hospitals), by quality control improvements, by industrial action of staff, and so on. The group likely will want to investigate the cause of any changes to overall expenses and may want to produce graphs of particular expenses, such as doctors' fees. They need to be clear what questions they actually want to ask — for example, are they concerned that medical treatment is getting more expensive for certain sections of the community who cannot afford it?

13a The three most common languages are Mandarin, Spanish and English. Together this gives $30\% + 15\% + 13\% = 58\%$ of the 40% who speak one of the three most common languages as a first language.

13b $40\% \times 7.7 \text{ billion} \doteqdot 3 \text{ billion}$

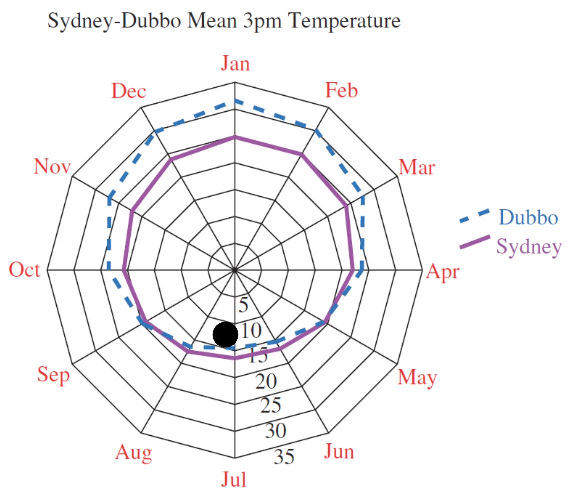
13c $30\% \times (40\% \times 7.7 \text{ billion}) = 30\% \times 3 \text{ billion} \doteqdot 0.92 \text{ billion}$

13d $40\% \times 13\% = 5.2\%$

13e It may be of some use if choosing a major world language is a consideration, but there are often other considerations in deciding what language to learn. For example, you may have relatives who speak Malay, or a girl-friend who is French, or you may want to learn Japanese because of Japan's importance to Australia's economy. Others learn languages for academic reasons, such as Latin because of its historical and linguistic importance, or Russian to study Russian literature. When deciding a language on the number of speakers, it is probably more useful to consider the total number of speakers, not merely those who speak it as a first language — close to a billion people speak English, but only a third of them do so as a first language.

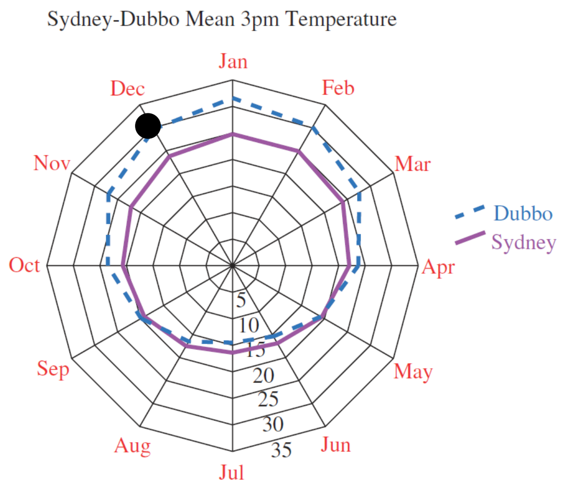
14a i 15°C

Note that this is established from the graph by reading the value from the point shown on the graph.



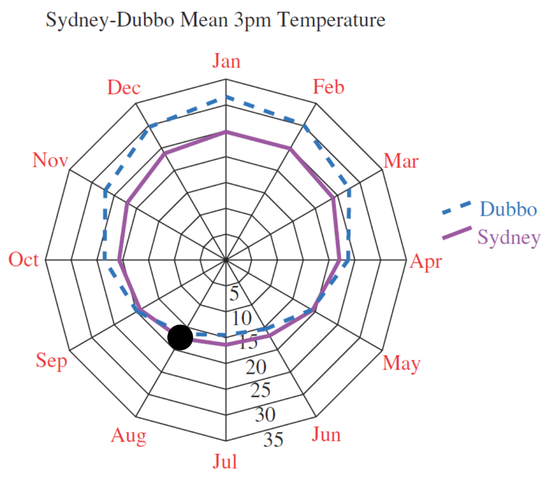
14a ii 30°C

Note that this is established from the graph by reading the value from the point shown on the graph.



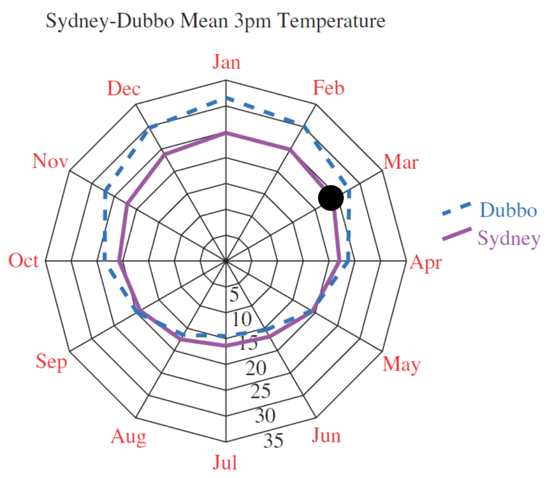
14b i 17°C

Note that this is established from the graph by reading the value from the point shown on the graph.



14b ii 23°C

Note that this is established from the graph by reading the value from the point shown on the graph.



14c Around 6-7°C in December-January.

14d September and May (when the lines intersect)

14e November-February

14f June-August (when the blue line is closer to the centre)

14g Various answers are acceptable.

14h Colour-blind readers may find the colours difficult to distinguish. Using dashes and colour also provides two visuals cues for the bulk of readers, making the graph easier to read.

15a There were 3 sections each out of 10, so the total possible score is $10 \times 3 = 30$.

15b The highest score is $\frac{22}{30} = 73\%$ and the lowest score is $\frac{8}{30} = 27\%$.

15c Bill on Essay writing, Claire on Interpretation, Ellie on all sections.

15d $\frac{4}{10} = 40\%$

15e $\frac{5}{10} = 50\%$

15f To meet the first condition of leaving the class, students must achieve a minimum total score of $55\% \times 30 = 16.5$. Aaron, Claire and Dion meet this first requirement. Students must also score at least $\frac{5}{10}$ for each section. Notice that Claire has not reached 50% in the Interpretation section, as she only scored $\frac{4}{10}$. Therefore, only Aaron and Dion will leave.

16a Yes, this is a reasonable interpretation, provided that similar levels of levels of postgraduates survive to the 55-64 age bracket (otherwise similar numbers may have attained postgraduate degrees in the past and a significant have passed since obtaining the qualification).

16b $100\% - 38.9\% = 61.1\%$

The number with post-school qualifications is $(100\% - \text{those without})$.

16c The number of Australians with post-school qualifications aged 15-64 is $(100\% - 43.5\%) \times 14\,848.1 = 8389.1765$ thousand.

Using part b, the number of Australians with post-school qualifications aged 45-54 is $61.1\% \times 2969.3 = 1814.2423$.

Hence, the probability that a randomly chosen 15-64 year old with a post-school qualification lies in the 45-54 age bracket is

$$\frac{1814.2423}{8389.1765} \doteqdot 21.6\%$$

.

Solutions to Exercise 15B

1a

$$\bar{x} = \frac{\sum xf}{n} = \frac{70}{10} = 7$$

$$\text{Var} = \frac{\sum(x-\bar{x})^2 f}{n} = \frac{36}{10} = 3.6$$

$$s = \sqrt{\text{Var}} = \sqrt{3.6} \doteqdot 1.9$$

x	f	xf	$(x-\bar{x})^2$	$(x-\bar{x})^2 f$
3	1	3	16	16
5	1	5	4	4
6	1	6	1	1
7	3	21	0	0
8	2	16	1	2
9	1	9	4	4
10	1	10	9	9
Total	10	70		36

1b

$$\bar{x} = \frac{\sum xf}{n} = \frac{70}{10} = 7$$

$$\text{Var} = \frac{\sum(x)^2 f}{n} - \bar{x}^2 = \frac{526}{10} - 7^2 = 3.6$$

$$s = \sqrt{\text{Var}} = \sqrt{3.6} \doteqdot 1.9$$

x	f	xf	x^2f
3	1	3	9
5	1	5	25
6	1	6	36
7	3	21	147
8	2	16	128
9	1	9	81
10	1	10	100
Total	10	70	526

2a

x	f	xf	x^2f
12	1	12	144
14	1	14	196
16	1	16	256
17	1	17	289
19	1	19	361
21	1	21	441
22	1	22	484
23	1	23	529
Total	8	144	2700

$$\bar{x} = \frac{\sum xf}{n} = \frac{144}{8} = 18$$

$$\text{Var} = \frac{\sum x^2 f}{n} - \bar{x}^2 = \frac{2700}{8} - 18^2 = 13.5$$

$$s = \sqrt{13.5} \doteqdot 3.67$$

2b

x	f	xf	x^2f
2	1	2	4
3	3	9	27
6	2	12	72
7	1	7	49
8	3	24	192
9	2	18	162
10	2	20	200
13	1	13	169
Total	15	105	875

$$\bar{x} = \frac{\sum xf}{n} = \frac{105}{15} = 7$$

$$\text{Var} = \frac{\sum x^2 f}{n} - \bar{x}^2 = \frac{875}{15} - 7^2 = 9.333\ldots \doteqdot 9.33$$

$$s = \sqrt{9.333\ldots} \doteqdot 3.06$$

2c

x	f	xf	x^2f
40	1	40	1600
49	1	49	2401
50	2	100	5000
51	1	51	2601
54	1	54	2916
57	3	171	9747
60	1	60	3600
65	1	65	4225

70	1	70	4900
Total	12	660	36 990

$\bar{x} = \frac{\sum xf}{n} = \frac{660}{12} = 55$

$\text{Var} = \frac{\sum x^2 f}{n} - \bar{x}^2 = \frac{36\,990}{12} - 55^2 = 57.5$

$s = \sqrt{57.5} \doteqdot 7.58$

2d

x	f	xf	$x^2 f$
7	1	7	49
8	1	8	64
9	2	18	162
10	4	40	400
11	5	55	605
12	4	48	576
13	2	26	338
14	1	14	196
15	1	15	225
Total	21	231	2615

$\bar{x} = \frac{\sum xf}{n} = \frac{231}{21} = 11$

$\text{Var} = \frac{\sum x^2 f}{n} - \bar{x}^2 = \frac{2615}{21} - 11^2 = 3.523\ldots$

$s = \sqrt{3.523\ldots} \doteqdot 1.88$

3a $\bar{x} \doteqdot 7.17, s \doteqdot 3.18$

3b $\bar{x} = 5.7, s \doteqdot 1.73$

3c $\bar{x} \doteqdot 3.03, s \doteqdot 0.91$

3d $\bar{x} \doteqdot 42.88, s \doteqdot 10.53$

4a $1 + 5 + 6 + 7 + 8 + 3 + 3 + 0 + 1 = 34$

4b $\mu \doteqdot 3.26, \sigma \doteqdot 1.75$

4c

class	0–2	3–5	6–8
centre	1	4	7
freq	12	18	4

4d $\mu \doteqdot 3.29, \sigma \doteqdot 1.93$

4e Information is lost when data are grouped, causing the summary statistics to change.

5a The data written in order is
21.5, 22, 22, 23, 23.5, 23.5, 25, 27, 27, 29.5, 30, 30, 32.5, 33, 34, 35.5, 37, 39, 42.

There are 19 values so the $\frac{19+1}{2} = 10$th value will be the median. This is 29.5.

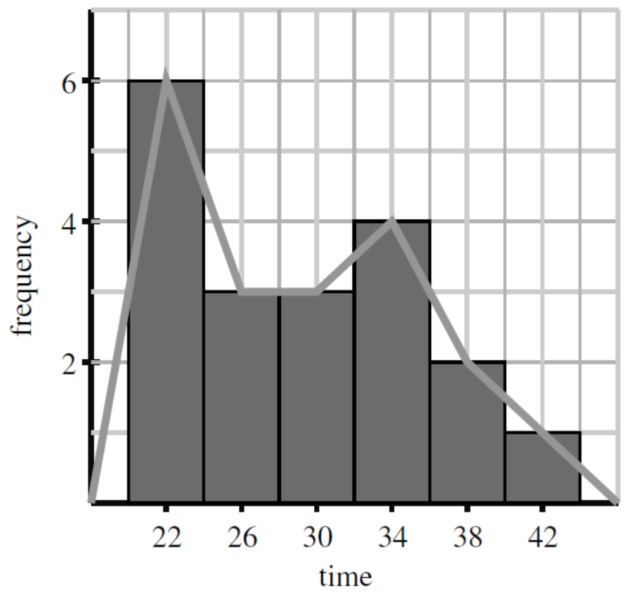
5b

class	20–24	24–28	28–32	32–36	36–40	40–44
centre	22	26	30	34	38	40
freq	6	3	3	4	2	1
c.f.	6	9	12	16	18	20

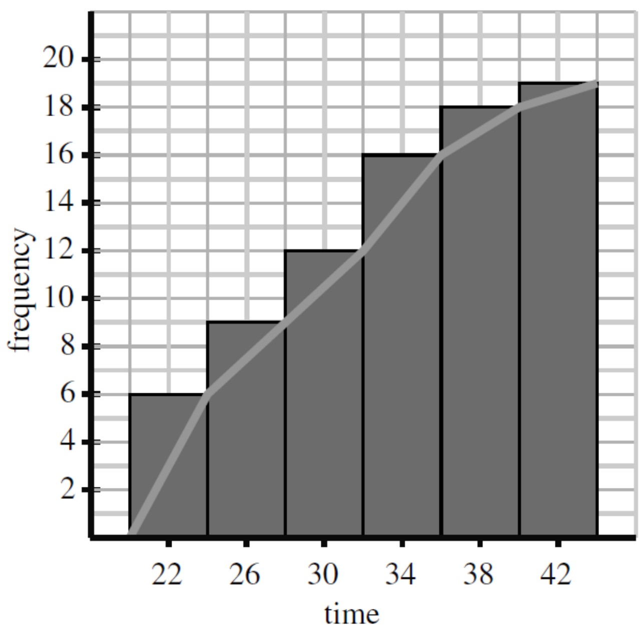
5c There are 19 values so the $\frac{19+1}{2} = 10$th value will be the median, which is 30.

No, this does not match part a, because information is lost when the data are grouped.

5d



5e



6a

x	152	154	155	157	158	159	162	163
f	1	2	1	1	2	3	2	3

x	164	165	166	168	170
f	2	2	3	1	1

6b There are 24 values so the median will be the average of the 12th and 13th value. Hence the median is $\frac{162+163}{2} = 162.5$ cm.

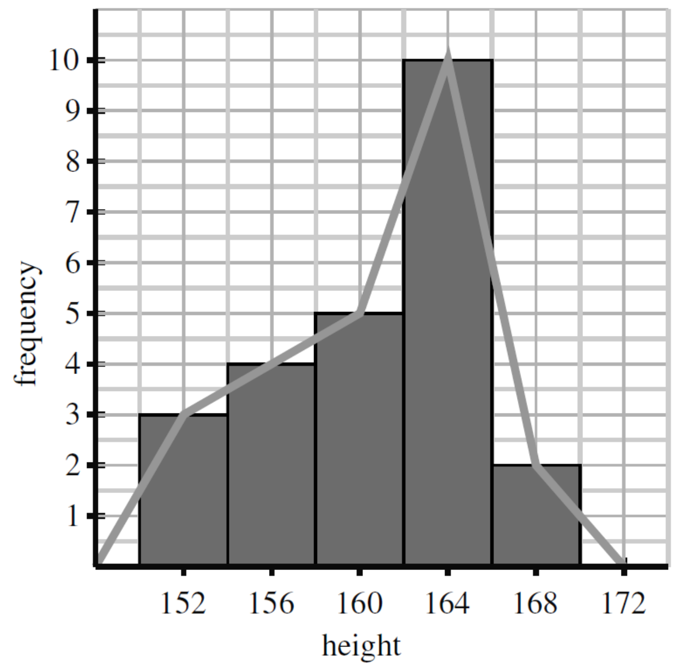
6c Trends are less clear when the data are not grouped, because it is less visually clear that the data are falling in certain zones on the domain.

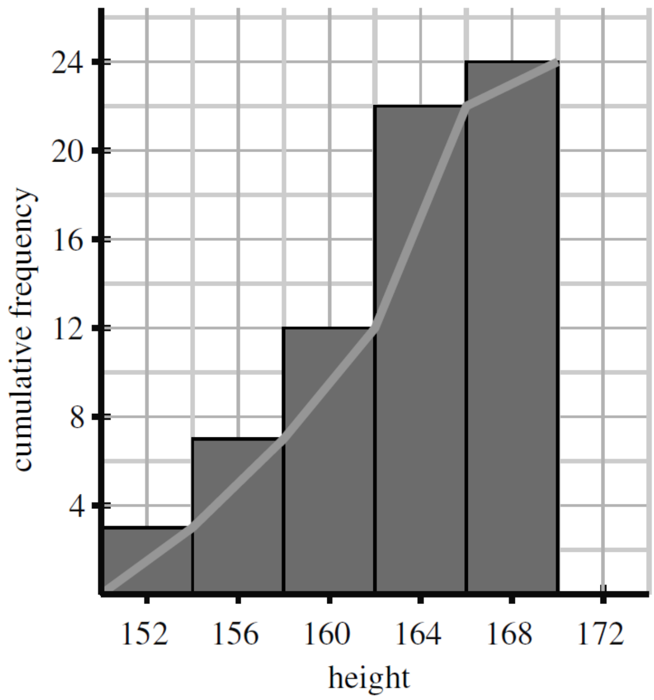
6d

group	150–154	154–158	158–162	162–166	166–170
centre	152	156	160	164	168
freq	3	4	5	10	2

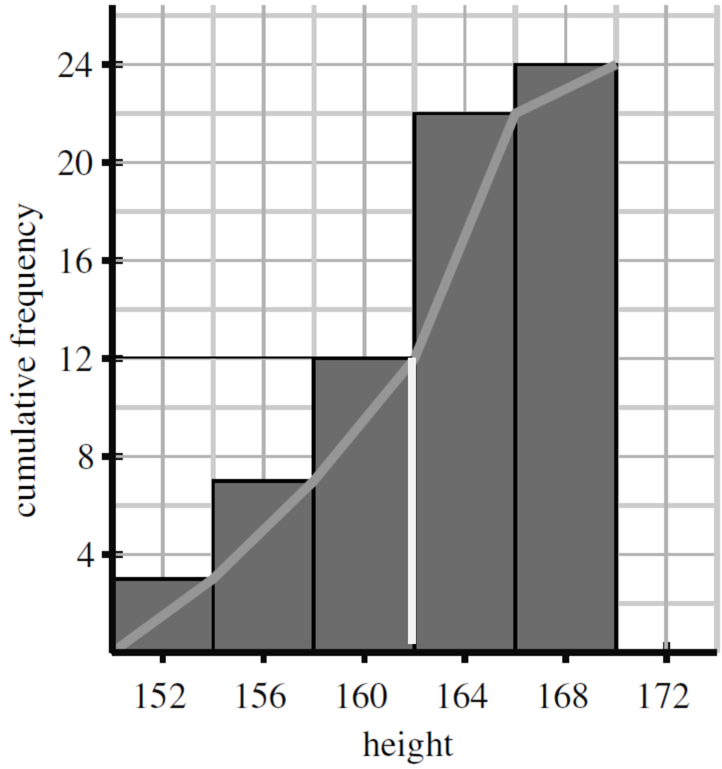
6e There are 24 values so the median will be the average of the 12th and 13th value. Hence the median is $\frac{160+164}{2} = 162$ cm

6f


frequency
10
9
8
7
6
5
4
3
2
1
152
156
160
164
168
172
height

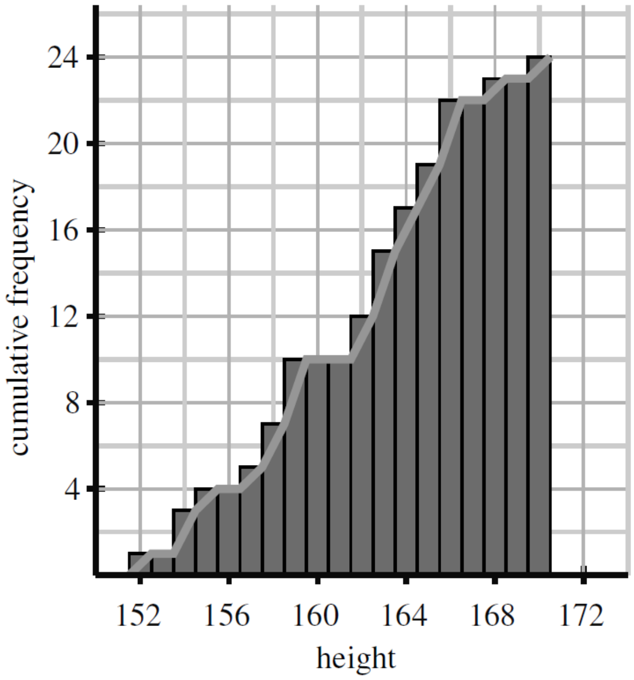

cumulative frequency
24
20
16
12
8
4
152
156
160
164
168
172
height

6g



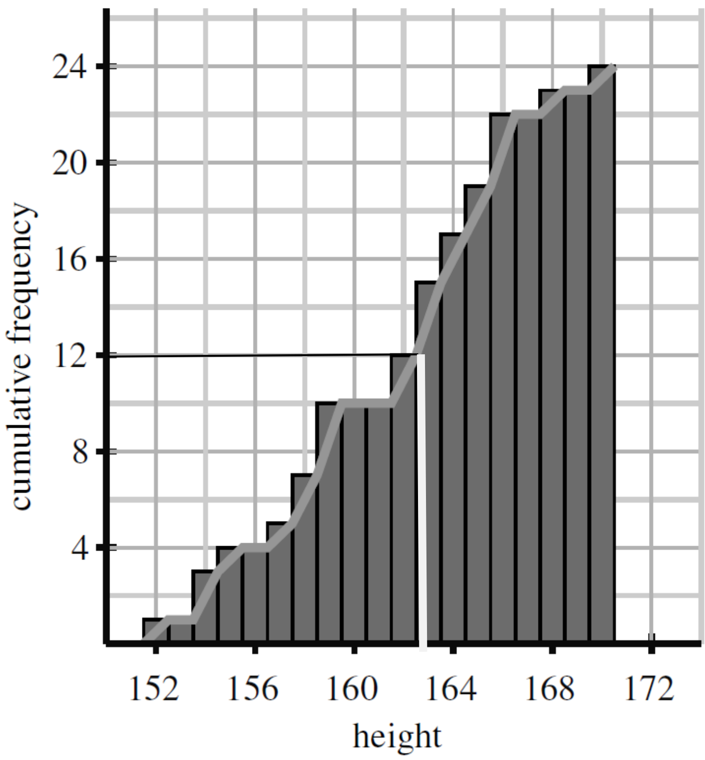
This gives the median height as 162 cm.

6h



6i The cumulative frequency polygon and ogive are much less sensitive to the grouping process than the frequency histogram and ogive. The graphs in parts g and h look very similar in shape.

6j



The line at frequency 12 meets the ungrouped data ogive at 162 cm, matching that in part g.

7a i Corrected sample standard deviation

$= 13.6 \times \frac{10}{10-1}$

$= 13.6 \times \frac{10}{9}$

$\doteqdot 15.1$

7a ii Corrected sample standard deviation

$= 13.6 \times \frac{100}{100-1}$

$= 13.6 \times \frac{100}{99}$

$\doteqdot 13.7$

7a iii Corrected sample standard deviation

$= 13.6 \times \frac{1000}{1000-1}$

$= 13.6 \times \frac{1000}{999}$

$\doteqdot 13.6$

7b Corrected sample standard deviation

$= 13.6 \times \frac{10\ 000}{10\ 000-1}$

$= 13.6 \times \frac{10\ 000}{9999}$

$= 13.601 \ldots$

Percentage change

$= \left(\frac{13.601 \ldots - 13.6}{13.6} \div 2\right) \times 100\%$

$\doteqdot 0.005\%$

Note that we divide by 2 as the distribution has two sides. We are only interested in the percentage change on the positive side of the distribution.

Solutions to Exercise 15C

1a Mean

$= \frac{4 + 8 + 5 + 2 + 9 + 12 + 8}{7}$

$\doteqdot 6.9$

Ascending order is 2, 4, 5, 8, 8, 9, 12

Median is 4th term $= 8$

Mode: 8

Range $= 12 - 2 = 10$

1b Mean

$= \frac{12 + 23 + 18 + 30 + 24 + 29 + 19 + 22 + 25 + 12}{10}$

$= 21.4$

Ascending order is 12, 12, 18, 19, 22, 23, 24, 25, 29, 30

Median is 5.5th term $= 22.5$

Mode: 12

Range $= 30 - 12 = 18$

1c Mean

$= \frac{7 + 6 + 2 + 5 + 7 + 3 + 4 + 5 + 7 + 6}{10}$

$= 5.2$

Ascending order is 2, 3, 4, 5, 5, 6, 6, 7, 7, 7

Median is 5.5th term $= 5.5$

Mode: 7

Range $= 7 - 2 = 5$

1d Mean

$$= \frac{54 + 62 + 73 + 57 + 61 + 61 + 54 + 66 + 73}{9}$$

$\doteqdot 62.3$

Ascending order is 54, 54, 57, 61, 61, 62, 66, 73, 73

Median is 5th term $= 61$

Trimodal: 54, 61, 73

Range $= 73 - 54 = 19$

2a $Q_1 = 7, Q_2 = 13, Q_3 = 17, \text{IQR} = 10$

2b $Q_1 = 12.5, Q_2 = 18.5, Q_3 = 25.5, \text{IQR} = 13$

2c $Q_1 = 7.5, Q_2 = 11, Q_3 = 18, \text{IQR} = 10.5$

2d $Q_1 = 5, Q_2 = 8.5, Q_3 = 13, \text{IQR} = 8$

2e $Q_1 = 4, Q_2 = 7, Q_3 = 13, \text{IQR} = 9$

2f $Q_1 = 10, Q_2 = 15, Q_3 = 21, \text{IQR} = 11$

2g $Q_1 = 5, Q_2 = 9, Q_3 = 13.5, \text{IQR} = 8.5$

2h $Q_1 = 12, Q_2 = 14, Q_3 = 18, \text{IQR} = 6$

3a $Q_1 = 4, Q_2 = 12, Q_3 = 16, \text{IQR} = 12$

3b $Q_1 = 1, Q_2 = 6.5, Q_3 = 11, \text{IQR} = 10$

3c $Q_1 = 7, Q_2 = 9, Q_3 = 12, \text{IQR} = 5$

3d $Q_1 = 2.5, Q_2 = 5, Q_3 = 7, \text{IQR} = 4.5$

3e $Q_1 = 7, Q_2 = 7, Q_3 = 10, \text{IQR} = 3$

3f $Q_1 = 4, Q_2 = 5, Q_3 = 9, \text{IQR} = 5$

3g $Q_1 = 2.5, Q_2 = 4, Q_3 = 9.5, \text{IQR} = 7$

3h $Q_1 = 4.5, Q_2 = 9, Q_3 = 12, \text{IQR} = 7.5$

4a Answers may differ here, but 40 and 92 are likely.

4b 40, 54, 59, 69, 92

4c $\text{IQR} = 15$, $Q_1 - 1.5 \times \text{IQR} = 31.5$ and $Q_3 + 1.5 \times \text{IQR} = 91.5$.

Thus 92 is the only outlier by the IQR criterion.

4d Some may identify 40 as an outlier by eye — this shows the advantage of plotting values, where it becomes evident that this score is well separated from other scores. A student receiving 40 in this cohort should be noted as someone needing extra attention and assistance.

4e i 54, 60, 70.5, $\text{IQR} = 16.5$

4e ii 53.5, 58, 68.5, $\text{IQR} = 15$

4e iii 54, 59, 68.5, $\text{IQR} = 14.5$

4f In this case, with a reasonably sized dataset, the middle of the data is fairly stable and removing an extreme value has only a small effect on the quartiles and IQR. With a large dataset and tightly clustered values in the middle two quarters of the data, the difference would be even smaller.

4g i 60.8, 11.1

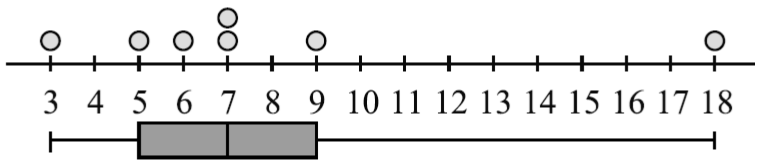
4g ii 61.6, 10.5

4g iii 59.5, 9.4

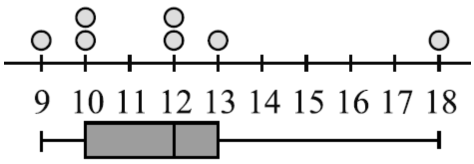
4g iv 60.3, 8.7

4h 2.4 is 22% of 11.1. Any deviation from the mean is exaggerated by the standard deviation because the deviation from the mean is squared when calculating the variance.

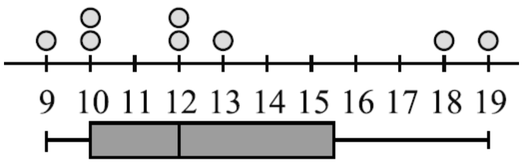
5a i IQR = 4, outlier 18



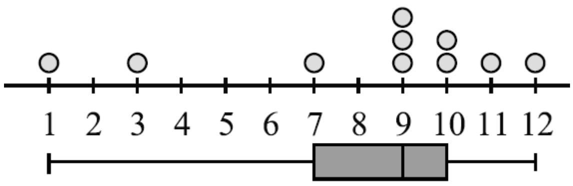
5a ii IQR = 3, outlier 18



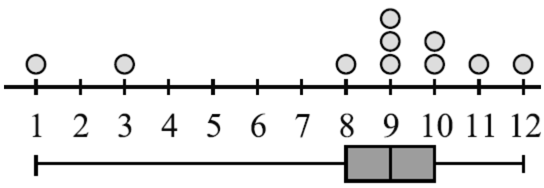
5a iii IQR = 5.5, no outliers



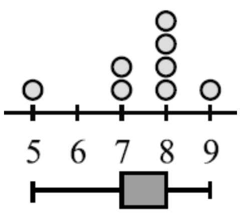
5a iv IQR = 3, outlier 1



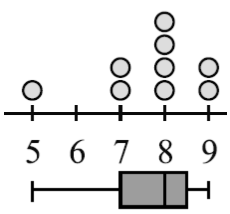
5a v IQR = 2, outliers 1, 3



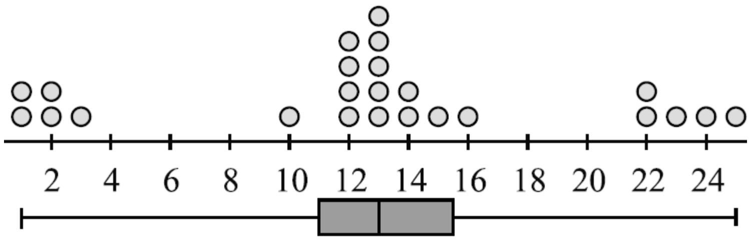
5a vi IQR = 1, outlier 5



5a vii IQR = 1.5, no outliers



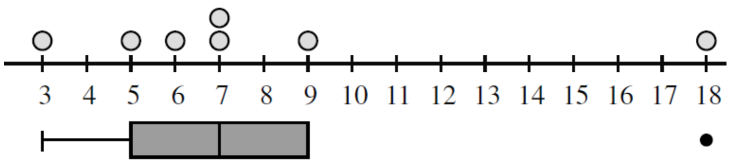
5a viii IQR = 4.5, outliers 1, 1, 2, 2, 3, 23, 24, 25



5b It must be noted that some of the pathologies in these examples come about because of the small datasets. Statistics is always more accurate and reliable with a large dataset.

Generally the definition picks up the values that appear extreme on the dot plots. Notably (in these small datasets), it picks up single extreme values — if more values are a long way from the mean, they may not be marked as outliers. Datasets with a small IQR may need a closer inspection — in parts vi and vii, the value at 5 is not so extreme and the datasets are not so different, yet in one case it is marked as an outlier, but in the other it is not. The final dataset has a very tight subset of data between the Q_1 and Q_3, giving a small interquartile range. This definition of outliers gives 8 values in 24 (one third of the data) as outliers. Furthermore, 23–25 are outliers, but 22 is not. The issue here is the unusual shape of the distribution. Rules such as this IQR criterion for outliers should be an invitation to inspect the values that have been flagged more closely, rather than following a rule blindly.

5c



6a

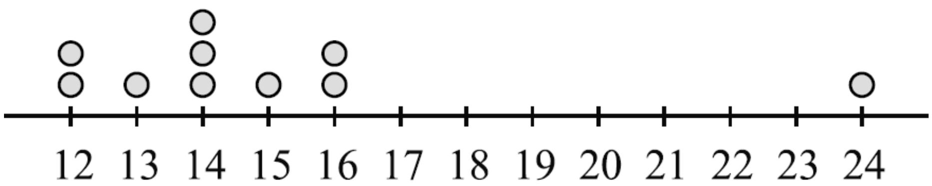
x	f	xf	x^2f
12	2	24	288
13	1	13	169
14	3	42	588
15	1	15	225
16	2	32	512
24	1	24	576
Total	10	150	2358

$$\bar{x} = \frac{\sum xf}{n} = \frac{150}{10} = 15$$

$$\text{Var} = \frac{\sum x^2 f}{n} - \bar{x}^2 = \frac{2358}{10} - 15^2 = 10.8$$

$$s = \sqrt{10.8} \doteqdot 3.29$$

6b The value 24 appears to be an outlier.



6c IQR $= 3$ and $Q_3 = 16$.

Because $24 > 16 + 1.5 \times 3$, this definition also labels 24 an outlier.

6d

x	f	xf	x^2f
12	2	24	288
13	1	13	169
14	3	42	588
15	1	15	225
16	2	32	512
Total	9	126	1782

$$\bar{x} = \frac{\sum xf}{n} = \frac{126}{9} = 14$$

$$\text{Var} = \frac{\sum x^2 f}{n} - \bar{x}^2 = \frac{1782}{9} - 14^2 = 2$$

$$s = \sqrt{2} \doteqdot 1.41$$

6e This does not have much effect on the mean, but it has a big percentage effect on the standard deviation — removing the outlier more than halves the standard deviation. The operation of squaring $(x - \bar{x})$ means that values well separated from the mean have an exaggerated effect on the size of the variance.

6f No effect at all. (The median is 14 and IQR is 3 with and without 24 in the dataset.)

6g If there are significant outliers, or at least values spread far from the mean, this can have a big influence on the IQR. The IQR is a good measure if you are more interested in the spread of the central 50% of the data.

7a Emily got less than 62.

7b Around 50% (and no more than 50%)

7c The mathematics results were more spread out, and the centre of the data (by median) was 5 marks higher. The interquartile range of both distributions, however, was the same. Clearly the mathematics cohort has some students who perform much more strongly, and others who perform much weaker, than the majority of their peers.

7d Xavier was placed in the upper half of the English cohort, but in the lower half of the mathematics cohort. The English result was thus more impressive.

7e i Outlier is 45.

7e ii The bottom 25% of English scores show a spread of 6 marks (51–57). The bottom 25% of mathematics scores show a spread of 8 marks ($53-61$). The spread of the lower half is now much more comparable.

8a The results are not paired. Just because Genjo received the lowest score in the writing task does not mean that he received the lowest score in the speaking task. Thus we cannot answer the question, although we might make conjectures, given that Genjo is obviously struggling significantly with English.

8b i mean 66.1, median 68, range 56

8b ii $\text{IQR} = 73 - 60 = 13$

$73 + 1.5 \times 13 = 92.5$

$60 - 1.5 \times 13 = 40.5$

91 and 35 are outliers.

8c i mean 64.4, median 65.5, range 56

8c ii IQR $= 71 - 57.5 = 13.5$

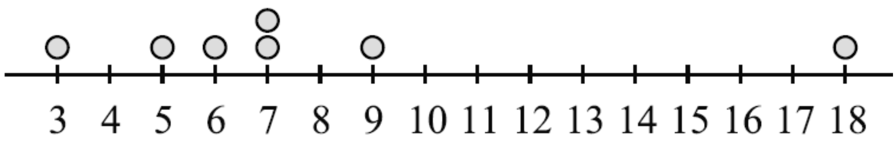
$71 + 1.5 \times 13.5 = 91.25$

$57.5 - 1.5 \times 13.5 = 37.25$

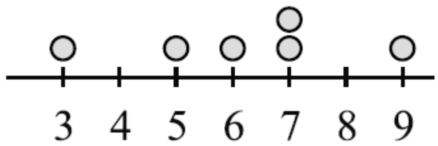
37 and 93 are outliers.

8d It is difficult to say. Students have found the second task more challenging, evidenced by the lower mean and median. This could be due to the construction of the task, or simply because it is a type of task that some students find more difficult.

9a Question 5a i contained one outlier. The dot plot for the original data is

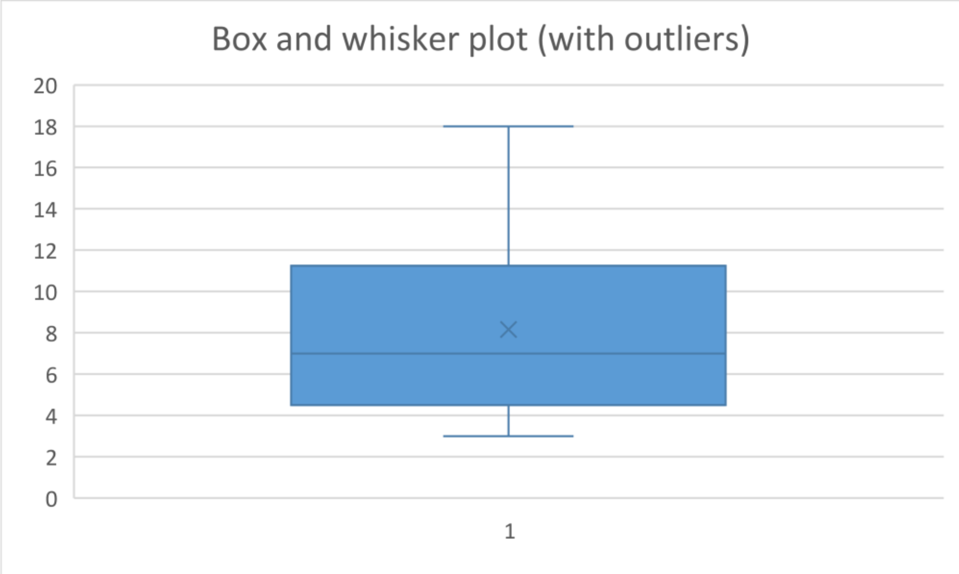


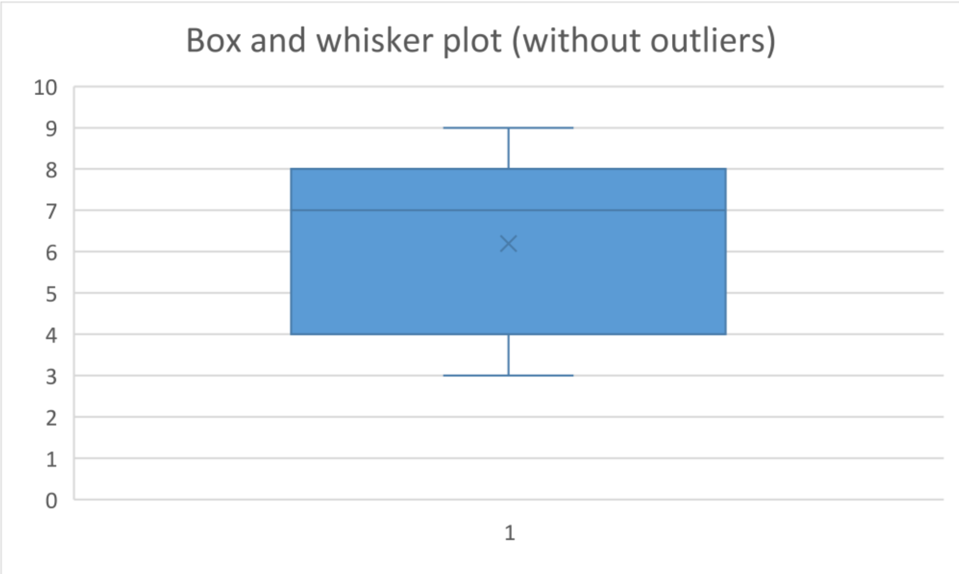
Whilst the dot plot without the outlier is given by



(You should try plotting these yourself in software like Desmos if you have not done so already.)

Here you can see that the plot without outliers is a lot thinner and all of the datapoints are significantly more condensed than in the graph excluding outliers.





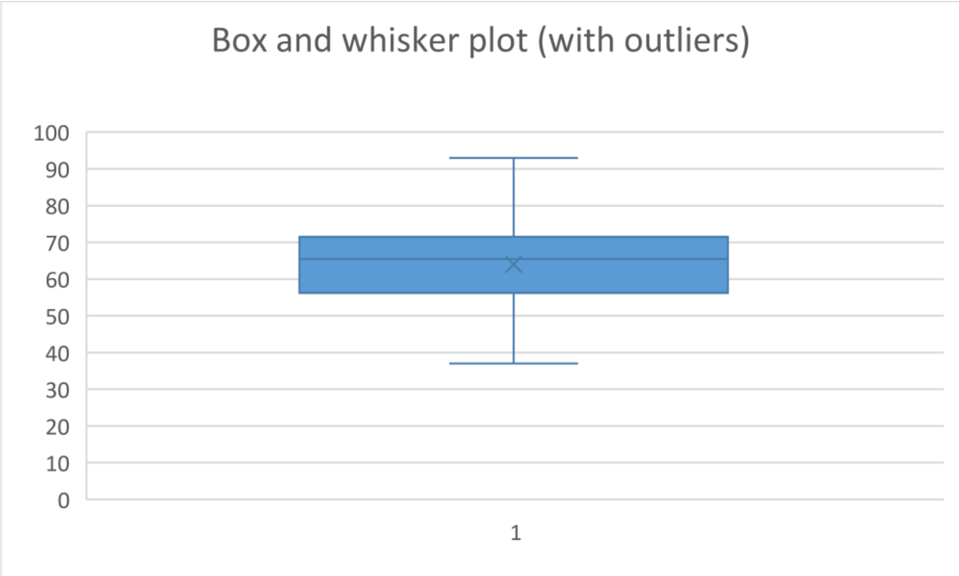
(You should try recreating these using Excel if you have not done so already.)

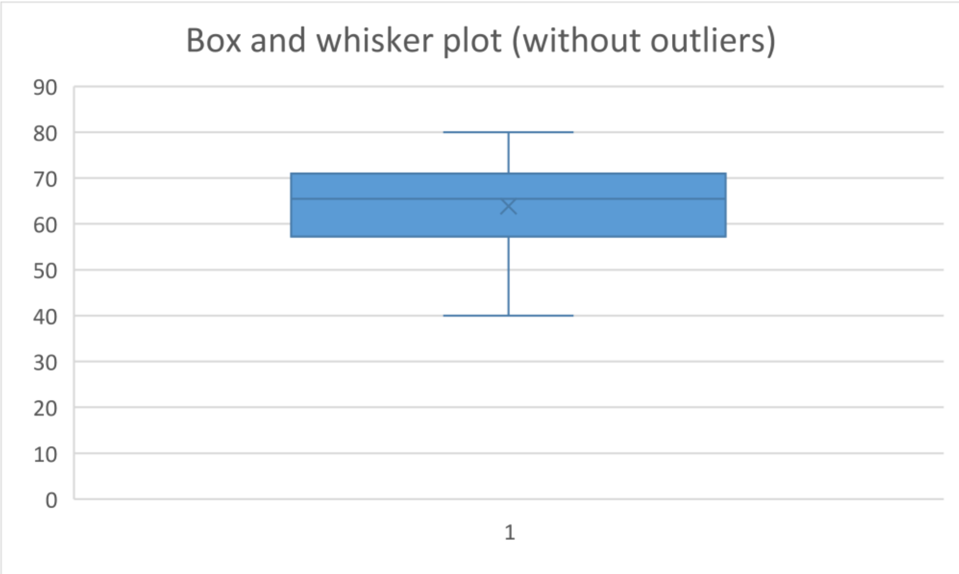
Note that the box and whisker plots have different endpoints (the one without outliers has a smaller range). Also note that whilst the median has remained at 7 (this is because coincidentally there were two datapoints with the value 7 in this dataset), Q_1 has changed from 4.5 to 4 and Q_3 has changed from 11 to 8. This in turn means that the IQR has gone from 6.5 (with the outlier) to 4 (without the outlier). This is a significant change.

9b An example of a question that contained two outliers is question 8.

93 was an outlier on the upper end of the scale and 37 was an outlier on the lower end of the scale. We shall look at the statistics for the speaking task in this question.

The box plots are as follows:



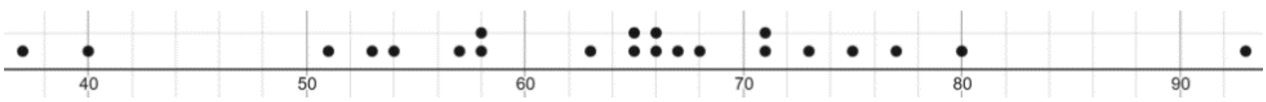


(You should try recreating these using Excel if you have not done so already.)

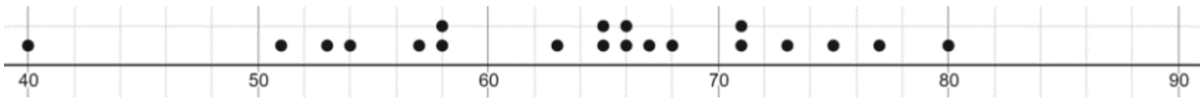
Note that the box and whisker plots have different endpoints (the one without outliers has a smaller range) but the median, Q_1 and Q_3 are unchanged.

Now the dot plots are:

With outliers



Without outliers



(You should try plotting these yourself in software like Desmos if you have not done so already.)

Here you can see that the plot without outliers is a lot thinner and all of the datapoints are significantly more condensed than in the graph excluding outliers. Hence, having two or more outliers can mask one another's existence from an IQR test if they are at both ends of the dataset.

9c Q5a i: $\sigma = 4.48, \bar{x} = 7.86$.

Hence the outlier 18 is $\frac{18-87.86}{4.48} = 2.26$ standard deviations from the mean.

Q5a ii: $\sigma = 2.78, \bar{x} = 12$.

Hence the outlier 18 is $\frac{18-12}{2.78} = 2.16$ standard deviations from the mean.

Q5a iv: $\sigma = 3.33, \bar{x} = 8.1$.

Hence the outlier 1 is $\frac{1-8.1}{3.33} = -2.13$ standard deviations from the mean.

Q5a v: $\sigma = 3.31, \bar{x} = 8.2$.

Hence the outlier 1 is $\frac{1-8.2}{3.31} = -2.17$ standard deviations from the mean and the outlier 3 is $\frac{3-8.2}{3.31} = -1.57$ standard deviations from the mean.

Q5a vi: $\sigma = 1.12, \bar{x} = 7.5$.

Hence the outlier 5 is $\frac{5-7.4}{1.12} = -2.23$ standard deviations from the mean.

Q5a viii: $\sigma = 7.02, \bar{x} = 12.8$.

Investigating just two of the six outliers:
The outlier 1 is $\frac{1-12.8}{7.02} = -1.68$ standard deviations from the mean.

The outlier 25 is $\frac{25-12.8}{7.02} = 1.73$ standard deviations from the mean.

This would indicate that any outlier more than ± 1.5 standard deviations from the mean should be considered an outlier.

Solutions to Exercise 15D

1a i height

1a ii weight

1b i radius

1b ii area. It is natural to think that the area of the circle is determined by the radius chosen when it is drawn, but mathematically we could write $r = \sqrt{\frac{A}{\pi}}$, reversing the natural relationship.

1c i weight

1c ii price. Note that the price may change when meat is bought in bulk, so there is a deeper relationship between these two quantities than simply
price $=$ weight $\times$ cost per kg.

1d i world rank

1d ii placing

1e i temperature

1e ii power consumption. Power consumption increases with the use of air conditioners (higher temperatures) or heaters (colder weather).

1f It is natural to take x as the independent variable and y as the dependent variable. Note in this case the relationship cannot naturally be reversed, because there are multiple x-values resulting from the same y-value.

2a strong positive

2b virtually none

2c strong negative

2d strong negative

2e moderate positive

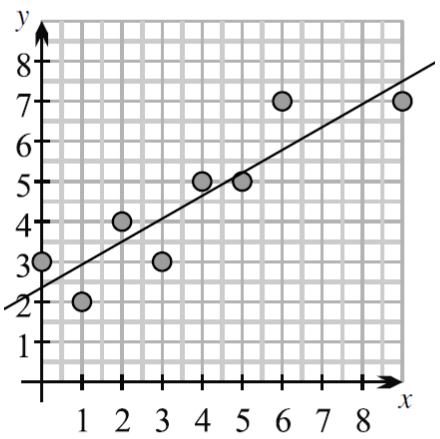
2f weak positive

2g strong negative

2h strong positive

2i moderate negative

3a



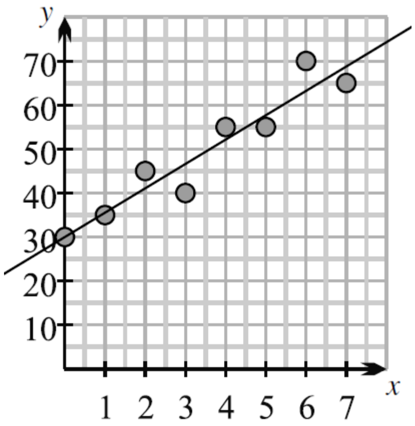
Line of best fit:

gradient $= \frac{0.5}{1} = 0.5$

y-intercept $= 2.5$

$y = 0.5x + 2.5$

3b



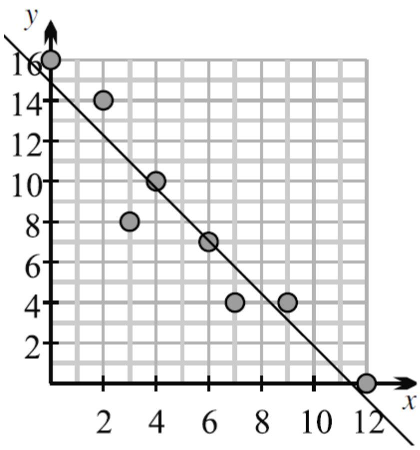
Line of best fit:

gradient $= \frac{5}{1} = 5$

y-intercept $= 30$

$y = 5x + 30$

3c



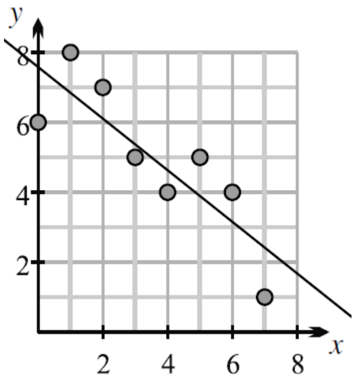
Line of best fit:

gradient $= -\frac{3}{2}$

y-intercept $= 15$

$y = -\frac{3}{2}x + 15$

3d



Line of best fit:

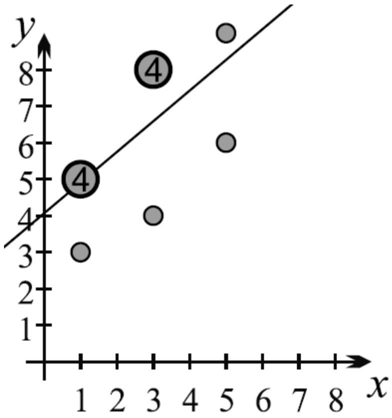
gradient $= -\frac{1.5}{2} = -\frac{3}{4}$

y-intercept $= 7$

$y = -\frac{3}{4}x + 7$

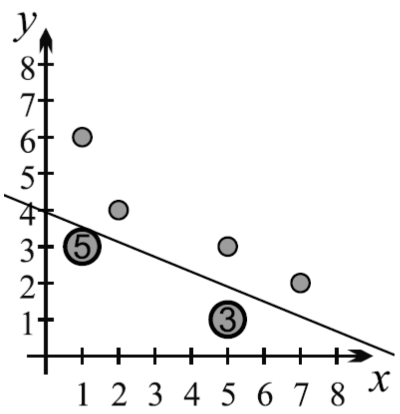
4a i Strong positive correlation

4a ii



4b i Strong negative correlation

4b ii



5a A quadratic relationship (a parabola)

5b A square root

5c A hyperbola

5d A circle

5e An exponential

5f No obvious relationship

6a i 6 L

6a ii 10 L

6b $V = 2t$

6c The V-intercept is zero. In no minutes, zero water will flow through the pipe.

6d This is the flow rate of the water, 2 L/s.

6e Negative time makes little sense here, because he cannot measure the volume of water that flowed for say -3 minutes.

6f Experimental error could certainly be a factor, but it may simply be that the flow rate of water is not constant. It may vary due to factors in, for example, the pumping system.

6g $V = 2t$

$= 2 \times 30$

$= 60$ L

The extrapolation seems reasonable provided that the half-hour chosen is at about the same time of day that he performed his experiment.

6h $V = 2t$

$t = \dfrac{V}{2}$

$= \dfrac{45}{2}$

$= 22.5$ minutes

6i Yasuf's experiments were all carried out in a period of several hours during the day. It may be that the flow rate changes at certain times of the day, for example, at peak demands water pressure may be lower and the flow rate may decrease. The flow rate may also be different at night — for example, the water pump may only operate during the day. More information and experimentation is required.

7a 1000 (from the graph)

7b i gradient

$= \frac{\text{rise}}{\text{run}}$

$= \frac{14 - 5}{10 - 0}$

$= \frac{9}{10}$

$= 0.9$

y-intercept $= 5$

$P = 0.9t + 5$

7b ii It looks fairly good for the seven year period.

7b iii Using $P = 0.9t + 5$,

$P = 0.9(9) + 5 = 13.1$

So, predicted population is 1310.

From the graph, $P = 15.4$

So, actual population is 1540.

Hence the error was $1540 - 1310 = 230$ people.

7c i The new model predicts $P = 5 \times 2^{0.19\times 9} = 16.4$, that is, 1640 people, so it is certainly much better.

7c ii Population is growing very strongly in Hammonsville. Investigators should be looking into the cause of the growth, which may change over the next few years. For example, it may be due to a short-term mining boom. Eventually there may be other constraining factors, such as available land for housing.

7d Extrapolation can be dangerous. Provided, however, that the independent variable is constrained to a small enough interval, linear predictions may well have validity. This is the idea behind calculus, where curves are approximated locally by a tangent.

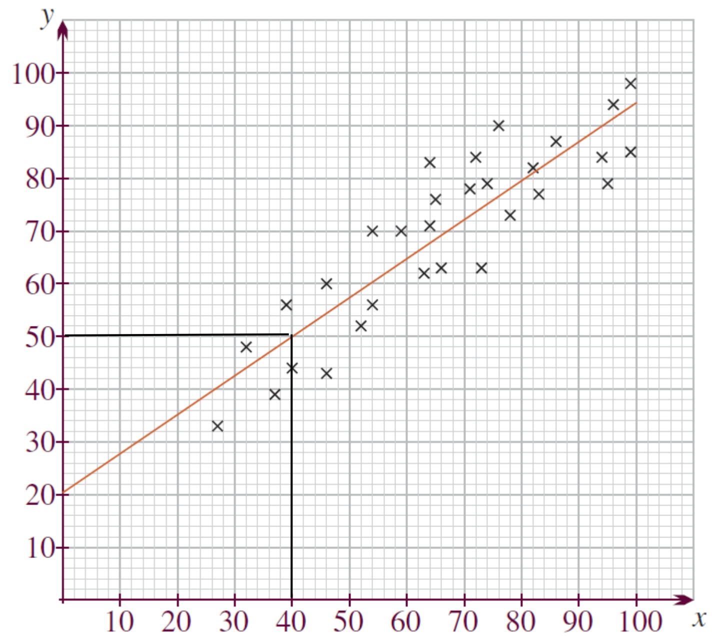
8a 99 in assessment 1, 98 in assessment 2. They were obtained by the same student, but another student also got 99 in assessment 1.

8b 27 in assessment 1, 33 in assessment 2. They were the same student.

8c Students getting below about 77 marks in assessment 1 do better in assessment 2, students above 77 marks in assessment 1 get a lower mark in assessment 2, according to the line of best fit. Perhaps the second assessment started easier, but was harder at the end.

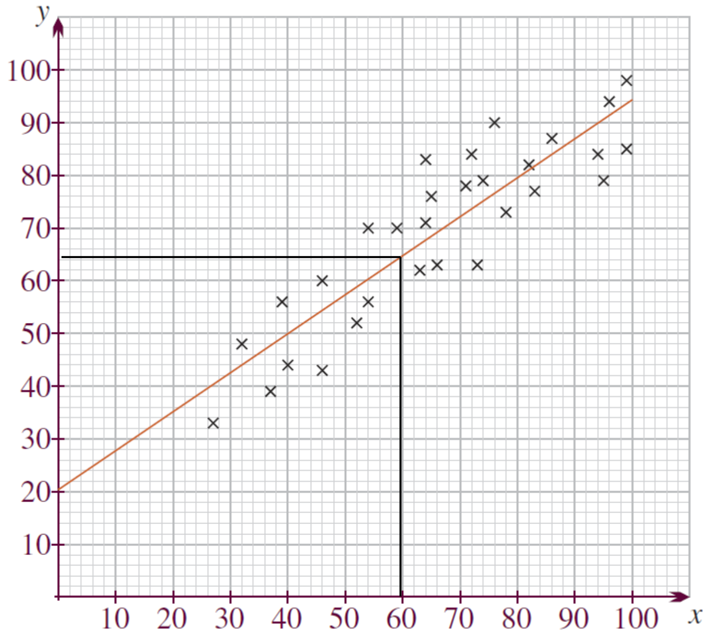
8d i 50

This is obtained from the graph as shown.



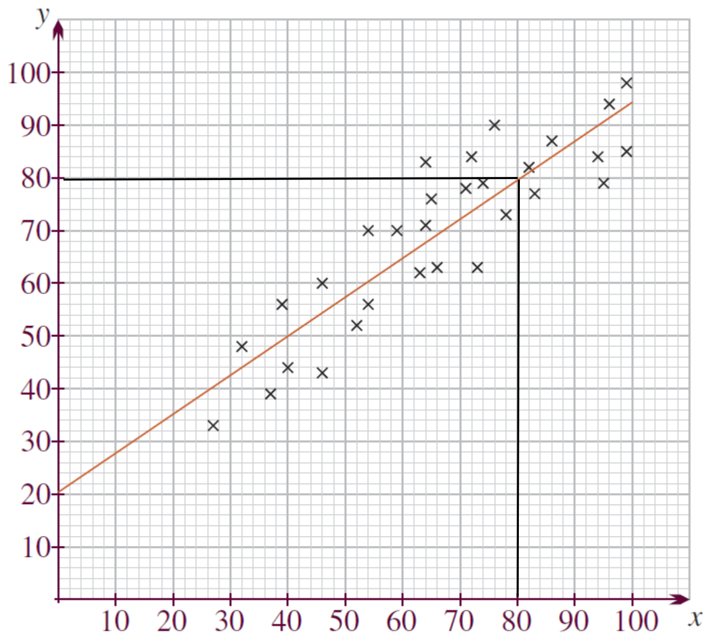
8d ii 65

This is obtained from the graph as shown.



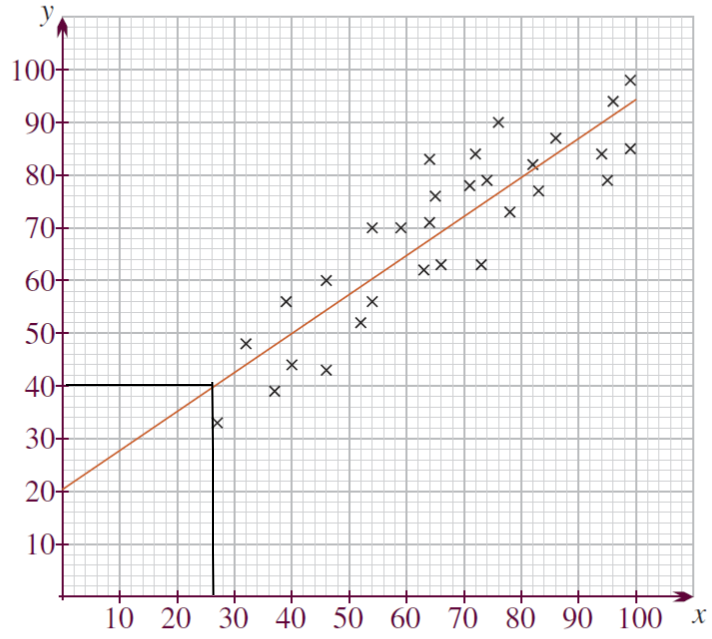
8d iii 80

This is obtained from the graph as shown.



8d iv 26

This is obtained from the graph as shown.



8d v A negative score! Clearly the model breaks down for small scores.

8e Using the points (0,20) and (100,94):

gradient

$= \frac{\text{rise}}{\text{run}}$

$= \frac{94 - 20}{100 - 0}$

$= \frac{74}{100}$

$= 0.74$

y-intercept $= 20$

$y = 0.74x + 20$

8f A more accurate method would incorporate data from more than one assessment task in estimating their missing score. This is a question better tackled using standard deviation and the techniques of the next chapter.

9a By observation of the graph, the maximum vertical difference between a plotted point and the line of best fit occurs when $L = 0.8$ m.

$T^2 = 4 - 3.2018 \doteqdot 0.8$ s²

9b It could be experimental error. For example, the string could have been twisted or released poorly, the experiment could have been incorrectly timed, or there could have been a recording error.

9c They may have measured 10 periods and then divided by 10 before recording the length of one period. Errors could then arise if the motion was damped, that is, if the pendulum slowed down significantly over a short time period.

9d The line of best fit in this model has $T^2 = 4.01L$.

Theory would predict a value of

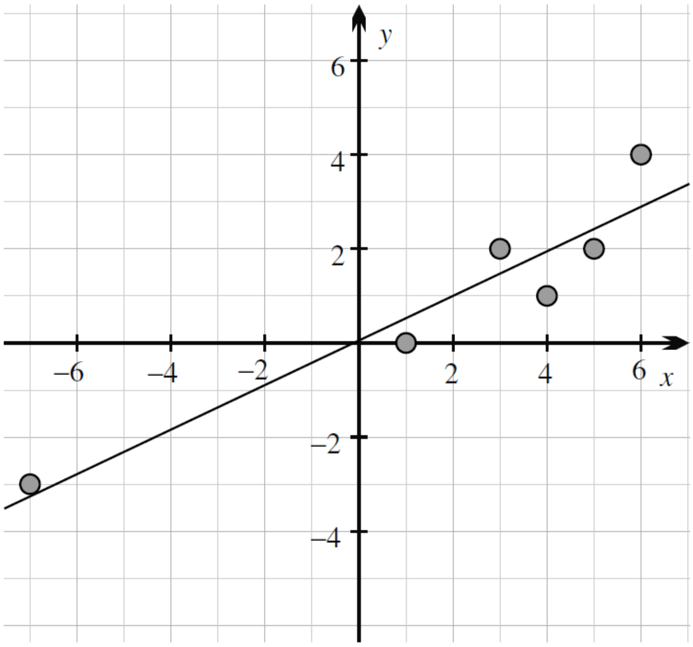
$$\begin{aligned} T^2 &= \frac{4\pi^2 L}{g} \\ &= \frac{4\pi^2}{9.8} L \\ &\doteqdot 4.03L \end{aligned}$$

These results are in pretty good agreement with the theory.

Solutions to Exercise 15E

1a There appears to be a fairly strong correlation, though note the small dataset.

1b



1c

							Sum
x	-7	1	3	4	5	6	12
y	-3	0	2	1	2	4	6
$x-\bar{x}$	-9	-1	1	2	3	4	0
$y-\bar{y}$	-4	-1	1	0	1	3	0
$(x-\bar{x})^2$	81	1	1	4	9	16	112
$(y-\bar{y})^2$	16	1	1	0	1	9	28
$(x-\bar{x})(y-\bar{y})$	36	1	1	0	3	12	53

1d $\bar{x} = 12 \div 6 = 2$

$\bar{y} = 6 \div 6 = 1$

$(\bar{x}, \bar{y}) = (2, 1)$

1e See above

1f

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2(y-\bar{y})^2}}$$

$$= \frac{53}{\sqrt{112 \times 28}}$$

$$\doteqdot 0.95$$

1g r is close to ± 1, so it is a good fit.

1h

$$m = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$$

$$= \frac{53}{112}$$

$$\doteqdot 0.47$$

1i $b = \bar{y} - m\bar{x}$

$\doteqdot 1 - 0.47 \times 2$

$= 0.06$

So y-intercept is 0.06.

1j Using $m \doteqdot 0.5$ (to one decimal place) and

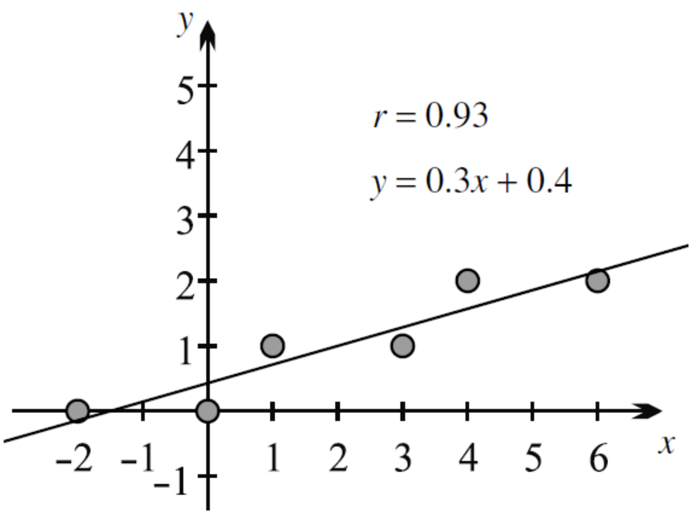
$b \doteqdot 1 - 0.5 \times 2 = 0$,

equation of line of best fit is

$$y = \frac{1}{2}x + 0$$

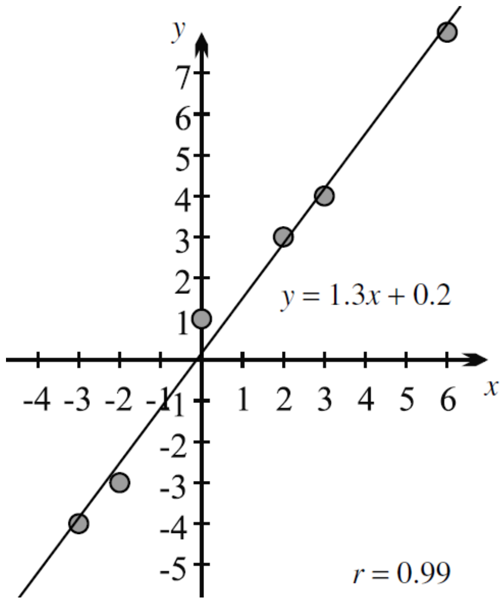
2a

							SUM
x	-2	0	1	3	4	6	12
y	0	0	1	1	2	2	6
$x-\bar{x}$	-4	-2	-1	1	2	4	0
$y-\bar{y}$	-1	-1	0	0	1	1	0
$(x-\bar{x})^2$	16	4	1	1	4	16	42
$(y-\bar{y})^2$	1	1	0	0	1	1	4
$(x-\bar{x})(y-\bar{y})$	4	2	0	0	2	4	12



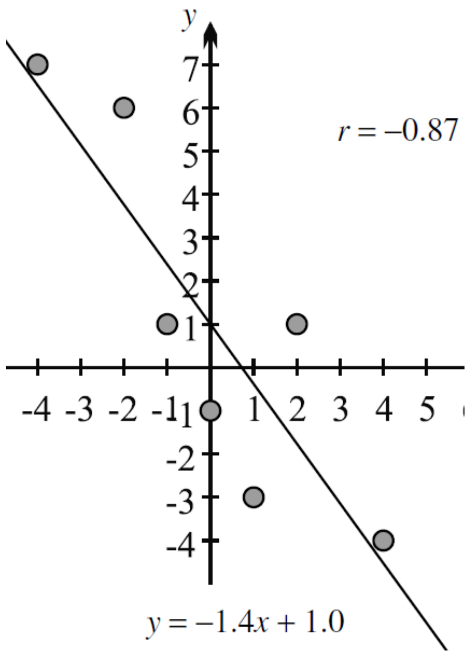
2b

							SUM
x	-3	-2	0	2	3	6	6
y	-4	-3	1	3	4	8	9
$x-\bar{x}$	-4	-3	-1	1	2	5	0
$y-\bar{y}$	-5.5	-4.5	-0.5	1.5	2.5	6.5	0
$(x-\bar{x})^2$	16	9	1	1	4	25	56
$(y-\bar{y})^2$	30.25	20.25	0.25	2.25	6.25	42.25	101.5
$(x-\bar{x})(y-\bar{y})$	22	13.5	0.5	1.5	5	32.5	75



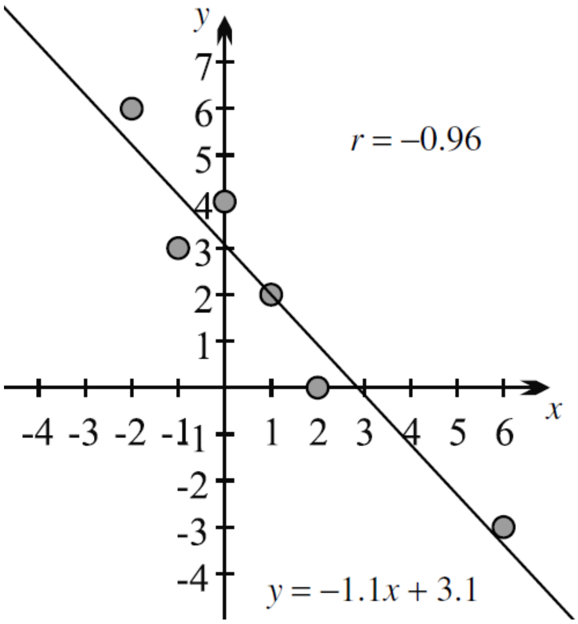
2c

								SUM
x	-4	-2	-1	0	1	2	4	0
y	7	6	1	-1	-3	1	-4	7
$x-\bar{x}$	-4	-2	-1	0	1	2	4	0
$y-\bar{y}$	6	5	0	-2	-4	0	-5	0
$(x-\bar{x})^2$	16	4	1	0	1	4	16	42
$(y-\bar{y})^2$	36	25	0	4	16	0	25	106
$(x-\bar{x})(y-\bar{y})$	-24	-10	0	0	-4	0	-20	-58



2d

								SUM
x	-2	-1	0	1	2	6	6	-2
y	6	3	4	2	0	-3	12	6
$x-\bar{x}$	-3	-2	-1	0	1	5	0	-3
$y-\bar{y}$	4	1	2	0	-2	-5	0	4
$(x-\bar{x})^2$	9	4	1	0	1	25	40	9
$(y-\bar{y})^2$	16	1	4	0	4	25	50	16
$(x-\bar{x})(y-\bar{y})$	-12	-2	-2	0	-2	-25	-43	-12


y
7
6
5
4
3
2
1
-4 -3 -2 -1
1 2 3 4 5 6
x
-1
-2
-3
-4
r = −0.96
y = −1.1x + 3.1

Solutions to Exercise 15F

1a $r = 0.96, y = 0.95x + 0.44$

1b $r = 0.79, y = 0.45x + 2.58$

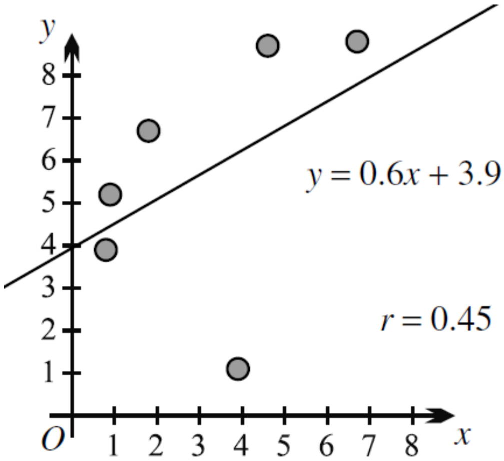
1c $r = -0.86, y = -1.05x + 8.75$

1d $r = -0.53, y = -0.41x + 4.70$

1e $r = 0.96, y = 1.38x + 0.75$

2a i $r = 0.45, y = 0.58x + 3.94$

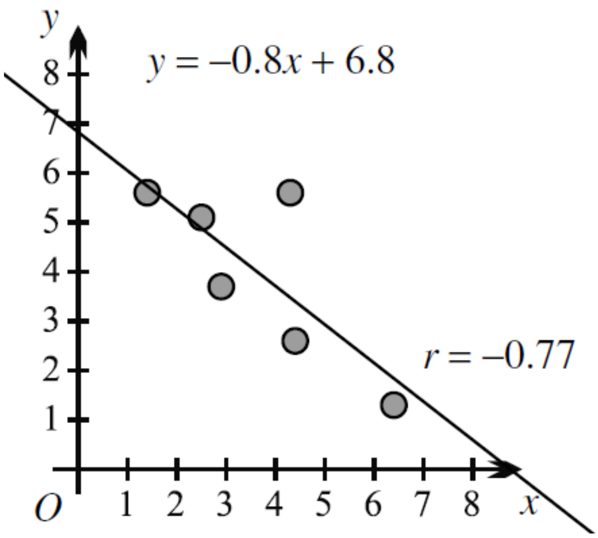
2a ii



2a iii If the outlier at $(3.9, 1.1)$ is removed, then $r = 0.91, y = 0.75x + 4.43$.

2b i $r = -0.77, y = -0.78x + 6.83$

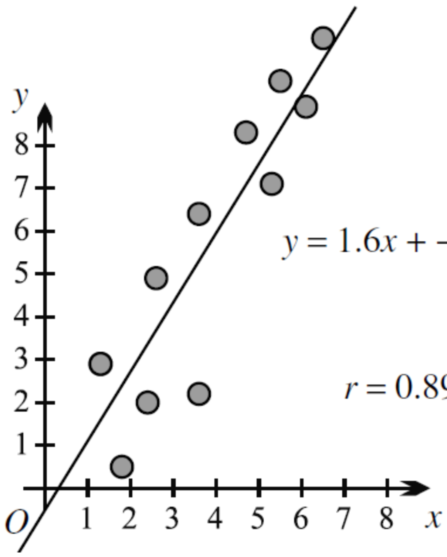
2b ii



2b iii If the outlier at $(4.3, 5.6)$ is removed, then $r = -0.97, y = -0.89x + 6.79$.

2c i $r = 0.89, y = 1.62x - 0.51$

2c ii



2c iii If the outlier at $(3.6, 2.2)$ is removed, then $r = 0.93, y = 1.61x - 0.19$.

3 Because the dataset was larger, the effect of the single outlier was mitigated by the other data points.

4a i Dataset 1: $y = 1x + 1.4, r = 0.86$, Dataset 2: $y = 0.7x + 3.0, r = 0.76$

4a ii Dataset 1: $y = 0.8x + 1.9, r = 0.79$, Dataset 2: $y = 0.7x + 2.5, r = 0.82$

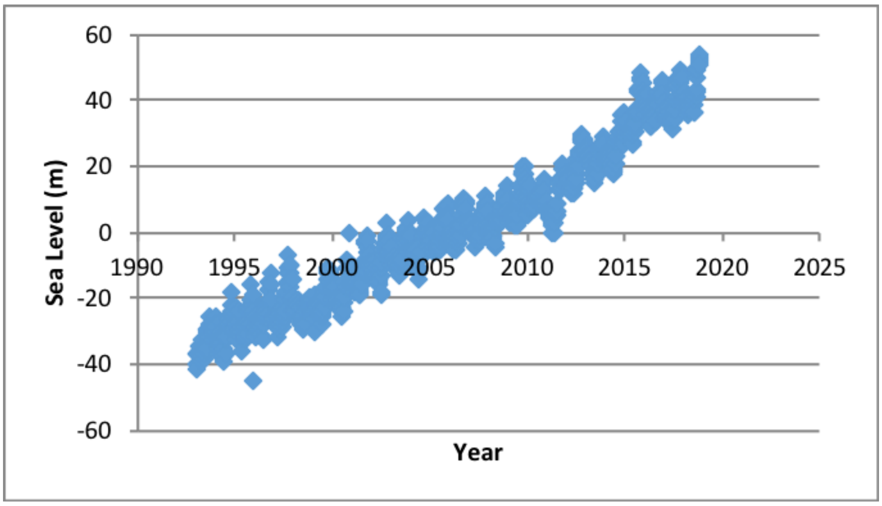
4b In all cases the correlation is strong. In part a, the repeated point has strengthened the correlation, but in the second example it has weakened it. Note that a strong correlation doesn't indicate that the data are correct. In part a, for example, leaving out 4 of the 9 points still gave a strong correlation, but a very different equation of line of best fit.

4c The effect is less in the larger dataset, as expected. The gradient is unchanged (correct to one decimal place) and the y-intercept only differs by 20%, rather than by 26%. In a larger (more realistically sized) dataset, the effect would likely be less again. The effect of the repeated point will also depend on its place on the graph (central versus on the extremes of the data) and how close it is to the line of best fit.

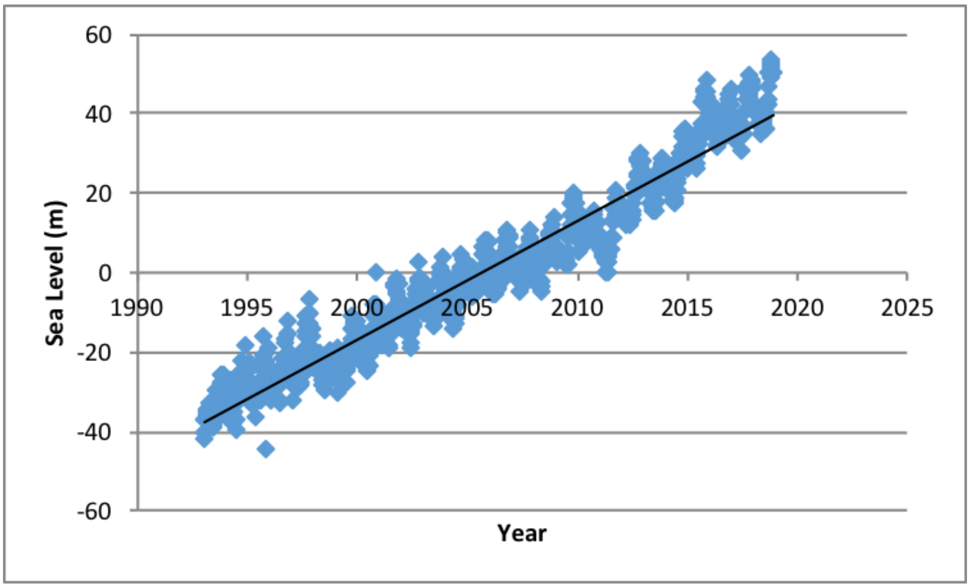
5-7 Answers will vary

8a, b Refer to the Excel spreadsheet provided in the digital textbook.

8c



8d Select the graph and then go into 'chart tools/analysis/trend line' and select the linear trend line which gives you:



8e To find its R^2 value go to 'more trend line options' and select 'display R^2' this will give you the equation $R^2 = 0.942$. This is a good fit due to the high value of R^2 (it is close to 1).

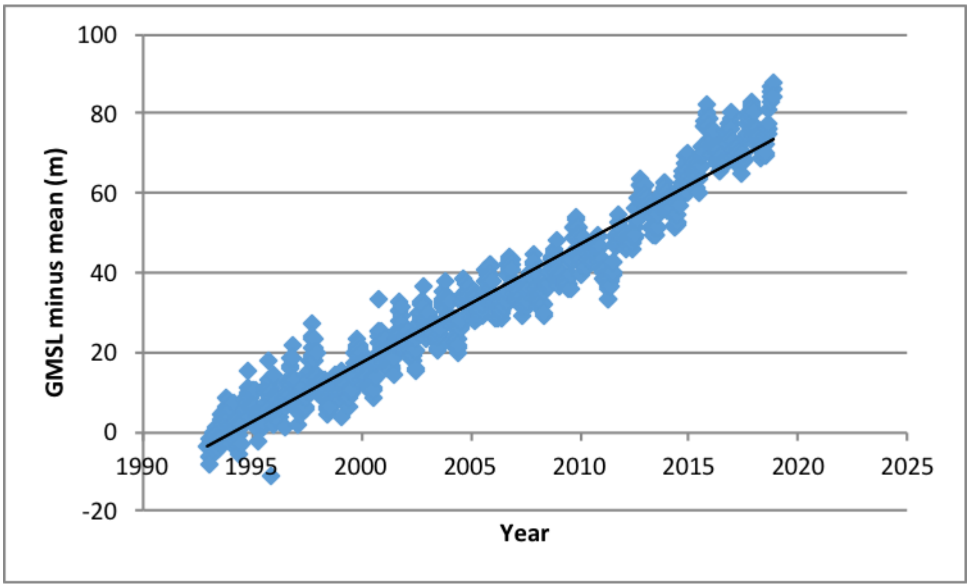
8f To find its equation go to 'more trend line options' and select 'display equation on chart' this will give you the equation $y = 2.9911x - 5999.1$.

The gradient is the annual increase which is approximately 3 metres per year.

8g i −33.8 metres

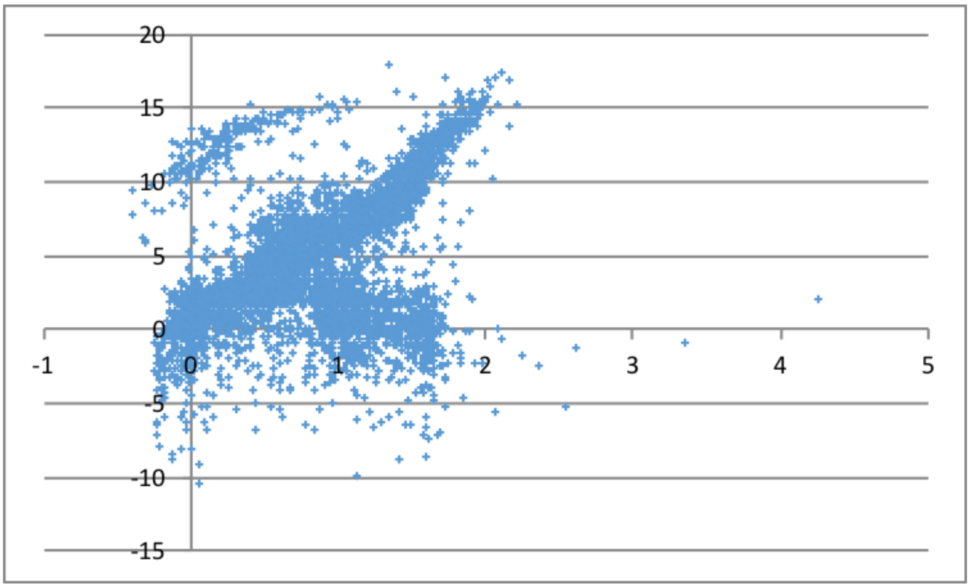
8g ii Follow the Excel instructions.

8g iii

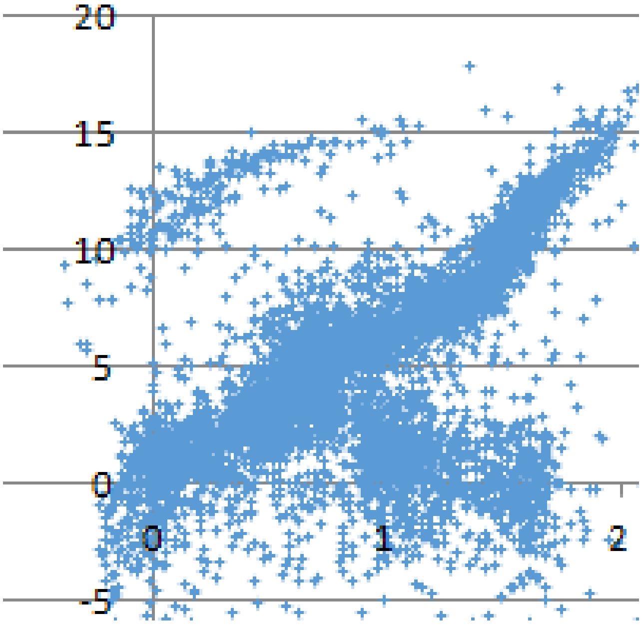


9 Answers will vary

11d



11e



12 Answers will vary

Solutions to Chapter review

1a Mean

$= \frac{4+7+9+2+4+5+8+6+1+4}{10}$

$= \frac{50}{10}$

$= 5$

Ascending order is 1, 2, 4, 4, 4, 5, 6, 7, 8, 9

Median is 5.5th term $= 4.5$

Mode: 4

Range $= 9 - 1 = 8$

1b Mean

$= \frac{16+17+14+13+18+15+16+15+11}{9}$

$= \frac{135}{9}$

$= 15$

Ascending order is 11, 13, 14, 15, 15, 16, 16, 17, 18

Median is 5th term $= 15$

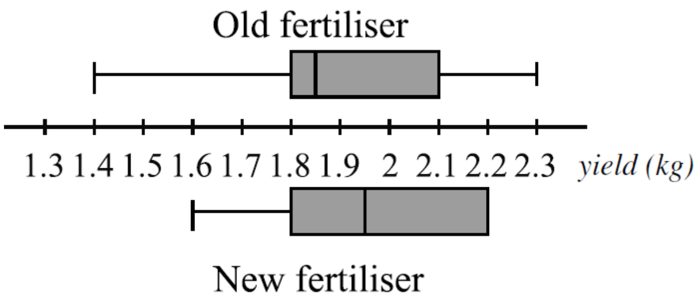
Mode: 15, 16 (bimodal)

Range $= 18 - 11 = 7$

2a Old fertiliser: median $= 1.85$, $Q_1 = 1.8$, $Q_3 = 2.1$

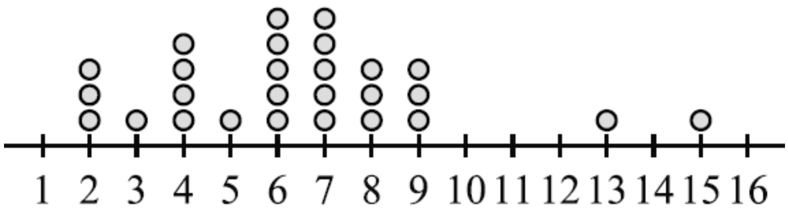
New fertiliser: median $= 1.95$, $Q_1 = 1.8$, $Q_3 = 2.2$

2b



2c The fertiliser does appear to increase his yield — the median yield has increased by 100 g. Probably more data are required because the lower quarter (0–25%) shows an increase, but the maximum has reduced. These claims, however, are each being made on the basis of one data point. 3

3a



3b By eye, 13 and 15 look like outliers.

3c median is 14th term $= 6$

Q_1 is 7th term $= 4$

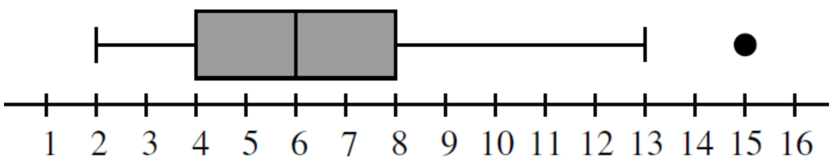
Q_3 is 21st term $= 8$

$\text{IQR} = 8 - 4 = 4$

$Q_3 + 1.5 \times \text{IQR} = 8 + 1.5 \times 4 = 14$

By the IQR criterion, 15 is an outlier, but 13 is not.

3d



4a mean ≑ 11.82 s, standard deviation ≑ 0.537 s

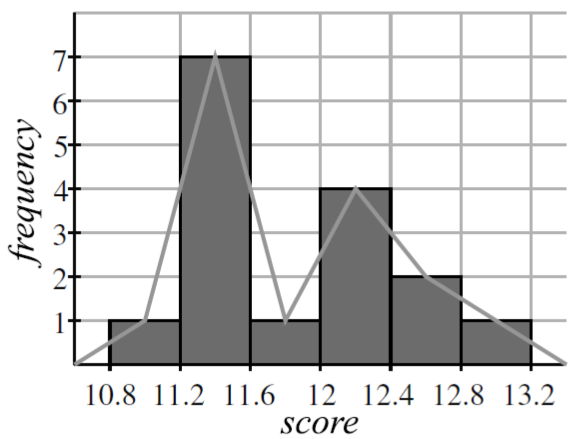
4b

group	10.8–11.2	11.2–11.6	11.6–12.0	12.0–12.4	12.4–12.8	12.8–13.2
centre	11.0	11.4	11.8	12.2	12.6	13.0
freq	1	7	1	4	2	1

4c mean ≑ 11.85 s, standard deviation ≑ 0.563 s.

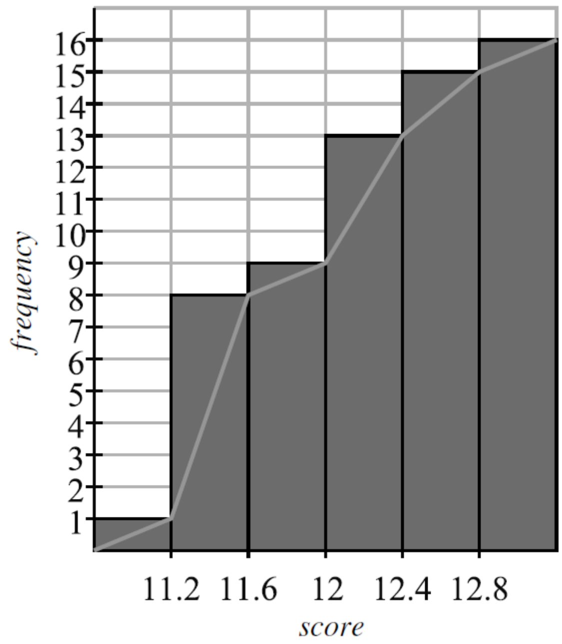
Agreement is reasonable, but as expected, the answers are not exactly the same.

4d



4e 0.5 seconds is a big difference in the time of a 100 metre sprint — the scale would be too coarse.

4f



4g The line at 50% of the data (frequency 8) meets the polygon where the sprint time is 11.6 seconds. You can confirm that this agrees with the result for splitting the grouped ordered data into two equal sets.

5a

	first	second	Total
order entrée	45	42	87
no entreé	38	28	66
Total	83	70	153

5b 153 people attended the restaurant that night.

5c $P(\text{customer ordered an entrée})$

$$= \frac{87}{153}$$

$\doteqdot 0.57$ or 57%

5d Percentage of customers who attended the first sitting

$= \frac{83}{153} \times 100\%$

$\doteqdot 54\%$

5e $P(\text{order entrée} \mid \text{attend first})$

$= \frac{45}{83}$

$\doteqdot 0.54$ or 54%

$P(\text{order entrée} \mid \text{attend second}) = 42 \div 70 = 60\%.$

$= \frac{42}{70}$

$= 0.60$ or 60%

No, it is not correct.

5f $P(\text{attend first} \mid \text{ordered an entrée})$

$= \frac{45}{87}$

$\doteqdot 0.52$ or 52%

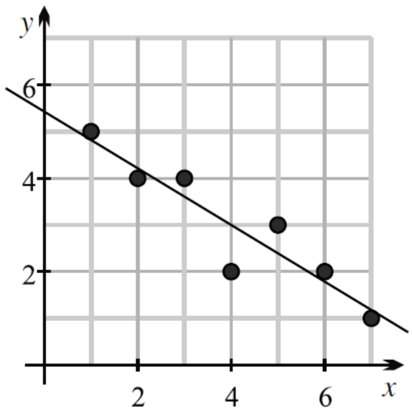
5g Using the answer to part e, 60% of customers in the second sitting ordered entrée.

60% of $90 = 0.6 \times 90 = 54$ entrées

5h Those attending the first session may prefer a quick meal before heading out to the theatre or some other event. There may also be more family groups operating on a tighter budget.

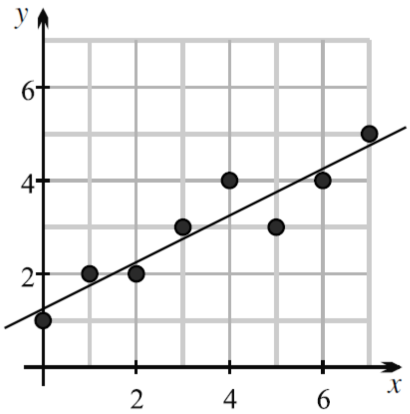
5i If they can estimate the demand on certain dishes, then they may be able to prepare parts of the dish in advance, for example, preparing the garnishes or chopping the ingredients.

6a i



6a ii $r = -0.93, y = -0.61 + 5.43$

6b i



6b ii $r = 0.94, y = 0.5x + 1.25$

7a 120 000 (from the graph)

7b 94 000, 62 000, 80 000, 80 000 (from the graph)

7c Total annual arrivals $= 94\ 000 + 62\ 000 + 80\ 000 + 80\ 000 = 316\ 000$

Average per quarter

$= \dfrac{316\ 000}{4}$

$= 79\ 000$

7d The arrivals may vary over the year because of seasonal or other effects. Government policy may consider an annual immigration quota, allowing a higher rate in one quarter to be balanced by a low rate in a subsequent quarter. As in 2000, examining the average for each quarter balances out such effects.

7e First quarter of 2000: 84 000 arrivals (from the graph)

7f It would be important to know the emigration rate of those leaving the country. The Net Overseas Migration (NOM) may be the better measure for many purposes. Other information of interest might include country of origin, destination within Australia, and whether they're intending to stay permanently or for a limited period.

7g i $y = 2.7(2000.16) - 5328.8$

$= 71.632$ (thousands)

$\doteqdot 71\ 600$

7g ii Rounding error has affected these calculations — a discrepancy in the second decimal place of the gradient is multiplied by 2000, resulting in an answer that is out by as much as $0.05 \times 2000 = 100$ thousand.

7g iii $y = 2.70633(2000.16) - 5328.8$

$= 84.293$ (thousands)

$\doteqdot 84\ 300$

This is in agreement with part d.

7g iv $y = 4(2.70633(2000.16) - 5328.8)$

$= 660.1996$ (thousands)

$\doteqdot 660\ 000$

7g v Percentage change

$= \dfrac{660 - 316}{316} \times 100\%$

$\doteqdot 109\%$ increase