

Statistics

2

Maths boosts sales



Name: Adam Driussi

Job: Director of The Quantum Group, an actuarial consulting firm

Qualifications:

Bachelor of Economics (Actuarial Studies), Fellow of the Institute of Actuaries of Australia and Graduate Diploma in Applied Finance and Investment

What do you do with a sharp analytical mind and great communication skills? You become an actuary!

Actuaries analyse large quantities of data and provide strategic advice to help businesses solve a wide range of practical problems. Adam Driussi is director of an actuarial consulting firm. He says, 'Move coming to work every day and having to solve a different problem to the day before.

The fact that you tend to get paid very well to do it is a very nice bonus.'

As an actuary, Adam and his team might analyse a client's customer database. Based on what they discover they will make a suggestion as to how a business can market to their customers more effectively. Adam explains: 'For example, eBay communicates with most of its customers via email. So we analysed the data to help eBay develop email campaigns that promoted new items that are relevant to each of eBay's millions of customers—as opposed to promoting the same items to every customer. The result was more than double the number of people responding to the campaigns and buying new items from eBay.'

Adam loved maths at school. 'My maths skills form the foundation for the problem solving and logic that I need to do my job every day. Without that foundation you could never solve the problems you get posed as an actuary.'

Why learn this?

We live in an age of 'Big Data', but we need to organise and analyse the data to draw conclusions or make decisions. Statistics are used by all sorts of organisations to understand complicated situations, make decisions and convince others of what to do. A knowledge of the processes used to gather and present data will help you to distinguish credible facts from unlikely fiction.

After completing this chapter you will be able to:

- find five-number summaries and use them to understand, analyse and compare data sets
- understand and construct data representations such as frequency curves, tables, box plots, dot plots, column graphs, histograms and scatter plots
- understand and construct data representations involving time-dependent bivariate data
- investigate and analyse the use of statistics in the media

10A

- draw and interpret lines of best fit
- calculate the mean and standard deviation and use these values to compare data sets.

Prepare for this chapter by attempting the following questions. If you have difficulty with a question, you can download a Recall Worksheet from the eBook or the Pearson Places website.

- 1 Find the mean, median (middle number), mode (most common number) and range (difference between largest and smallest numbers) for each of these sets of data.

(a) The number of siblings for your five best friends:
2 4 5 1 4

(b) The shoe size of your nine young cousins:
3.5 4.5 5.0 5.5 5.0 4.5 5.5 5.0 3.5

- 2 Classify each of the following as being nominal, ordinal, discrete or continuous data.

(a) Number of goals scored by a hockey team each week.

(b) Heights, in metres, of the members of your maths class.

(c) Sizes of shirts (S, M, L) on a rack.

- 3 The heights (in centimetres) of 20 students are as follows:

179.7 167.3 168.6 164.7 182.8 160.5 179.2 163.0 177.7 171.3
169.8 173.2 164.7 172.1 165.6 171.9 164.8 166.7 169.1 166.7

Round the heights to the nearest cm and then:

- (a) draw a stem-and-leaf plot using intervals of 5 cm (b) find the median.

- 4 The distance jumped (in metres) by each competitor in the long jump was recorded and is shown in the frequency table.

(a) Find an estimate for the value of the mean.

(b) In which class interval is the median?

Class interval (m)	Frequency
5.75–<6.00	4
6.00–<6.25	6
6.25–<6.50	9
6.50–<6.75	13
6.75–<7.00	5
7.00–<7.25	3

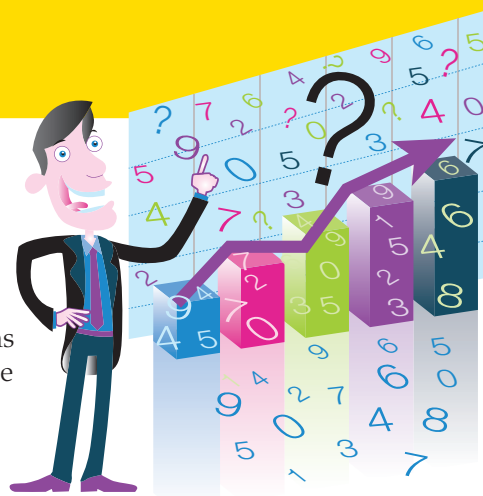
Exploration Task



You can download this activity from the eBook or the Pearson Places website.

Problems with problems

In this activity you will explore how to work with statistics problems that have more than one possible answer. How can the answers be checked? What might all the answers have in common?



Cumulative frequency curves

2.1

Measures of centre for ungrouped data

Various **measures of centre** (sometimes called **measures of location**) can be used to represent the 'centre' of a set of numerical data. Each measure gives a useful piece of information about the data. Three measures of centre are the **mean**, the **median** and the **mode**.

The **mode** is the most common data value, so it is the data value with the highest frequency. Showing all the data values in a table can make it easier to identify the mode. If two (or more) data values have the same highest frequency, then the data is called **bimodal** (or **multimodal**) and all data values with the highest frequency are modal values.

The **median** is the middle value when the data set is written in order. When the number of data values n is odd or even:

- odd—the median is the middle data value, which is the $\left(\frac{n+1}{2}\right)$ th value.
- even—the median is the value between the two middle values, which is the mean of the two middle values. You find this by adding together the $\left(\frac{n}{2}\right)$ th value and the $\left(\frac{n}{2} + 1\right)$ th value and dividing by 2.

Showing all the data values in a table together with their **cumulative frequency** values is a useful way to identify the median. The cumulative frequency is the sum of the frequencies up to each data value.

The **mean** is the sum of all the data divided by the number of data values n .

$$\bar{x} = \frac{\sum x}{n}$$

Showing all the data values in a table can make it easier to calculate the mean.

- Each data value x can be multiplied by its frequency f to find the product xf .
- The column of frequency values f is added to find the sum $\sum f$, which is the number of data values n .
- The column of xf values is added to find the sum $\sum xf$, which is the sum of all the data
- The mean is the sum of the xf values (sum of all the data) divided by the sum of frequency values (number of data values).

$$\bar{x} = \frac{\sum xf}{\sum f}$$

The mean is often called the **average**, which is used in everyday language to describe something that is typical or unexceptional.

Measures of centre for ungrouped data

Mode: most frequently occurring data value

Median: middle data value

- n is odd: $\left(\frac{n+1}{2}\right)$ th data value
- n is even: mean of the $\left(\frac{n}{2}\right)$ th and $\left(\frac{n}{2} + 1\right)$ th data values.

$$\text{Mean: } \bar{x} = \frac{\Sigma x}{n} = \frac{\Sigma xf}{\Sigma f}$$

Worked example 1**W.E. 1**

Find the mode, median and mean for each set of data correct to 1 decimal place, if necessary.

(a) 6.2, 5.7, 2.1, 8.7, 1.3

(b)	x	16	17	18	19	20
	f	8	20	15	4	1

Thinking

- 1 Check for repeated values.
- 2 State the value of the mode.
- 3 Order the data and locate the middle.
- 4 State the value of the median.
- 5 Write the formula for the mean.
- 6 Substitute the values and evaluate the mean.

Working

(a) There is no mode.

The mode does not exist.

1.3, 2.1, 5.7, 6.2, 8.7

 $n = 5$ (odd)

$$\frac{n+1}{2} = \frac{5+1}{2} = 3$$

The 3rd data value is 5.7.

The median is 5.7.

$$\bar{x} = \frac{\Sigma x}{n}$$

$$\begin{aligned} \text{mean} &= \frac{1.3 + 2.1 + 5.7 + 6.2 + 8.7}{5} \\ &= \frac{24}{5} \\ &= 4.8 \end{aligned}$$

- (b) 1 Find the highest frequency and matching data value.
- 2 State the value of the mode.
- 3 Extend the original table to include a cumulative frequency column.

(b) $f = 20$, $x = 17$

The mode is 17.

x	f	Cumulative frequency
16	8	8
17	20	$8 + 20 = 28$
18	15	$28 + 15 = 43$
19	4	$43 + 4 = 47$
20	1	$47 + 1 = 48$

4 Decide on the middle position.

$$n = 48 \text{ (even)}, \frac{n}{2} = 24$$

The middle is between the 24th and 25th data values.

5 Find the value of the median.

The 9th–28th data values are all 17.
The median is 17.

6 Extend the original table to include an xf column.

x	f	xf
16	8	128
17	20	340
18	15	270
19	4	76
20	1	20
Total	48	834

7 Find the total of the f and xf columns.

8 Calculate the mean and round to the specified number of decimal places.

$$\begin{aligned} \bar{x} &= \frac{\sum xf}{\sum f} \\ &= \frac{834}{48} \\ &= 17.4 \text{ (1 d.p.)} \end{aligned}$$

Measures of centre for grouped data

Large numerical data sets may be presented in a frequency table with the data grouped into class intervals. Although this makes the data set easier to deal with, it means that the original data values are lost.

When dealing with grouped data, the mode, median and mean can be found in terms of class intervals rather than individuals values.

- A modal class interval can be found rather than a modal value. The modal class interval is the most common class interval with the highest frequency.
- For the mean, the midpoint of each class interval is used as its representative data value. The mean can then be estimated by finding the mean of these representative values. However, this estimate will not be as accurate as a calculation from the raw data values.
- The median class interval is the ‘middle’ of the class intervals, but it is more useful to find an estimated median by using a **cumulative frequency curve** (sometimes called an *ogive*).

A cumulative frequency curve, graphed against the data values, will show the sum of frequencies ‘up to’ each data value. This means it shows the number of data values that are ‘less than’ each particular data value.

By definition, half of the data values will be less than the median, so you can estimate the median by finding the half-way point on a cumulative frequency curve, as in the following example.

Note that when the number of data values is large, then the cumulative frequency values

associated with the data values $\frac{n}{2}$ and $\frac{n+1}{2}$ will each give very similar values for the median.

Worked example 2

W.E. 2

The table shows the weights (kg) of 50 crates of equipment.

- (a) Find the modal class.
 (b) Calculate an estimate for the mean weight.
 (c) Use a cumulative frequency curve to find an estimate for the median weight.

Class interval (kg)	Frequency
45–<50	3
50–<55	13
55–<60	8
60–<65	10
65–<70	4
70–<75	12

Thinking

- (a) 1 Find the highest frequency and matching interval.
 2 State the modal class.
- (b) 1 Find the midpoint (x) of each interval.
 2 Include an xf column.
 3 Find the total of the f and xf columns.

- 4 Calculate the mean.

Working

(a) $f = 13$, 50–<55

The modal class is 50–<55 kg.

(b)

Class interval (kg)	x	f	xf
45–<50	47.5	3	142.5
50–<55	52.5	13	682.5
55–<60	57.5	8	460
60–<65	62.5	10	625
65–<70	67.5	4	270
70–<75	72.5	12	870
Total		50	3050

$$\begin{aligned}\bar{x} &= \frac{\sum xf}{\sum f} \\ &= \frac{3050}{50} \\ &= 61\end{aligned}$$

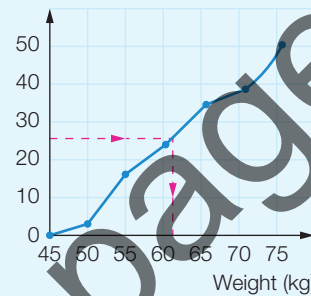
The estimated mean is 61 kg.

- (c) 1 Extend the original table to include a cumulative frequency column and an individual data (x) column with an extra row at the top under the headings.
- 2 Let the x values be the same as the upper values of each interval with the extra place at the top matching the lower end of the first interval.
- 3 Fill in the cumulative frequency column with the number of data 'less than' each x value (Here, 0 then 3, then $3 + 13 = 16$ and so on.)
- 4 Graph the cumulative frequency against x values with data values on the horizontal axis.
- 5 Join the points with straight lines (although sometimes a smooth curve is used).
- 6 Locate the half-way position on the cumulative frequency axis ($\frac{n+1}{2}$ gives 25.5 but $\frac{n}{2} = 25$ is simpler and sufficiently accurate).
- 7 Draw a horizontal line to the graph, then a vertical line to the data axis.
- 8 State the estimated median value.

(c)

Class interval (kg)	x	f	Cumulative frequency
<45	45	–	0
45–<50	50	3	3
50–<55	55	13	16
55–<60	60	8	24
60–<65	65	10	34
65–<70	70	4	38
70–<75	75	12	50

Cumulative frequency



The median is about 61 kg.

Measures of centre for grouped data

The modal class is the interval with the highest frequency.

The mean can be estimated by using the midpoint of each class interval as its representative value. Mean: $\bar{x} = \frac{\sum x}{n} = \frac{\sum xf}{\sum f}$

The median can be estimated from the cumulative frequency curve:

- Use the endpoints of the intervals as the data values.
- Write the number of data 'less than' each of the endpoints, which is the cumulative frequency.
- Plot the points and join with straight lines.
- Read across from half-way up the cumulative frequency axis to ($\frac{n+1}{2}$, or $\frac{n}{2}$ is usually close enough), then down to the data axis.

Grouped discrete data

For discrete data that has been grouped into intervals, the cumulative frequency still represents the number of data values 'less than' each data value.

Consider the following grouped discrete frequency table and cumulative frequency table. There are no data values less than 20, there are five data values less than 30, etc. All 34 values are less than 70.

Class interval	Frequency
20–29	5
30–39	8
40–49	17
50–59	3
60–69	1

Class interval	Cumulative frequency
<20	0
<30	5
<40	13
<50	30
<60	33
<70	34

2.1 Cumulative frequency curves

Navigator

Answers
p. 771

1, 2, 3, 4, 5, 6, 7 (a–b), 8

1, 2, 3, 4, 5, 6, 7 (a–b), 8

1, 2, 3, 4, 6, 7, 8

Fluency

W.E. 1

- 1 For each of the following data sets, calculate the (i) mode, (ii) median and (iii) mean. Round the mean to 1 decimal place where necessary.

(a) 5, 12, 2, 7, 8, 11, 4

(b) 22, 27, 35, 42, 15, 18, 30, 44, 28, 17

(c)

x	25	26	27	28	29
f	1	5	12	18	10

(d)

x	0	1	2	3	4	5	6
f	8	10	14	2	6	0	1

W.E. 2

- 2 For each set of grouped data (i) find the modal class, (ii) find an estimate for the mean and (iii) use the cumulative frequency curve to find an estimate for the median. Round the mean value correct to 1 decimal place where necessary.

(a)

Class interval	Frequency
0–<10	4
10–<20	5
20–<30	8
30–<40	11
40–<50	7

(b)

Class interval	Frequency
250–<300	37
300–<350	58
350–<400	26
400–<450	16
450–<500	9
500–<550	4

3 For each set of discrete grouped data find (i) the modal class, (ii) an estimate for the mean and (iii) use the cumulative frequency curve to find an estimate for the median. Round the mean value correct to 1 decimal place where necessary.

(a) The number of students present each day.

Class interval	Frequency
450–459	2
460–469	7
470–479	8
480–489	15
490–499	10
500–509	8

(b) The number of points scored by a rugby team.

Class interval	Frequency
0–9	2
10–19	3
20–29	3
30–39	4
40–49	1
50–59	1

4 The frequency table on the right lists the lengths of fish caught, measured in centimetres. Find the median length by constructing a cumulative frequency curve.

Class interval (cm)	Frequency
30–<35	7
35–<40	10
40–<45	5
45–<50	12
50–<55	13
55–<60	4

5 The following questions relate to the cumulative frequency graph shown on the right.

(a) The number of values in the data set is:

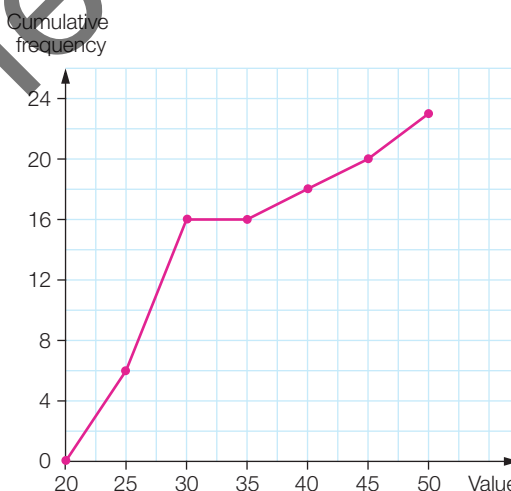
- A 23 B 24
- C 26 D 50

(b) The median is closest to:

- A 12 B 27
- C 28 D 29

(c) The class interval that has a frequency of zero is:

- A 20–<25 B 30–<35
- C 30–<36 D 31–<36



Understanding

6 The number of people boarding each of 25 successive trains at a station was as follows:

6 16 21 29 39 46 42 29 21 29 35 21 17
 10 8 18 17 26 7 9 12 19 26 20 21

- (a) Find the median of this raw data.
- (b) How many trains have fewer than 18 people?
- (c) Group the data into the intervals 0–9, 10–19 etc. to draw a cumulative frequency curve.
- (d) Use the graph to estimate the median number of people boarding each train and the number of times there were fewer than 18 people.
- (e) Compare your answers in (d) with your answers from the raw data in (a) and (b).

Reasoning

- 7 There is another way to estimate the median for grouped data, using the class interval that contains the median. If you know that the median is the third of eight results in a class interval $20 < 25$ (or $20-24$ for discrete values), then the median is approximated by the value $20 + \frac{3}{8} \times 5 \approx 21.9$. Round to 1 decimal place in a calculation such as this.
- Use this method to find an approximation for the median for Question 2(a).
 - Compare this with the value obtained using the cumulative frequency curve.
 - Use this method to find an approximation for the median for Question 3(a).
 - Compare this with the value obtained using the cumulative frequency curve.

Open-ended

- 8 Use the following information to construct a discrete data set with at least 19 results and then calculate the mean.
- median = 39, odd number of results
 - median = 72.5, even number of results

Problem solving

Product codes

Most products that you buy have a barcode attached to them. This barcode represents the International Article Number (EAN-13) and is a 13-digit unique code for that product. As part of the system to ensure the correct numbers are scanned by a barcode reader, the 13th digit is a 'check' digit which is not actually part of the product identification.

The following are all legal product codes:

9316090060604	9337090060609
9326090060603	9338090060606
9336090060602	9339090060603

- In the first column of codes, apart from the check digit, which digit changes? What do you notice about the sum of the two digits that are changing?
- Predict the 'check' digit in the code 934609006060^* .
- In the second column of codes, apart from the check digit, which digit changes? Do these digits follow the same pattern as the first column of codes?

Not every digit acts in the same way. In fact the even-placed digits all act in the same way and the odd-placed digits all act in the same way, as shown:

$\times 1$	$\times 3$	$\times 1$	$\times 3$	$\times 1$	$\times 3$	$\times 1$	$\times 3$	$\times 1$	$\times 3$	$\times 1$	$\times 3$	$\times 1$
9	3	1	6	0	9	0	0	6	0	6	0	4

- Calculate the sum of these multiplied values.
- Do the same for each of the EAN-13 codes given. What conclusion can you reach?
- Now calculate the check digit in each of the following EAN-13 codes:
 - 933919006060^*
 - 933929006060^*
 - 933928006060^*
 - 933927006060^*

Strategy options

- Work backwards.
- Test all possible combinations.

2.7

Statistics in the media

Discussions in the media often involve statistics, especially when someone wants to justify an opinion. You need to think carefully about what you hear and read, because statistics are not always as they seem. Some of the issues that might arise include:

- insufficient evidence being provided to support the conclusions made
- percentages used without an indication of the actual size of the survey
- raw figures used that are based on very different populations
- oversimplification of the results.

This is not to say everything you read or hear is incorrect, but it is a good idea to think critically about the key statistical issues involved. As an example, consider the two tables that follow, which accompanied a newspaper article.

Dressing with danger			
% of clothes that cause injuries		% of injuries caused by clothes treated at Victorian hospitals	
Shoes	48.1	Fracture, excluding tooth	18.8
Jumper, jersey, cardigan, shirt	13.0	Sprain/strain	17.2
Buttons	9.2	Open wounds	13.3
Jeans/pants	8.3	Foreign body	9.6
Socks	7.4	Burn	9.0
Clothes, not specified	5.8	Superficial	8.4
Zippers	4.7	Muscle/tendon injury	5.2
Nightwear	3.5	Dislocation	4.9
		Eye	2.9
		Bites	1.2
		Other	9.5

Source: Victorian Injury Surveillance and Applied Research System

At first glance, it appears that clothing injuries could be quite a serious problem. However, when you consider how many injuries in total get treated at Victorian hospitals, very few of them would be caused by clothing.

2.7 Statistics in the media

Navigator

Answers
p. 781

1, 2, 3, 4, 5, 6, 8

1, 2, 3, 4, 5, 6, 7, 8

1, 3, 4, 5, 6, 7, 8

Fluency

- 1 A section of an article from the London *Daily Mail* is shown on the following page.
 - (a) True or false: The article rates sports stars and managers in terms of their earnings in sport. Explain your answer.
 - (b) Who of the top 10 rich list, if any, experienced a drop in wealth in the year shown?

(c) Apart from the new people on the list, which sportsperson experienced the biggest percentage increase in their wealth over the year? State the percentage increase to the nearest whole percentage.

Lewis Hamilton remains Britain's richest sports star with £88m fortune as Wayne Rooney overtakes Jenson Button into second

- Lewis Hamilton's fortune has spiralled by £20m to £88m in the past year
- Wayne Rooney is now second in the Sunday Times Sport Rich List
- The Manchester United captain's fortune is up £12m to £72m
- F1 driver Jenson Button is £1m behind Rooney in third place

By Martyn Ziegler, Press Association

Published: 09:01 EST, 25 April 2015

Updated: 10:51 EST, 25 April 2015

Lewis Hamilton remains the richest sportsman in Britain with a fortune of £88 million but Wayne Rooney has overtaken Jenson Button to be in second place in the 2015 Sunday Times Sport Rich List.

Formula One world champion Hamilton, who lives as a tax exile in Monaco, remains on pole position in the list but his fortune has spiralled by £20 million on last year.

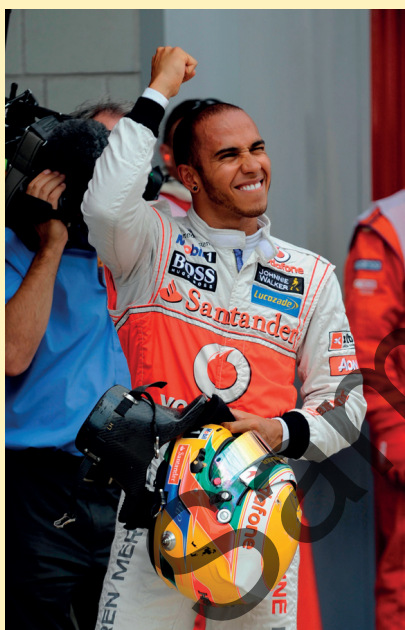
Hamilton has his sights on retaining his world championship title

after winning three of the first four races of the season, and is set to sign a new contract with Mercedes worth £27 million a year.

The list was released just hours after Hamilton was pictured on the Rome set for the filming of comedy sequel Zoolander 2 on Friday.

Hamilton was seated front row with Olivia Munn and busily chatting in between takes for the film, which will be released in February 2016.

Hamilton is making his first live-action cameo on the big screen having previously lent his voice to Disney animation Cars 2.

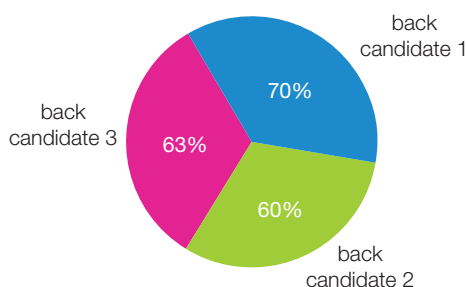


Position	Sportsperson	Sport	2015	2014
1	Lewis Hamilton	Motor racing	£88m	£68m
2	Wayne Rooney	Football	£72m	£60m
3	Jenson Button	Motor racing	£71m	£63m
4	Andy Murray	Tennis	£48m	£40m
5	Rio Ferdinand	Football	£44m	£44m
6	Steven Gerrard	Football	£42m	£37m
7=	Luoi Deng	Basketball	£40m	£36m
7=	Jose Mourinho	Football	£40m	New
9	Frank Lampard	Football	£39m	£37m
10=	Sir Nick Faldo	Golf	£38m	New
10=	Ryan Giggs	Football	£38m	£36m
10=	Rory Mclroy	Golf	£38m	£28m
10=	Arsene Wenger	Football	£38m	New

2 (a) In the early stages of the nomination process for the Republican Party leading up to the 2012 election for the US President, an American news network produced a graph similar to the one shown here.

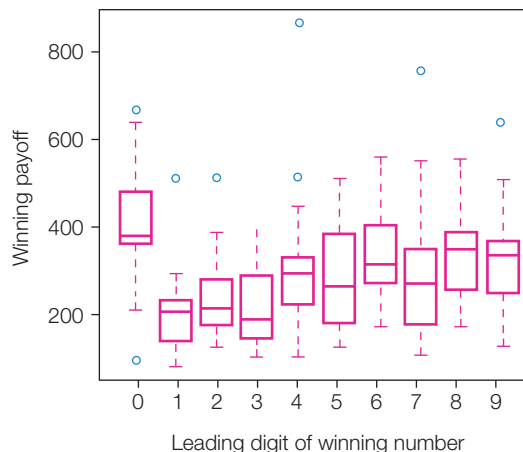
What problems can you see with this graph? Why do you think the graph has been drawn this way? Could you draw an accurate graph from the information provided?

2012 Presidential Run
Republican Candidate



- (b) Some graphs provide a simple and accurate visual representation of information. This box plot shows the winnings (\$) in the New Jersey Pick-It Lottery. In this lottery you pick any three-digit number you like and the total prize pool is divided by the number of winning entries.

What do you think this graph is telling you and why do you think that is the case?



Understanding

- 3 Consider the following article and graph from Roy Morgan Research.

The brew crew: Australia's heaviest coffee drinkers

Is Australia kicking the caffeine habit? In the last decade, the average coffee consumption by Australian adults has declined slowly but steadily, from 10.5 cups to 9.2 cups per week. Or could we just be drinking a stronger brew, thus reducing the need for so many cups? After all, café visitation is on the rise, as is ownership of coffee-making machines.

The proportion of Australians aged 18+ who go to a café for a coffee or tea in an average 3-month period has grown gradually from 54% in the year to December 2009 to 56% in the year to December 2013. Meanwhile, the increase in people who own coffee makers has shot up, from 28% in 2009 to 36% in 2013.

Among Australia's biggest coffee drinkers are people who work long hours. In the year to December 2013, those who worked 60+ hours in any given week consumed an average 10.1 cups weekly, compared to 8.6 for non-workers or 8.8 for those who worked 35–39 hours.

In news that probably won't surprise the parents out there, it seems that having children also increases our need for caffeine. Whereas the average weekly coffee consumption for people who don't have kids is 7.2 cups, it rises to 9.6 cups for parents.

Perhaps a little more surprising is that consumption increases with the age of the kids. So while parents of infants under 2 years old actually drink less coffee (8.8 cups per week) than the national average, those with kids aged between 12 and 15 drink an average of 10.3 cups.

The rise of the home coffee maker

Even though it's hard to imagine when they'd get time to use it, because they work such long hours, people who work 60+ hours per week are significantly more likely than the average Aussie to have a coffee maker at home, with 44% of them owning one (up from 38% in the year to December 2009). People with kids also over-index on coffee-maker ownership at 39%, up from 32% in 2009.

Angela Smith, Group Account Manager—Consumer Products, Roy Morgan Research, says:

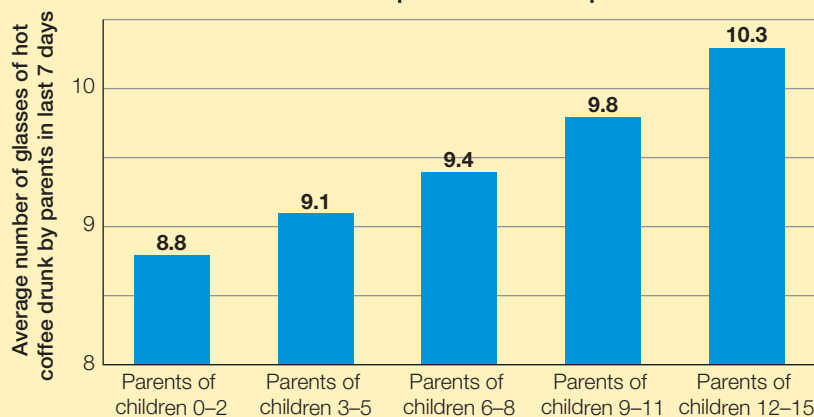
'The increased presence of coffee makers—which can be anything from stovetop cafetières to 'pod' machines such as Nespresso and Expressi—in Australian households is good news for the manufacturers of these items.

Interestingly, it hasn't impacted adversely on café visitation: possibly because people are developing a taste for 'real' coffee over instant.

'While it makes sense that people who work long hours would consume more coffee, their need for caffeine goes beyond this, to the point where they also drink more Cola and energy drinks than people who work fewer hours.

'The news that parents of older children drink more coffee in an average week than those of infants may seem surprising, considering the stereotype of the sleep-deprived new parent, but this is simply a function of age. Our data shows that older people drink more coffee, and parents of older children are typically older than those of infants. Mind you, their extra caffeine intake could also be linked to the sleep they lose through lying awake at night, worrying about where their kids are or what they're up to on Snapchat...!'

Kids and caffeine: average weekly coffee consumption of Australian parents



Source: Roy Morgan Single Source (Australia), Jan–Dec 2013

Base: Australians 18+

Margin of Error

The margin of error to be allowed for in any estimate depends mainly on the number of interviews on which it is based. Margin of error gives indications of the likely range within which estimates would be 95% likely to fall, expressed as the number of percentage points above or below the actual estimate. Allowance for design effects (such as stratification and weighting) should be made as appropriate.

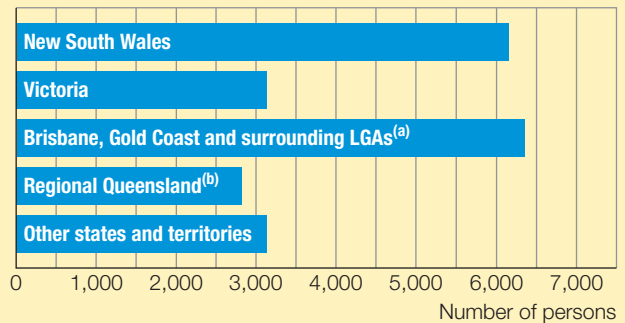
Sample Size	Percentage Estimate			
	40%–60%	25% or 75%	10% or 90%	5% or 95%
5000	±1.4	±1.2	±0.8	±0.6
7500	±1.1	±1.0	±0.7	±0.5
10 000	±1.0	±0.9	±0.6	±0.4
20 000	±0.7	±0.6	±0.4	±0.3
50 000	±0.4	±0.4	±0.3	±0.2

- (a) Considering the graph, what is the trend of coffee consumption when comparing parents with children in different age groups?
 - (b) According to the article, why is this trend surprising?
 - (c) What alternative reason was given to explain the trend?
 - (d) When parents are compared to non-parents, which group tends to drink more coffee? What is the difference in the number of cups consumed per week?
 - (e) Which group of people drink the greatest amounts of coffee? What else, according to the article, does this group of people have more than the average Australian? Why do you think this is so?
 - (f) According to the article, coffee consumption has decreased but visiting cafés and owning coffee makers has increased. What reasons can you think of to explain this?
- 4 The following article appeared on the Australian Bureau of Statistics website in 2015. (Note: 'LGA' stands for 'local government area'.)

The graph on the right shows the number of older Australians who moved to selected 'sea change' areas of the Queensland coast and their place of origin.

Around 70% of the sea-changers moved into separate houses. Almost half of these people moved into a house with 3 bedrooms. The sea-change moves were generally made by those in the younger end of the age range, with those between 55 to 64 in 2006 making up over 14,000 (or 67%) of the 22,000 movers.

Sea change moves are often associated with changes in people's employment circumstances, for example following retirement. Amongst the 14 000 55–64 year old movers to selected coastal LGAs in Queensland, 37% were employed in 2006 but were not in the labour force in 2011. This was higher than when compared to all 55–64 year olds in Australia, where 19% were employed in 2006 but were not in the labour force in 2011.



Footnote(s): (a) Includes LGAs of Brisbane, Gold Coast, Ipswich, Morton Bay, Redland and Logan. (b) Includes all other LGAs in Queensland outside of Brisbane, Gold Coast, Ipswich, Morton Bay, Redland and Logan and the selected coastal local government areas.

Source(s): Microdata: Australian Census Longitudinal Dataset, 2006-2011

- (a) Explain why this article only appeared in 2015, when the initial data dates from 2006.
- (b) How many people aged 55 and over shifted to the coast from somewhere else in Queensland in 2006? Estimate how many of these people retired during the period 2006–2011. Comment on the reliability of any assumptions you made.
- (c) Did the article suggest that a person was more likely or less likely to retire after a sea change, compared to over 55s generally? Explain your reasoning.

5 The following article appeared on-line in *Medical Xpress* in November 2015.

Parents' top fears about teens' cellphone use—are they justified?



Parents' fears about their teenagers' heavy use of cell phones and social media may be exaggerated, according to a new report from Duke University researchers. However, there are important exceptions in the areas of cyberbullying and sleep disruption.

'Each generation worries about how young people are using their time,' said Candice Odgers, associate professor in Duke's Sanford School of Public Policy and associate director of the Duke Center for Child and Family Policy. 'We see young people constantly on their phones and assume ill effects, but much of the research to date tells a more positive story.'

Teenagers' online lives closely resemble their experiences, connections and risks in the offline world, and cellphone use alone poses few entirely new dangers, Odgers said.

The article by Odgers and Duke Ph.D. candidate Madeleine J. George,

titled 'Seven Fears and the Science of How Mobile Technologies May Be Influencing Adolescents in the Digital Age,' appears online 18 Nov in the journal *Perspectives on Psychological Science*.

The review weighs commonly expressed fears regarding teenagers' use of mobile devices against existing research evidence. It calls for more rigorous research to evaluate how these quickly evolving technologies are impacting young people's lives.

'We tend to count hours spent using technology, rather than seek to understand the reasons teens are immersed in the digital world. When we look closely, we see considerable overlap between the underlying motivations and content of online versus offline communications and activities,' Odgers said.

There is little question that American adolescents are constantly connected. Almost 90% of adolescents

own, or have access to, a mobile phone. They spend an average of 1.5 hours a day text messaging and the vast majority access the internet from their phones. They devote an average of 7.5 hours a day to digital media of all kinds.

But, contrary to the early internet age—when a small minority of teens were online and heavy internet use was a sign of offline problems—now, teens' online worlds mirror their offline lives.

Teens with strong offline social networks tend to reinforce and strengthen their relationships through online interactions, the review found. Rather than connecting with strangers, most adolescents use digital media to interact with friends and acquaintances already in their face-to-face social networks.

'The overlap between offline and online connections is so striking that viewing what happens online as somehow separate from teens' real lives' is a false distinction,' said George.

For adolescents struggling with existing relationships, though, more time spent online can predict declines in well-being.

'If parents have concerns about their teen's face-to-face social interactions or activities, they probably have more reason to be concerned about online activities,' said George.

Parent-adolescent relationships online also appear to mirror offline relationships, the authors found. Although cellphone use may take away time spent with parents, if the existing relationship is strong, the new technology can allow more frequent, positive parent-child contact.

- The article explores the idea that adolescents are 'constantly connected'. Find three pieces of information used in the article to back this up.
- Is there evidence of how the data was collected? How would you imagine the data to be collected? Do you have any reason to suspect bias in the findings?

Reasoning

6 The following article appeared in the business section of the *Australian* in November 2015.

Can cabbages to Korea cover for coal's comedown?

THE AUSTRALIAN
12:00AM NOVEMBER 17, 2015

Rowan Callick
Asia-Pacific Editor

The South Korea story offers a salutary warning against expectations that Australia's bright new free-trade agreements will turn around our slumping export figures any time soon.

But it also showcases the fearlessness of Australian exporters—whose sales of cabbages and cauliflowers to the famous home of kimchi soared by 7944 per cent in the first half of this year.

Early data from the South Korea FTA, the first of the succession of deals concluded by Trade Minister Andrew Robb and his team of trade

negotiators, shows that while producers have rapidly seized their new opportunities, especially in agribusiness, the bottom line for Australian exports remains in decline. For the value of iron ore and coal, especially, within the trade profile with South Korea—half the value of all exports—is such that it will take years for other sales to bulk up sufficiently to compensate.

The same story applies to our other big export markets except the US. Those commodities accounted last year for 42 per cent of our exports to Japan, 58 per cent to India, and an astounding 65 per cent of our sales to China. In the first half of this year, iron ore receipts from sales to South Korea were 34 per cent down, to \$2.25 billion—a fall of \$1.2bn.

The good news is that the big rises in exports following the South Korea FTA, representing the most impressive increases across the board, comprise \$1bn.

Some of the agricultural successes are attributable not just to the FTA but at least as much to a shift in South Korea's import risk assessments, essentially the quarantine process, which formerly delayed accepting products for years.

This provides a huge boost, which surely is also driving a rise in jobs, especially in regional Australia.

Yet those combined revenues—not just the levels of increase but the entire sales—still fall \$200m short of the drop in value of Australia's iron ore exports alone to South Korea in the same six months.

Food boom		
Australian exports to Korea in the six months to July 1 post free trade agreement		
Fish exports: 2838% to \$852,000	Cabbages and cauliflowers: 7944% to \$724,000	Straw: 482% to \$2.3m
Wine: 39% to \$7.5m	Fresh or chilled beef: 26% to \$195m	Nuts: 234% to \$10.8m
Olive oil: 1067% to \$70,000	Other liquor: 5000% to \$2.65m	Frozen beef: 37% to \$476m
Citrus: 887% to \$1.16m	Other oils: 167% to \$917,000	Dentists' preparations: 1520% to \$243,000
Sheep and goat meat: 30% to \$35m	Grapes: 1357% to \$1.95m	Sausages: 1389% to \$804,000
Blankets: 4571% to \$1.3m	Butter sales: 93% to \$8m	Dates, figs and tropical fruit: 175% to \$250,000
Bread, cakes and biscuits: 930% to \$2.8m	Gold: 153% to \$240m	Potatoes: 64% to \$6.4m
Other fruit: 4616% to \$3.5m	Confectionery: 2500% to \$104,000	Furniture: 84% to \$991,000

- (a) An increase that 'soared by 7944 per cent' means it increased by how many times? Round your answer to the nearest whole number. How is it possible that such a huge percentage increase will not be enough to make up for the fall in value of other types of export?
- (b) Potatoes were up '64% to \$6.4m' and cabbages and cauliflowers were up '7944% to \$724 000'. What were the values of these exports 6 months previously?
- (c) What was the value of iron ore exports to South Korea 6 months previously?

- 7 The following fictional article is of a type often seen around election time. Read it and answer the questions that follow.

Poll – Voters back Smith

Mr John Smith's return as Prime Minister appears to have buoyed the government's electoral chances, with one in four voters saying they are at least 'a bit more likely' to vote for the government at the next federal election. Mr Smith was clearly preferred by female voters (47%

support) over the leader of the opposition, Ms Rosa Casey (41%).

Men preferred Ms Casey (54%) to Mr Smith (45%). Support for the opposition among men was 47%, compared to the government's (45%), while from women the government

attracted 47% support compared to the opposition's 44%. The survey also asked voters to identify those characteristics they believed Ms Casey or Mr Smith possessed. The results of the survey are summarised below.

SUNDAY BUGLE POLL					
Voting intentions		%	Preferred PM		%
ALP		46	Casey		48
Coalition		46	Smith		46
Other/Don't know		8	Neither/Don't know		6
SUNDAY BUGLE POLL: RATING THE LEADERS					
	Casey (%)	Smith (%)		Casey (%)	Smith (%)
Decisive	68	53	Sincere	22	55
Forward thinking	68	48	Too arrogant	75	15
Good for Australia	43	55	Honest	20	62
Weak	10	30	Capable	72	72
A good leader	66	43	None of the above	1	2

Source: Corio Market Research 1030 voters from Vic., NSW. Poll done by telephone.

- Do the survey results agree with the heading 'Voters back Smith'? Discuss.
- The survey was done by telephone—is this the best way to do such a survey? Discuss.
- Why have the results for choice of preferred PM been separated into male and female opinions in the article?
- Why don't the totals of male and female voters' opinions add up to 100%?
- Discuss the categories that describe Rosa Casey and John Smith, such as 'decisive' and 'forward thinking'. Comment on the appropriateness of these categories. What other categories might be appropriate?
- Is this survey just about the preferred Prime Minister or is it about something else?

Open-ended

8 The following advertisement was produced by the South Australian Government in 2014.

www.problemgambling.sa.gov.au

THE POKIES: THE ODDS

What are the odds? Odds are they are against you

If you gamble for the chance to win money, do you know what the odds are of actually winning?

Every form of gambling has an element of chance that helps determine the outcome. How much a person will win or lose is determined by the odds of the game and how much money is staked. It is important to remember that the 'house' (eg the casino, hotel or club) always has the advantage and that the odds are against the gambler.


THE ODDS ARE CLEARLY STACKED AGAINST THE GAMBLER - MEANING YOU SHOULD ALWAYS EXPECT TO LOSE.

THIS TABLE SHOWS THE ODDS OF WINNING ON SOME FORMS OF GAMBLING IN SOUTH AUSTRALIA

Bet	Odds of Winning
Poker Machines - Getting 5 Black Rhinos on Black Rhinos Machine (Top Prize) (\$1 Bet per line)	1 in 9,765,625
Lotto - Winning First Division (playing 1 game)	1 in 8,145,060
Oz Lotto - Winning First Division (playing 1 game)	1 in 45,379,620
Powerball - Winning First Division (playing 1 game)	1 in 54,979,155
Keno - Winning First Division (Spot 10, playing 1 game)	1 in 8,911,711
Keno - Winning First Division (Spot 9, playing 1 game)	1 in 1,380,687
Super 66 - Winning First Division (playing 1 game)	1 in 1,000,000
The Pools - Winning First Division (playing 1 game)	1 in 2,760,681
Instant Scratchies - Winning First Division with \$1 High Tier (playing 1 game)	1 in 1,500,000
Trackside - picking, at random, a trifecta in a 13 horse race	1 in 1,715

NOW SEE HOW YOUR ODDS OF WINNING COMPARE WITH NON-GAMBLING RELATED ACTIVITIES

Odds of non-gambling activities	Odds of Occurrence
Marriage ending in divorce	1 in 2.3 marriages
Chances of going bald - if you're a man	3 in 4 men
Dying from heart disease	1 in 4 people
Having your car stolen	1 in 142 cars
Dying from a venomous bite or sting	1 in 1,000,000 people
Being killed by lightning	1 in 1,603,250 people


GAMBLINGHELPLINE
1800 858 858
 24/7 • FREE • CONFIDENTIAL

* Some figures taken from Centre for Gambling Research Fact Sheet - Gambling Odds 2003

The South Australian Government developed this advertisement in 2014 with funds from the Gamblers Rehabilitation Fund.

What do you think about this advertisement? What do you think about the statistics that are quoted in the advertisement? How would they have been calculated? Do you think this would be effective in reducing problem gambling?



Newspaper polls

Equipment required: newspapers, internet access

Some newspapers conduct a daily poll on an issue of current interest. A couple of examples are shown below, taken from the *Herald Sun*.

Should politicians be banned from shopping centres?		Should Australia abandon the Union Jack and create a new flag?	
Total: 1695 votes		Total: 16 404 votes	
YES	NO	YES	NO
1266 votes	429 votes	5561 votes	10 843 votes
74.6%	25.4%	33.9%	66.1%

The Big Question

Are these polls reliable? To answer this you will have to consider what reliable means. This is not always a simple matter and you need to address the following issues.

- Is the question a fair question?
- Is the polling method reasonable?
- If the question was asked again the next day, would a similar result be expected?

Sometimes questions are phrased in such a way as to lead to a particular answer. In the polling industry this is known as a push poll.



Engage

To participate in these polls you can either call a telephone vote line or vote online.

- 1 What type of poll, of those you have studied, are these closest to in their design?
- 2 Why do you think there is such a difference in the total number of votes for the two examples shown? Is this likely to influence you in how reliable you think the results might be?

Explore

See if you can find some similar examples online or in another newspaper.

Explain

- 3 Is there a problem with these polls from the point of view of how the sample is obtained?
- 4 Do a statistical analysis of the results obtained in these two polls, or on a couple of other similar polls that you have found for yourself. Do the results appear to be reasonable?

Evaluate

- 5 Do you think the wording of the questions should be changed? If so, write the question the way you would like it to be asked.
- 6 Why do you think newspapers conduct these polls on a daily basis?

Extend

- 7 Find out more about 'push polling'.
- 8 Make up some examples of these one-question polls for yourself, based on some issues of local or current interest. Decide how you would go about conducting your poll so that the results are reliable.

Strategy options

- Break the problem into manageable parts.
- Have I seen a similar problem?

Standard deviation

2.8

One of the most useful statistical measures of spread is the **standard deviation**. The symbol σ or the letters SD are often used to represent the standard deviation.

Consider the following data set:

5, 6, 7, 10, 10, 12, 14, 16

The mean of the data is 10.

You can now calculate how far each of these values is away from the mean.

This is referred to as the score's *deviation* from the mean. The deviation values are:

-5, -4, -3, 0, 0, 2, 4, 6

If you add these together you get 0, which is not very helpful. In fact, you will *always* get 0. (You might like to think about why this is true.) To avoid this zero, the deviations can all be squared (to give positive values) before they are added.

The squared deviation values are:

25, 16, 9, 0, 0, 4, 16, 36

These add to 106.

You can then divide this by the number of values, in this case 8, to find the average deviation from the mean, which is called the **variance**. Here, the variance is 13.25.

To find the standard deviation you then take the square root of the variance, which reverses the effect of the squaring. The standard deviation here is approximately 3.64.

Note that previously you used two measures of spread: range and interquartile range. The standard deviation is a more accurate measure of spread because it uses all data points, not just the difference between the extreme points.

$$\begin{aligned}\text{Variance } \sigma^2 &= \frac{\text{sum of (deviations from mean)}^2}{\text{number of values}} \\ \text{Standard deviation } \sigma &= \sqrt{\text{variance}} \\ &= \sqrt{\frac{\text{sum of (deviations from mean)}^2}{\text{number of values}}} \\ &= \sqrt{\frac{\sum(x - \bar{x})^2}{n}}\end{aligned}$$

Note: σ is the lower-case Greek letter sigma and Σ is the capital Greek letter sigma.

10A

Worked example 14

W.E. 14

Find the standard deviation, rounded to 2 decimal places, for the following data.

2, 3, 3, 3, 5, 7, 8, 11, 12

Thinking

- 1 Find the mean of the data set.
- 2 Find the deviations from the mean and then square them.
- 3 Add the squares of the deviations and divide by the number of values to find the variance.
- 4 Take the square root of the variance.

Working

$$\text{Mean: } \frac{2+3+3+3+5+7+8+11+12}{9} = 6$$

$$\text{Deviations: } -4, -3, -3, -3, -1, 1, 2, 5, 6$$

$$\text{Square of deviations: } 16, 9, 9, 9, 1, 1, 4, 25, 36$$

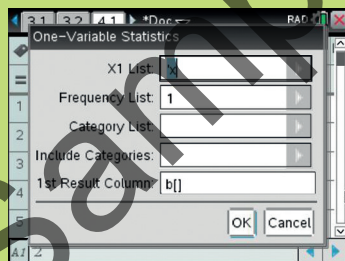
$$\text{Variance} = \frac{110}{9}$$

$$\text{Standard deviation} = \sqrt{\frac{110}{9}} = 3.50 \text{ (2 d.p.)}$$

This calculation is not practical if there are too many data values, but in that case technology can be used. Standard deviation can be calculated using your CAS as shown below.

Using TI-Nspire CAS

Add **Lists & Spreadsheet** to your document and enter the data in column **A**. Select **menu > Statistics > Stat Calculations > One-Variable Statistics...** and set **Num of Lists** to 1. Then set the **One-Variable Statistics** settings as shown.



The statistics data should appear as shown. Scroll through the list to find the σ_x standard deviation value.

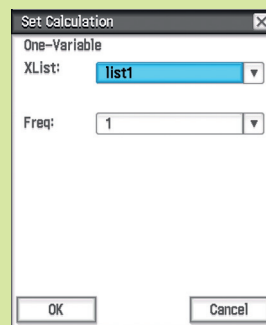
A	x	B	C	D
			=OneVar(
3	3	Σx	54.	
4	3	Σx^2	434.	
5	5	$s_x := s_n \dots$	3.7081	
6	7	$\sigma_x := \sigma_n \dots$	3.49603	
7	8	n	9.	

C6 = 3.4960294939005

This is 3.50, rounded to 2 decimal places, just as found when calculating by hand.

Using Casio ClassPad CAS

From the menu select **Statistics** and enter the data in **list1**. Select **Calc > One-Variable** and set the **XList** as **list1** and **Freq** as 1 as shown.



The statistics data should appear as shown. Scroll down to find the σ_x standard deviation value.

Stat Calculation	
One-Variable	
x	=6
Σx	=54
Σx^2	=434
σ_x	=3.4960295
s_x	=3.7080992
n	=9
minX	=2
Q_1	=3
Med	=5
Q_3	=8

This is 3.50, rounded to 2 decimal places, just as found when calculating by hand.

Many scientific calculators can be used to calculate the mean and standard deviation. You can also use spreadsheet software.

In Excel you use **average** to calculate the mean and **stdevp** to calculate the standard deviation for a population.

	A	B	C
1	2		
2	3		
3	3		
4	3		
5	5		
6	7		
7	8		
8	11		
9	12		
10	3.496029		
11			

If you have a large data set in a frequency table you can still use your CAS to simplify the calculations. If the data is grouped, then use the class centre as the value in the first column.

The following example shows how to find the standard deviation for the following list of times taken by students to complete the 100 m sprint at the School Sports Carnival.

Time (s)	Frequency
10.50–<10.75	2
10.75–<11.00	4
11.00–<11.25	7
11.25–<11.50	3
11.50–<11.75	5
11.75–<12.00	10

Using TI-Nspire CAS

Enter the class centre values in column **A** and the frequencies in column **B**. Then select **menu > Statistics > Stat Calculations > One-Variable Statistics...** and set **Num of Lists** to 1. Then set the **One-Variable Statistics** settings as shown.

A	time	B	freq	C	D
1	10.625	22			
2	10.875	4			
3	11.125	7			
4	11.375	3			
5	11.625	5			

Using Casio ClassPad CAS

Enter the class centre values in **list1** and the frequencies in **list2**. Select **Calc > One-Variable** and set the **XList** as **list1** and **Freq** as **list2** as shown.

Set Calculation

One-Variable

XList: list1

Freq: list2

OK Cancel

Using TI-Nspire CAS

The statistics data should appear as shown. Scroll through the list to find the σ_x standard deviation value. You can also confirm that $n = 31$, the sum of the frequencies.

A	B	C	D
time	freq		=OneVar(
10.875	4		11.4073
11.125	7	Σx	353.625
11.375	3	Σx²	4039.23
11.625	5	sx := sn...	0.422009
11.875	10	σx := σn...	0.415146

Using Casio ClassPad CAS

The statistics data should appear as shown. Scroll through the list to find the σ_x standard deviation value.

You can also confirm that $n = 31$, the sum of the frequencies.

One-Variable	Value
\bar{x}	=11.407258
Σx	=353.625
Σx^2	=4039.2344
σ_x	=0.4151464
s_x	=0.4220088
n	=31
minX	=10.625
Q_1	=11.125
Med	=11.375
Q_3	=11.875

Interpreting the standard deviation

For two data sets that have the same mean, a smaller standard deviation value means that the data set is more closely packed. The data set with the smaller standard deviation is not necessarily bigger, stronger, longer (or whatever the variable represents), but just more consistently located near the mean value.

The mean and the standard deviation both use all of the data values, so they are often linked together. In a similar way, the median and the IQR are usually not affected by the extreme data values, so they are often linked together.

Worked example 15

W.E. 15

The table below shows the test results for 15 students in English and Mathematics. Both tests were scored out of 50. Find the mean and standard deviation for each data set and use these values to compare them. Round the values to 2 decimal places where necessary.

English	42	44	31	20	25	34	45	50	44	12	13	23	44	47	42
Maths	38	34	42	39	40	44	43	27	34	27	40	32	31	22	23

Thinking

- 1 Calculate the mean and standard deviation.
- 2 Comment on the statistical values calculated.

Working

English: mean = 34.4 SD = 12.39

Maths: mean = 34.4 SD = 6.98

The two tests had the same mean score but the standard deviation was much less for the Mathematics test. This means the scores for the Mathematics test were more consistent. However, you can see that in the English test both the best and worst results, overall, were recorded.

To analyse data, you need to decide which statistical values will be important to consider. Remember that the mean and standard deviation use all the data values. The median and IQR effectively use only the middle 50% of the data values. If there are outliers, it may be better to use the median and IQR. If the data set seems relatively uniformly distributed, then it may be better to use the mean and standard deviation.



Constructing distributions for the mean and standard deviation of simple random samples

It is often impractical to collect data from a whole population, so a random sample is taken that represents the population. The mean and standard deviation of the sample is calculated, and these statistics can be used to estimate the mean and standard deviation of the population. Small samples can be unreliable, but larger samples should give values close to the population.

The formula for the mean of a sample is the same as for the mean of the population.

$$\text{Mean: } \bar{x} = \frac{\sum x}{n} = \frac{\sum xf}{\sum f}$$

The formula for the standard deviation of a sample is slightly different from the population formula, using $n - 1$ instead of n , to help correct for mathematical sampling bias.

$$\text{Population standard deviation: } \sigma_n = \sqrt{\frac{\sum(x - \bar{x})^2}{n}}$$

$$\text{Sample standard deviation: } s \text{ or } \sigma_{n-1} = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

Note that your CAS offers both of these values each time you select 'calculate' for a set of data. Simply choose the ' $n - 1$ ' or ' s ' alternative for sample standard deviation. The sample standard deviation value will always be larger than the population standard deviation value.

Worked example 16

W.E. 16

The data below gives the mass in grams of a population of 50 river rocks to be used in an ornamental garden. In this question, round mean values to 1 decimal place and standard deviations to 2 decimal places.

13 37 44 48 31 17 63 67 37 32 35 13 91 65 23 28 18 32 39 42 54 23 21 29 26
22 18 34 36 38 24 18 17 18 17 30 18 18 30 25 29 26 29 34 20 22 27 29 23 17

- Find the mean and standard deviation of the mass of the river rocks.
- Use random sampling methods to create 12 samples of 5 values each. Label the samples A, B, C, ... L. Find the mean and standard deviation of each sample.
- Combine the samples of 5 in pairs to create 6 samples of 10: A and B; C and D; etc. Find the mean and standard deviation of each sample of 10.
- Combine the samples of 10 in pairs to create 3 samples of 20: A, B, C, D; E, F, G, H; etc. Find the mean and standard deviation of each sample of 20.
- Arrange the values of the sample means and standard deviations around the population values for each of the three different sample sizes. Comment on the distributions.

Thinking

Working

- Enter the data into a CAS.
- Read the mean value. Round correct to 1 decimal place, if necessary.
- Read the population standard deviation. Round correct to 2 decimal places, if necessary.

$$\begin{aligned} \text{(a) Mean} &= 30.9 \text{ g} \\ \text{SD} &= 15.25 \text{ g} \end{aligned}$$

(b) 1 Use a random number generator to produce 12 sets of 5 whole numbers from 1 to 50. Go through the set of 50 numbers picking out the selected data. For example, in Set A you pick the 26th, 39th, 22nd, 3rd and 2nd score.

(b) This is a sample answer only.

A: Random numbers: 26, 39, 22, 3, 2
Data: 36, 30, 23, 44, 37

B: Random numbers: 25, 28, 12, 36, 21
Data: 26, 24, 13, 30, 54

C: Random numbers: 1, 29, 5, 13, 8
Data: 13, 34, 31, 91, 67

D: Random numbers: 48, 49, 38, 18, 44
Data: 29, 23, 18, 32, 34

E: Random numbers: 45, 17, 25, 19, 21
Data: 20, 18, 26, 39, 54

F: Random numbers: 14, 34, 8, 9, 12
Data: 65, 18, 67, 37, 13

G: Random numbers: 22, 37, 7, 10, 36
Data: 23, 18, 63, 32, 30

H: Random numbers: 36, 15, 23, 17, 2
Data: 30, 23, 21, 18, 37

I: Random numbers: 34, 1, 22, 28, 25
Data: 18, 15, 23, 24, 26

J: Random numbers: 40, 11, 48, 24, 19
Data: 25, 35, 29, 29, 39

K: Random numbers: 12, 29, 22, 25, 18
Data: 13, 34, 23, 26, 32

L: Random numbers: 13, 10, 30, 33, 32
Data: 91, 32, 18, 17, 18

- 2 Enter each set of data into a CAS.
- 3 Read the mean value. Round correct to 1 decimal place, if necessary.
- 4 Read the sample standard deviation. Round correct to 2 decimal places, if necessary.

This is a sample answer based on the samples drawn in part (b).

A: Mean = 34 g, SD = 7.91 g

B: Mean = 29.4 g, SD = 15.13 g

C: Mean = 47.2 g, SD = 31.29 g

D: Mean = 27.2 g, SD = 6.61 g

E: Mean = 31.4 g, SD = 15.06 g

F: Mean = 40 g, SD = 25.38 g

G: Mean = 33.2 g, SD = 17.57 g

H: Mean = 25.8 g, SD = 7.66 g

I: Mean = 20.8 g, SD = 5.26 g

J: Mean = 31.4 g, SD = 5.55 g

K: Mean = 25.6 g, SD = 8.32 g

L: Mean = 35.2 g, SD = 31.81 g

- (c) 1 Enter each set of data into a CAS.
 2 Read the mean value. Round correct to 1 decimal place, if necessary.
 3 Read the sample standard deviation. Round correct to 2 decimal places, if necessary.

(c) This is a sample answer based on the samples drawn in part (b).

A, B: Mean = 31.7 g, SD = 11.63 g

C, D: Mean = 37.2 g, SD = 23.79 g

E, F: Mean = 35.7 g, SD = 20.19 g

G, H: Mean = 29.5 g, SD = 13.36 g

I, J: Mean = 26.1 g, SD = 7.56 g

K, L: Mean = 30.4 g, SD = 22.50 g

- (d) 1 Enter each set of data into a CAS.
 2 Read the mean value. Round correct to 1 decimal place, if necessary.
 3 Read the sample standard deviation. Round correct to 2 decimal places, if necessary.

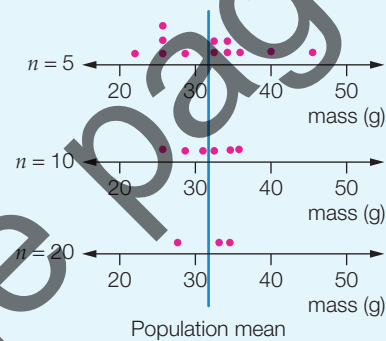
A, B, C, D: Mean = 34.5 g, SD = 18.44 g

E, F, G, H: Mean = 32.6 g, SD = 16.96 g

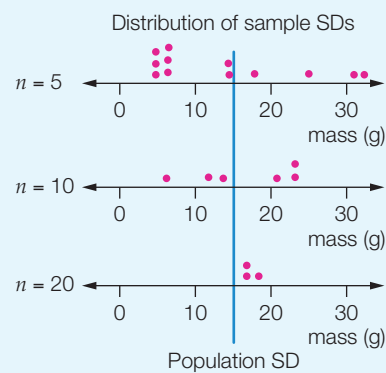
I, J, K, L: Mean = 28.3 g, SD = 16.48 g

- (e) 1 Draw an axis for each set of mean values and mark the population mean.
 2 Mark each sample mean on the axis, like a dot plot. Mark the position of the population mean.

(e) Distribution of sample means



- 3 Repeat the process for the standard deviations.



- 4 Comment on the distributions saying how they are similar and how they are different depending on sample size.

The mean and SD values for the samples were distributed either side of the population values. For the larger samples the sample values clustered closer to the population values. The larger the sample size, the closer the sample statistics are to the population.

2.8 Standard deviation

Navigator

Answers
p. 782

1, 2, 3, 4, 5, 6, 8, 10

1, 2, 3, 4, 5, 6, 7, 8, 10

2, 3, 4, 6, 7, 8, 9, 10

Fluency

W.E. 14

- 1 Find the standard deviation, rounded to 2 decimal places, for the following data.

11 12 15 15 17 19 21 24 26 29

W.E. 15

- 2 The table below shows the test results for 15 students in English and in Mathematics. Both tests were scored out of 50. Find the mean and standard deviation for each data set and use these values to compare them. Round the means to 1 decimal place and the standard deviations to 2 decimal places.

English	40	40	34	23	24	39	46	47	41	19	23	23	34	37	45
Maths	33	44	32	49	36	38	41	29	37	29	41	35	30	22	24

- 3 The grouped data shown below represents the distances, in km, travelled to school by a group of students. Estimate the mean and standard deviation (both in km) of the distances, using the centre of each class interval to represent the data points. Round the mean to 1 decimal place and the standard deviation to 2 decimal places.

Distance (km)	Frequency
0–<2	22
2–<4	28
4–<6	15
6–<8	7
8–<10	4
10–<12	2



Understanding

- 4 The following table shows the heights, in m, of the ten highest mountains in each of Australia, New Zealand, Brazil and Sri Lanka.

Australia	2220	2185	2178	2166	2132	2048	2047	2045	2050	1979
New Zealand	3754	3500	3440	3279	3201	3183	3176	3163	3160	3157
Brazil	2973	2890	2882	2861	2849	2798	2791	2770	2734	2033
Sri Lanka	2524	2243	2240	2100	2100	2076	2036	2016	2010	1896

- (a) Calculate the range of the data for each country.
 (b) Calculate the mean and standard deviation of the data for each country.
 (c) Write a sentence or two about what you have found.

- 5 (a) Find the mean, rounded to 1 decimal place, and standard deviation, rounded to 2 decimal places, for each of the following data sets:
- (i) 13 14 14 15 15 15 16 16 19 20 22 25 26 28
 - (ii) 13 14 14 15 15 15 16 16 19 20 22 25 26 48
 - (iii) 3 14 14 15 15 15 16 16 19 20 22 25 26 28
 - (iv) 3 14 14 15 15 15 16 16 19 20 22 25 26 48
- (b) Write a sentence or two about what this tells you.
- 6 You have recorded a data set as: 23 45 46 47 55 57 58 59 62 65 and calculated the mean and standard deviation. When checking your work you discover that the '23' should have been '43'. Changing this will:
- A increase both the mean and standard deviation
 - B increase the mean and decrease the standard deviation
 - C decrease the mean and increase the standard deviation
 - D decrease both the mean and the standard deviation.

Reasoning

- 7 The screenshot shows some information taken from a music playlist. You will need to be careful how you deal with this data. Give your answers correct to the nearest second.
- (a) Find the mean and standard deviation for songs listed as *Rock* (including *Rock*, *Classic Rock* and *RockNRoll*).
 - (b) Find the mean and standard deviation for songs listed as *Pop*.
 - (c) Find the mean and standard deviation for songs listed as neither *Rock* nor *Pop*.
 - (d) Draw a parallel box plot showing the three categories *Rock*, *Pop* and *Other*. Use decimal time values here.
 - (e) Write a couple of sentences comparing the three categories. Is there any evidence that *Rock* songs are a different length to *Pop* songs, for example?

Genre	Time	Genre	Time
Rock	2:07	Pop	2:25
Blues	3:25	Soundtrack	2:12
RnB	2:26	Rock	4:15
Blues	5:27	R&B	3:59
Rock	3:35	Blues	2:33
Country & F...	2:16	Blues	15:30
Rock	5:28	Rock	2:47
Rock	4:02	Rock	5:30
Blues	3:31	Rock	3:32
Blues	2:51	Rock	12:28
Rock	3:26	R&B	4:36
Rock	2:34	Rock	3:30
Pop	3:46	Pop	2:15
Classic Rock	4:01	Blues	3:17
Grunge	3:50	Country & F...	2:48
Pop	3:13	RockNRoll	3:05
Pop	2:25	Pop	3:00
Pop	2:30	Rock	3:51
Blues	2:09	Rock	3:43
Rock	3:31	R&B	3:56
Rock	5:51	Soundtrack	5:56
RockNRoll	1:59	Rock	2:52
Rock	3:32	R&B	4:10
Pop	3:49	Pop	3:52
Blues	4:20	Pop	2:40
Blues	2:46	Rock	3:34
Rock	3:20	Blues	6:00
Rock	3:32	R&B	4:47
Rock	3:32	Rock	5:55
Blues	4:07	Classical	3:57
Rock	8:33	Rock	3:42
Rock	8:53	Pop	4:14

Hint



Change the time in seconds to decimal parts of a minute.

Vic

W.E. 16

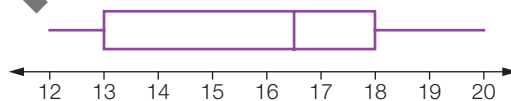
- 8 The data below gives the mass in grams of a population of 35 pebbles to be used in a flower arrangement. In this question, round mean values to 1 decimal place and standard deviations to 2 decimal places.

3 6 7 7 5 5 3 8 7 6 6 6 6 5 5 6 5 5
6 5 5 5 5 5 3 5 3 5 7 7 4 6 6 5 7

- Find the mean and standard deviation of the mass of the pebbles.
- Use technology to generate random numbers, and use these to randomly select 12 samples of 5 values each from the data. Label the samples A, B, C, ... L. Find the mean and standard deviation of each sample.
- Combine the samples of 5 in pairs to create 6 samples of 10: A and B; C and D; etc. Find the mean and standard deviation of each sample of 10.
- Combine the samples of 10 in pairs to create 3 samples of 20: A, B, C; D; E, F, G, H; etc. Find the mean and standard deviation of each sample of 20.
- Use a parallel dot plot to compare the distribution of sample means around the population mean for each of the three different sample sizes. Repeat the process for the distribution of sample standard deviations around the population standard deviation. Comment on the distributions.

Open-ended

- Construct a data set with at least 20 values that has a mean of between 20 and 22, a range of at least 15 and a standard deviation of between 2 and 4.
 - Write a couple of sentences about how you went about this task.
 - What effect did the condition about the range have on this task?
- Construct a data set, of at least 20 values, that would fit the following box plot and then find the mean and standard deviation of your data set.



Problem solving

Keeping score

From her attempts to shoot 40 points in the season's first basketball game, Angela scored only 10 points. This gave her a scoring average of 25%. In the second game she had the chance to score 30 points but only scored enough to raise her scoring average over the two games to 50%.

- How many points did she score in the second game?
- What was her scoring average for the second game?
- What is the greatest scoring average Angela could have had after the second game, assuming her maximum in the second game was still 30 points?

Strategy options

- Guess and check.
- Work backwards.



Chapter review

2

Maths literacy

bivariate data	interquartile range	primary data
box plot	outlier	quantile
cumulative frequency curve	parallel box plot	quartile
dependent variable	parallel dot plot	range
five-number summary	percentage cumulative frequency curve	scatter plot
independent variable	percentile	secondary data
10A extrapolation	line of best fit	standard deviation
interpolation	regression line	variance

Copy and complete the following using the words and phrases from the list, where appropriate. A word or phrase may be used more than once.

- Data taken from other people's research is _____.
- The difference between the upper quartile and the lower quartile is called the _____.
- A _____ is a graph that shows the sum of frequencies up to each data value.
- Data collected from your own observations is _____.
- A _____ is a graph that shows the minimum, lower quartile, median, upper quartile and maximum values for a data set.
- The value below which 80% of values are on a cumulative frequency curve can also be referred to as the 0.8 _____.
- The difference between the largest observed value and the smallest is called the _____.
- If a value is more than $1.5 \times \text{IQR}$ above Q_U or more than $1.5 \times \text{IQR}$ below Q_L then it is an _____.

10A 9 When you use a regression line to predict a value within the known range you are using _____. If the value is outside the known range you are using _____.

10A 10 The _____ is the square root of the _____.

Fluency

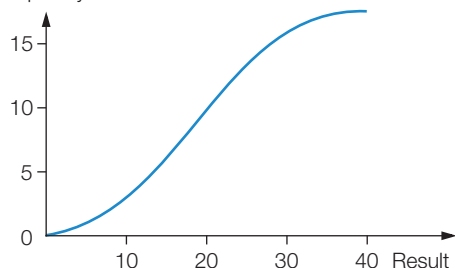
1 For each of the following cumulative frequency curves:

(i) What is the median?

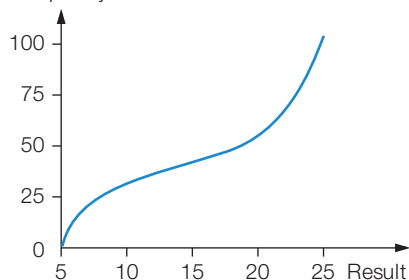
(ii) What is the interquartile range?

(iii) Draw a box plot.

(a) Cumulative frequency



(b) Percentage cumulative frequency



2.1, 2.2

2 For the set of data shown at right:

- (a) find the median by drawing a percentage cumulative frequency curve
- (b) find the interquartile range
- (c) draw a box plot.

Class interval	Frequency
120–<125	5
125–<130	17
130–<135	32
135–<140	54
140–<145	24
145–<150	13

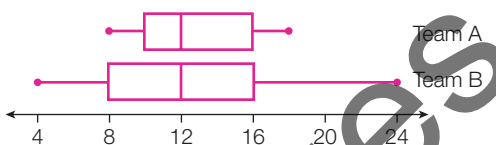
2.2

3 A frog in training achieves the following jump distances (in centimetres). On the basis of these jumps, how long would a jump need to be, for it to be considered an outlier?

19 25 17 22 20 19 16 15 17 21

2.2

4 You have been working in a group on an analysis task. One member of the group has produced the following parallel box plot for the data that has been collected. Your task is to write the first draft of the analysis. This should include a comment about which team you think has been more successful.



2.3

5 For each of the following occupations, state whether you think they deal with primary data or secondary data.

- (a) Meteorologist
- (b) Television weather forecaster
- (c) Insurance broker
- (d) Cancer research scientist

2.4

10A 6 Find, rounded to 2 decimal places, the standard deviation for the following data sets.

(a) 3 4 4 5 6 7 7 7 8 9 10 11 11 12 12 12 14 15 16

2.8

(b)

Number of flaws	Frequency
0	2
1	19
2	16
3	15
4	7

(c)

Weight (kg)	Frequency
0–<5	6
5–<10	11
10–<15	15
15–<20	8
20–<25	2

10A 7 Find the equation of the line of best fit, using a spreadsheet or some other form of technology, for each of the following sets of data. Round your answers to 3 decimal places, where necessary. In each case, the required rule is of the form $y = \dots$

2.6

(a)

x	0	1	2	3	4	5	6
y	3	7	13	20	26	30	35

(b)

x	0	2	5	7	10	12	14
y	18	15	11	5	0	-6	-11

Understanding

8 Consider the following data:

72.8 64.8 55.4 44.3 61.8 52.9 62.0 73.6 34.5 72.3
65.9 57.7 64.4 55.3 45.2 71.2 69.6 54.2 59.9 51.1

- Group the data into intervals of 10 ($30 < 40$, $40 < 50$ etc.) and draw a cumulative frequency curve.
- Calculate the five-number summary for the data.
- Find the median from the cumulative frequency curve and compare it with the median calculated from the raw data.

2.1

9 The rainfall (in millimetres) recorded on each day of a month in a particular location was:

12 15 17 16 25 2 44 3 2 0
4 6 8 10 0 11 3 6 3 8
6 15 15 17 0 22 17 21 30 33

- Calculate the range for the rainfall figures.
- Calculate the five-number summary for the data.
- Draw a box plot to represent the data.

2.2

10 For each of the following box plots:

- indicate what values would be considered as outliers
- decide whether there are any outliers in the data set displayed.



2.2

Reasoning

11 Some of the data available from the Bureau of Meteorology website for Perth (Metro) is shown below.

2.4

Temperature (°C)	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Annual
Mean maximum	30.9	31.3	29.5	25.6	22.4	19.3	18.3	18.8	20.0	22.9	26.2	28.8	24
Mean minimum	17.9	18.0	16.5	13.5	10.5	8.4	7.8	8.1	9.5	11.2	14.1	16.2	12

- Find the mean difference in temperature (°C) for each month.
- Draw a scatter plot showing mean maximum temperature, mean minimum temperature and mean difference in temperature.
- Describe the temperature pattern for Perth (Metro) using the graph as an aid.
- Could you use this graph to describe the temperature pattern for Kalgoorlie, in inland Western Australia? Give a reason for your answer.
- The daily maximum and minimum temperatures in Perth for June 2010 are given.

Minimum:	10.9	6.8	5.6	4.8	9.9	5.7	6.5	6.6	8.7	7.3	9.6
	6.9	4.3	2.2	2.3	6.1	12.0	9.7	10.8	13.1	6.4	9.4
	8.2	10.8	11.1	6.4	3.1	6.3	8.4	5.2			
Maximum:	16.2	17.9	17.6	17.0	16.9	14.3	15.5	15.9	15.1	16.3	16.1
	16.0	13.4	13.8	14.0	16.9	16.3	16.5	17.0	17.3	16.4	17.3
	18.2	18.2	15.2	14.2	14.5	14.6	11.9	14.0			

How did June 2010 compare to the overall data?

- 12 The number of colony forming units per 100 mL of water of two different bacteria taken from a particular lake at weekly periods is recorded in the following table. The table also shows the maximum temperature ($^{\circ}\text{C}$) for the day of recording.

Fecal coliform (CFU/100 mL)	350	73	180	150	200	130	2200	6300
<i>E. coli</i> (CFU/100 mL)	350	90	120	330	200	120	1600	5800
Temperature ($^{\circ}\text{C}$)	15.3	14.6	16.8	16.8	18	24	19.6	21.7

Fecal coliform (CFU/100 mL)	360	1200	560	1700	19	430	320	87
<i>E. coli</i> (CFU/100 mL)	300	980	630	1800	16	270	270	160
Temperature ($^{\circ}\text{C}$)	20.1	19.4	22.1	24.1	22	23	19	18

- (a) In how many different ways can the data be grouped to compare a pair of variables? Use technology to:
- (b) draw a scatter plot with fecal coliform on the x -axis and *E. coli* bacteria on the y -axis
- (c) find the equation of the line of best fit, assuming the data is linear.

- 13 The following fictional article contains information about the opinions of voters on the introduction of tolls on certain roads around Melbourne.

Toll roads – poll reaction

The State Government is facing a big electoral backlash over its new toll roads according to a *Sunday Bugle* poll conducted by Corio Market Research Centre in marginal seats along the route.

The poll found that at least half the voters surveyed avoid paying the tolls on the Tullamarine Freeway and Monash Freeway.

Residents of the affected areas have increased worries about traffic on their local suburban roads growing massively. The poll shows that the Government could lose three of the four seats in which the survey was conducted.

SUNDAY BUGLE POLL: TOLL ROADS

How marginal electorates would vote ...

	Government %	Opposition %
Tullamarine	33	47
Essendon	43	43
Oakleigh	38	38
Knox	46	32

... and who avoid tolls

Tullamarine	59%
Essendon	50%
Oakleigh	52%
Knox	49%

Source: Telephone poll of 1200 voters conducted by Corio Market Research Centre

- (a) Do the survey results support the article's suggestion of a 'big electoral backlash'?
- (b) Add up the percentages for Government and Opposition voters in the Tullamarine electorate. Why don't they add up to 100%? Why is this important to the judgement of the poll?
- (c) The poll was done by telephone. Which sampling methods would you use to get the best indication of how voters in each electorate are reacting to the toll roads?
- (d) What is the significance of the four different locations?

- 14 The owner of a company holds a record of the salaries paid to all the employees. Annual salaries 3 years ago and today are given in the following table.

2.3

Salary range (× \$10 000)	Number of employees (3 years ago)	Number of employees (currently)
1-<3	12	9
3-<5	15	10
5-<7	9	9
7-<9	5	8
9-<11	3	7
11-<13	0	5

- Convert the data for employees to percentages, and then draw a graph that emphasises the difference between salaries paid 3 years ago and those paid currently.
 - Contrast the means, medians and interquartile ranges for the two sets of data.
 - Draw a box plot for each set of data.
 - Describe, in your own words, the change in salaries over the past 3 years.
 - How much more is the company paying in salaries today than it was 3 years ago?
 - By what percentage has the average salary risen?
- 15 The grouped frequency table below shows the age distribution of the employees of a local company. The age given is the nearest whole number of years on 1 January of the current year. (So, if a worker was 22 years and 5 months old they would be recorded as 22, whereas a worker 22 years and 7 months old would be recorded as 23.)

2.1, 2.2

Age at 1 Jan (years)	16–20	21–25	26–30	31–35	36–40	41–45	46–50	51–55
Frequency	25	30	32	28	33	42	55	10

- Estimate the mean age, to 1 decimal place, giving a reason for your estimate.
- Calculate the mean and standard deviation for the ages of the employees, giving your answers rounded to 2 decimal places, where appropriate.
- Draw a cumulative frequency curve and use it to find the five-number summary.
- Draw a box plot of the data.
- Use your cumulative frequency curve to estimate the number of employees who are:
 - under 32
 - over 48.

2.8

10A