15 Displaying and interpreting data

Data bombard us constantly from every direction — prices of cars, daily temperatures and rainfall, birth weights of babies, marks in different subjects, how much milk people have for breakfast — and we struggle to make sense of it all. The subject of statistics begins as *descriptive statistics*, which develops various systematic approaches, using tables, graphs and summary statistics, so that we can see the big picture. Very quickly, however, statistics is combined with *probability theory*, inviting the language of *prediction* to be used, and leading perhaps to a discussion of *causation*, which is so fundamental in science.

In the Year 11 book, the foundations of probability theory and discrete probability distributions were developed, ending with some limited discussion of sampling and the way in which probability theory is related to statistical observations. This chapter and the next work the other way around, beginning in Chapter 15 with the raw data and ways of organising raw data. Then Chapter 16 moves to the relationship of that data to probability theory and continuous distributions.

Sections 15A–15C deal with *univariate* data, meaning that there is just one variable involved, such as the prices of cars or the birth weights of babies. Data may also be *bivariate*, such as when we measure the heights and weights of people to investigate how height and weight are related. Sections 15D–15F introduce the possible *correlation* between two statistical variables and the associated *line of best fit*.

Many opportunities are provided for investigations, for serious use of technology, and for possible projects, particularly in Exercise 15F.

Digital Resources are available for this chapter in the **Interactive Textbook** and **Online Teaching Suite**. See the *overview* at the front of the textbook for details.

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

15A Displaying data

Raw data come in small, large and huge unsorted lists, mostly of numbers, but also of categories. It is usually unrewarding to make much sense of such a list just by scanning through it. The first task of *statistics* is to provide tools for the analysis of data — Chapters 12 and 13 of the Year 11 book began this task.

There are three successive stages to analysing raw data.

- Display the raw data in various *tables* and *graphs* (or *charts*) to gain some overview of it, and perhaps some initial insight into what is happening.
- Carry out calculations of *summary statistics*. For univariate data, the most important are the *measures of location*, such as the mean and median, and the *measures of spread*, such as the variance, standard deviation and interquartile range. For bivariate data, we also use the *correlation* and the *line of best fit*.
- Speculate about patterns, predictions, and possible causal factors of the data, using *probability theory* to calculate theoretical probability distributions, followed by tests as to how well the data fit any suggested distribution.

All this may then be followed by suggestions for further experiments, in which the statistician may well be heavily involved in *designing the experiments* that will yield the next sets of raw data.

This chapter is about data — the table, the graphs, and the summary statistics. The next chapter deals with continuous probability distributions, and in particular with the normal distribution. Discrete probability distributions were discussed in Chapter 13 of the Year 11 book.

The Desmos graphing calculator which is embedded in the interactive textbook has a number of features to help visualise data and calculate summary statistics. Refer to the interactive textbook and teacher resources for Desmos guides and statistics activities.

A review of random variables

Here is a quick review of random variables from Section 13A of the Year 11 book.

1 EXPERIMENTS AND RANDOM VARIABLES

Random variables

- A *deterministic experiment* is an experiment with one possible outcome.
- A *random experiment* is an experiment with more than one possible outcome.
- A *random variable*, usually denoted by an upper-case letter such as X, is the outcome when a random experiment is run, and the various possible outcomes of the experiment are called the *values* of the random variable.

Scores and frequency

- When an experiment is run many times, the outcomes are called *scores*, and the (finite) list of all the scores is called a *sample*.
- The *frequency* of an outcome or value is the number of times it occurs.

Types of random variables

- A random variable may be *numeric* (if its values are numbers) or *categorical*.
- A random variable is called *discrete* if it is numeric and its values can be *listed*, meaning that it is possible to write them down in a sequence $x_1, x_2, x_3, \ldots$

ISBN 978-1-108-76630-2

Recording the country of birth of a person chosen at random in Australia is a categorical variable. Recording the number of overseas countries visited by that person is a numeric variable, which is discrete because its possible values 0, 1, 2, . . . can be listed.

Recording the height of a person is a numeric variable, but if we regard peoples' heights as real numbers rather than rounded measurements, then the variable is not discrete because we cannot list the set of possible values. This is a *continuous variable*, whose precise definition will be given in Chapter 16.

Frequency tables and cumulative frequency tables

The most basic object for organising and inspecting raw data is a *frequency table*, as introduced in Chapter 13 (Year 11). This table of frequencies can be produced digitally using a spreadsheet or database, or by hand using tallies.

When the data are numeric, we can also produce a *cumulative frequency table*, which tells us the number of scores less than or equal to a given score. For example, at the start of Year 7, Cedar Heights High School gave 40 students a spelling test marked out of 10. Here are the raw results, presented as univariate data:

4, 7, 2, 8, 7, 6, 3, 2, 8, 2, 9, 5, 8, 5, 8, 3, 6, 7, 5, 2,
10, 6, 7, 5, 6, 6, 9, 1, 5, 7, 8, 1, 6, 5, 7, 10, 6, 7, 8, 6,

and here are the tallies, the frequencies and the cumulative frequencies.

Mark	0	1	2	3	4	5	6	7	8	9	10	Sum
Tally		𝍸	𝍸𝍸	𝍸	𝍸	𝍸𝍸	𝍸𝍸	𝍸𝍸	𝍸𝍸	𝍸	𝍸	—
Frequency	0	2	4	2	1	6	8	7	6	2	2	40
Cumulative frequency	0	2	6	8	9	15	23	30	36	38	40	—

A quick glance at the cumulative frequencies suggests that 8–9 students have poor spelling, or perhaps they had little experience in earlier years doing tests.

2 CUMULATIVE FREQUENCY

- For numeric data, the *cumulative frequency* is the number of scores that are less than or equal to a given score.
- A frequency distribution table can be extended to a cumulative frequency distribution table by taking the accumulating sums of the frequencies.

If the values of a categorical dataset have been sorted into some sort of meaningful order, then a cumulative frequency table can also be produced — see the Pareto charts later in this section.

Finding the median from the cumulative frequencies

Two questions dominate the discussion of univariate data, as we saw in Chapter 13 (Year 11).

- Measures of location: Where is the centre of the distribution?
- Measures of spread: How spread out are the data?

We begin with the median, which is a measure of location.

ISBN 978-1-108-76630-2

3 THE MEDIAN — A MEASURE OF LOCATION

The *median* of numeric data — with symbol Q_2 meaning second quartile — is the middle score, when the scores are written out in ascending order. More specifically:

- For an odd number of scores, the median is the middle score.
- For an even number of scores, the median is the average of the two middle scores.

A cumulative frequency table makes it easy to pick out the median.

It may be helpful to review how to find the median by writing the scores out in order. Suppose first that there are an odd number of scores:

4 7 8 10 **10** 11 13 15 21 (9 scores)

↑

The median of the 9 scores is the middle score. This is the 5th score in the row, which is 10. Now suppose that there are an even number of scores:

4 7 8 10 10 **11** 11 13 13 21 (10 scores)

↑

There are now two middle scores — the 5th, which is 10, and the 6th, which is 11. The median is their average, which is $10\frac{1}{2}$.

Example 1 15A

Explain how to use the cumulative frequency table above of the 40 spelling test marks to calculate the median of the 40 scores.

SOLUTION

There are 40 scores, so the median is the average of the 20th and 21st scores.

From the table of cumulative frequencies, the 15th score is 5, and all scores from the 16th up to the 23rd score are 6.

Hence the 20th and 21st scores are both 6, so the median is 6.

Thus someone with a score of 6 is in the middle of the class.

Displaying categorical data

A glance at some newspapers shows the wide variety of tables and graphs that are used to display data. Such tables and graphs should be easy to read because one of their purposes is to display data to people who are interested in the subject matter of the table or graph, but may or may not know any statistics.

Tables and graphs may be intended as the first step of an open-minded analysis of an unknown situation. They may be a quick visual illustration of ideas expounded in accompanying text. They may be intended to argue a contentious point. Whatever their purpose, the first job of the reader is to look out for intended or unintended distortions created by the display. Statistics wrongly used routinely lead people astray.

Two examples displaying categorical data are given here — Pareto charts and two-way graphs. The exercises have examples of other displays, such as pie charts and bar charts. In every example, interpretation is vital.

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

Pareto charts

Any set of categorical data, or even discrete data, can be represented on a *Pareto chart*. Its main purpose, however, is to identify which problems in a business are most urgent, and it is classified as one of 'seven basic tools of quality'.

For example, Secure Roofs often arranges a repair, but for various reasons that repair does not take place on the scheduled day, causing loss of income for the company while salary and other expenses still have to be paid. The manager organised the last 200 such failures into six categories, as in the table to the right.

Problem	Frequency
Blackout	4
Employee not arriving	6
Illness of employee	16
Owner not home	64
Rain	88
Truck breakdown	22
Total	200

To construct a *Pareto chart*, first arrange the categories into descending order of frequency — this places the most serious issues first, because they are the first problems that need to be addressed. Then add a cumulative frequency column.

The Pareto chart consists of two graphs drawn together on the same chart:

- a frequency histogram with columns arranged in this descending order,
- a cumulative frequency polygon.

The chart usually has two vertical axes, one on the left and one on the right. On the left are the frequencies, on the right are the percentage frequencies.

Example 2 15A

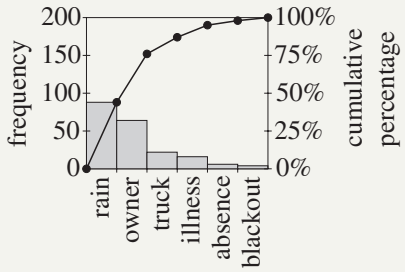
a Draw a Pareto chart of the data gathered by Secure Roofs.

b Describe what actions the manager may decide to take using the chart.

SOLUTION

a Here is the cumulative frequency table, with the categories arranged in descending order of frequency, and the cumulative frequencies calculated corresponding to that order. The Pareto chart is on the right.

Problem	Frequency	Cumulative
Rain	88	88
Owner	64	152
Truck	22	174
Illness	16	190
Absence	6	196
Blackout	4	200
Total	200	



ISBN 978-1-108-76630-2

b With this chart, the manager can go through the issues from left to right and attempt to deal with what is causing problems for the business, from the most serious to the least serious.
- First, perhaps, he will first decide only to schedule external roof repairs three days ahead, when forecasts are more reliable.
- Then perhaps he will personally ring each owner two days ahead with a friendly reminder, following up with an SMS the evening before.
- Perhaps he has budgeted for a new truck next year.
- Perhaps he knows that he has little control over the other three issues.

There are no firmly established conventions for drawing the details of a Pareto chart, and the conventions we have chosen are certainly not universal. Here are some details about the chart as we have drawn it.
- The rectangles of the histogram join up with each other.
- The cumulative frequency polygon starts at the left-hand bottom corner of the left-hand rectangle because the initial sum is zero.
  — The next plot is at the right-hand top corner of the first rectangle.
  — Each subsequent plot is above the right-hand side of each rectangle.

We will have more to say about histograms and polygons in Section 15B.

The cumulative frequency polygon in a Pareto chart is always concave down because the issues have been arranged in descending order of frequency. (The term 'concave down' is being used more loosely here than in Chapter 9 — here every chord lies 'below or on the curve' rather than 'below the curve'.)

Two-way tables (contingency tables)

A *two-way table* or *contingency table* consists of two or more related frequency tables put together. In its simplest form, it has only four numbers in the table, yet it is surprisingly complicated to interpret. The topic anticipates the discussion of bivariate data in Section 15D, and it involves also conditional probability from Section 12G of Year 11.

A survey asked 200 adults what colour phone cover they preferred. Responses were recorded as black–brown (Dark) or as coloured (Colour), and the gender of the person was also recorded. The resulting *joint frequencies* are tabulated to the right.

	Dark	Colour
Men	38	12
Women	56	94

Let us ask the question, 'Do men prefer dark colours more than women do?' If we glanced just at the joint frequencies 38 and 56 under 'Dark', we might conclude that women prefer dark colours more than men. The data are deceptive, however, because far more women were questioned than men — for reasons that we have not been told. We need to take this into account.

ISBN 978-1-108-76630-2

Example 3 15A

Explain how to analyse the two-way table above to answer the question, 'Do men prefer dark colours more than women do?'

SOLUTION

We can find the sums 50 and 150 of the two rows, and the sums 94 and 106 of the two columns. The grand total is 200, which checks the additions. These five numbers are called *marginal frequencies*. The bias of the sample towards women is now clear.

	Dark	Colour	Sum
Men	38	12	50
Women	56	94	150
Sum	94	106	200

Each of the three rows, and each of the three columns, is a frequency table. The last row and the last column are called *marginal distributions*, and the inner two rows and columns are called *conditional distributions* (for reasons explained below).

Now we can answer the question. The proportion of men preferring dark covers is $\frac{38}{50} = 76\%$, and the proportion of women preferring dark covers is $\frac{56}{150} \div 37\%$, so the survey definitely says that men prefer dark covers more than women do. (Because the survey was biased towards women, the proportion of people surveyed preferring dark covers is $\frac{94}{200} = 47\%$, which is not the mean of 76% and 37%).

Similarly, the proportion of men preferring coloured covers is $\frac{12}{50} = 24\%$, and the proportion of women preferring coloured colours is $\frac{94}{150} \div 63\%$ (and the proportion of people surveyed preferring coloured covers is $\frac{106}{200} = 53\%$).

Conditional probability in two-way tables

The percentages in the worked example above are probabilities. If we choose a person from the survey at random,

$$P(\text{person prefers dark covers}) = \frac{94}{200} = 0.47.$$

Section 12G in the Year 11 book introduced *conditional probability*. To find the probability that the person prefers dark covers given that the person is a man (or a woman), we use the *reduced sample space* of 50 men (or 150 women),

$$P(\text{prefers dark} \mid \text{man}) = \frac{38}{50} = 0.76$$
$$P(\text{prefers dark} \mid \text{woman}) = \frac{56}{150} \div 0.37.$$

We can also find conditional probabilities the other way around. For a person chosen at random from the survey,

$$P(\text{person is a man}) = \frac{50}{200} = 0.25$$
$$P(\text{man} \mid \text{prefers dark}) = \frac{38}{94} \div 0.40$$
$$P(\text{man} \mid \text{prefers coloured}) = \frac{12}{106} \div 0.11.$$

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

4 TWO PARTICULAR DATA DISPLAYS

Histograms and polygons, and cumulative histograms and polygons, will be discussed further in the next section. In this section we have particularly looked at:

- *Pareto charts*, which consist of a frequency histogram and a cumulative frequency polygon drawn together, after the categories have been arranged in decreasing order of frequency. They are normally used with categorical data, for the purpose of displaying issues in descending order of importance.
- *Two-way tables* (or *contingency tables)*, by which we can investigate whether two variables are related, and make estimates of conditional probability.

The mode and the range

The *mode* is the most popular score, meaning the score with the greatest frequency ('mode' means 'fashion'). It is the simplest measure of location to identify because it is immediately obvious from the frequency table. It is even more obvious from the resulting histogram.

For example, in the frequency table of problems experienced by Secure Roofs, the mode is the problem 'Rain', with a frequency of 88. In the earlier table of spelling test scores, the mode is 6, which happens to coincide with the median, but this is not always the case.

Some frequency tables have two or more scores with the same maximum frequency, and are called *bimodal* or *trimodal* or *multimodal*.

The *range* is only defined for numeric data. It is the difference between the minimum and maximum scores. For example, with the 40 spelling test scores,

$$\text{minimum} = 1, \qquad \text{maximum} = 10, \qquad \text{range} = 10 - 1 = 9.$$

The range is the simplest measure of spread of a dataset.

This meaning of the word 'range' in statistics is quite different from its meaning in the language of functions, where it means the set of output values of a function.

5 MODE (A MEASURE OF LOCATION) AND RANGE (A MEASURE OF SPREAD)

- The *mode* of a dataset is the most popular score, that is, the score with the greatest frequency. A dataset may be *bimodal*, *trimodal* or multimodal.
- The *range* of a dataset is the difference between the minimum and the maximum scores.
- The mode is a measure of location, and the range is a measure of spread.

Exercise 15A FOUNDATION

1 State whether each random variable is numeric or categorical. If it is numeric, state whether it is discrete or continuous. Comments may be appropriate.

a The favourite day of the week for a person chosen at random in Australia

b Height of Australian professional basketball players

c Age

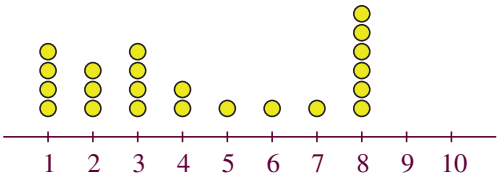
d Political affiliation

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

e Colour of a counter drawn from a cup containing 5 red and 6 blue counters.
f Sex (male or female) of a child attending a particular primary school
g Sum of the numbers when two dice are thrown
h Shoe size
i Examination scores

2 Find the median, mode and range of each dataset.
a 10 13 14 14 15 17 18
b 5 7 8 9 10 12 13 15 17
c 3 3 4 5 7 9 10 12 13 15
d 4 4 4 6 6 6 7 7 8 8 9 10
e 4 2 6 4 7 3 4 6 3 4 3 5 2 1 5 7 8
f 2 9 7 6 4 3 2 7 8 9 10 5 4 2 3 6 9 3

3 A shop sells individual cupcakes and keeps a record of how many cupcakes each customer purchases. The results are shown in the dot plot to the right.



a Construct a cumulative frequency table from the data.
b Find the median sales of cupcakes.
c Find the mode of the data.
d The shop intends to pre-package cupcakes to streamline its sales for many customers. Discuss the advantage of selling the cupcakes in packages of **i** 3, **ii** 4, **iii** 8.

4 A basketball coach begins each training session by challenging his star player to shoot as many hoops as he can in two minutes. Over twenty-one sessions he records these results:

4 3 5 4 5 4 6 7 5 6 8 6 6 8 9 10 7 8 9 7 9

a Construct a dot plot of the data. (This can be used as an alternative to a tally.)
b Copy and fill in the following frequency table.

score x	3	4	5	6	7	8	9	10
frequency f								
cumulative								

c What is the median number of hoops shot by the star player?
d In his twenty-second session he shoots 11 hoops in the 2 minutes allowed. What is his new median score?
e Is this frequency table a helpful way of displaying the scores?

5 An operator tracks the number of customers who pay the $4 fee to take his amusement ride on each day of the week. His data is displayed in the following table.

Mon	Tue	Wed	Thu	Fri	Sat	Sun
13	32	35	38	57	75	65

a Draw a bar chart showing the data, with days of the week on the horizontal axis. Use a scale of 1 cm per 10 rides.
b Construct a table of cumulative frequencies, and draw a cumulative bar chart showing the number of rides sold up to and including that day of the week.

ISBN 978-1-108-76630-2

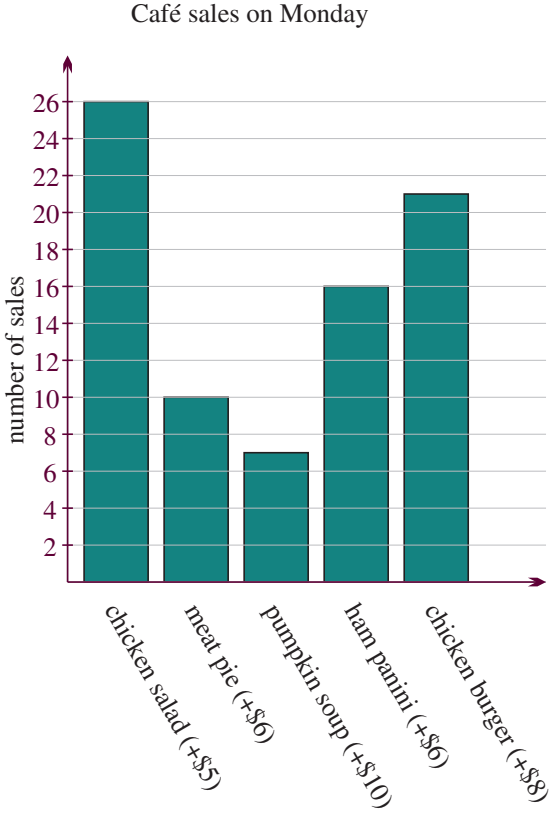
6 A survey of 1000 people in the Netherlands generated the following data shown in a contingency table relating hair and eye colour.

Colour	Blond hair	Red hair	Brown hair	Black hair	Total
Brown eyes	78	4	65	25	172
Blue eyes	324	10	46	9	389
Grey eyes	252	8	47	10	317
Green eyes	74	3	35	10	122
Total	728	25	193	54	1000

a What is the most common hair and eye colour combination in the study?
b What was the least common combination?
c What is the probability that a blond-haired person also has blue eyes?
d What is the probability that a black-haired person also has blue eyes?
e What percentage of people with black hair had brown eyes?
f What percentage of people with dark hair (brown or black) had brown eyes?
g What percentage of people with light hair (blond or red) had lighter coloured eyes (blue, grey or green)?
h Does there appear to be a link between hair colour and eye colour?
i This study was carried out from a particular genetic population. Is it likely that similar results hold everywhere?

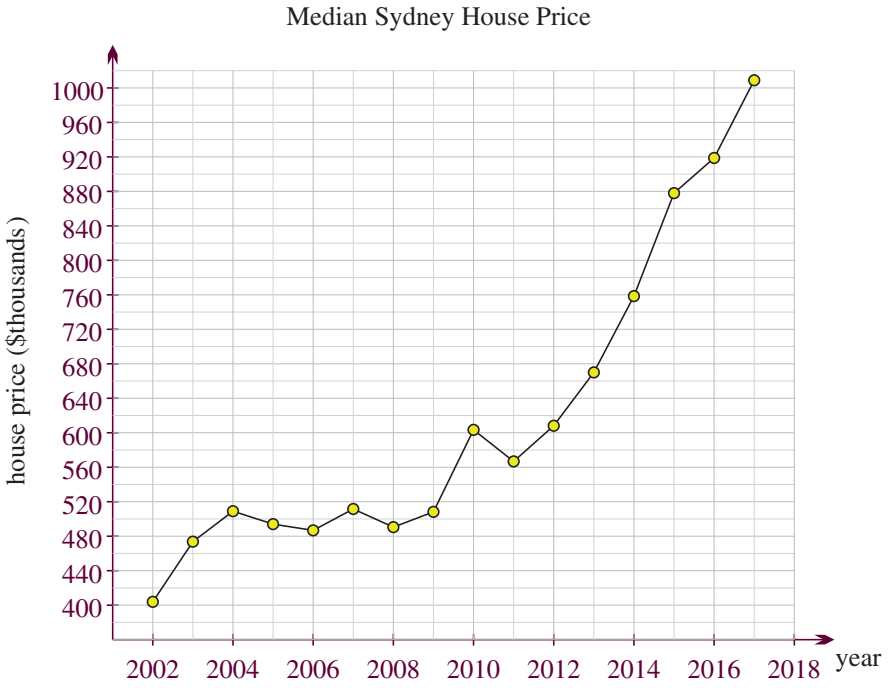
7 A café tracks its sales on a certain Monday to find what menu items are selling. It has a limited menu: chicken salad, meat pie, pumpkin soup, ham panini, and chicken burger. The café's results are shown on the graph to the right. The graph also records the markup (profit) on each choice, shown as (+$ markup).

a What is the total number of menu orders for the café on Monday?
b Determine what percentage the sale of each menu option is of the total.
c What is the profit, in dollars, obtained from each of the choices on the menu on the Monday?
d What is the total profit, in dollars, for the café on the day?
e The café has a policy to drop from the menu any choice with sales below 10%. Give two reasons why they should not drop the pumpkin soup from the menu.



ISBN 978-1-108-76630-2

8 The median house price in Sydney from 2002 to 2017 is recorded in the line graph.



a Write down the median price of a Sydney house in 2002 and 2017, correct to the nearest ten thousand dollars.

b What is the percentage increase in house prices in Sydney from 2002 to 2017?

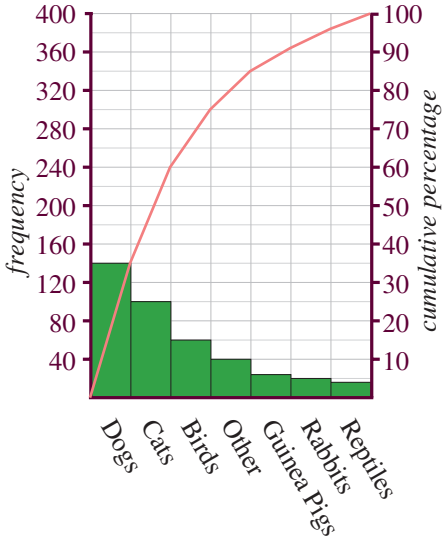
c What was the average increase in house price per year over this time?

d If this trend continues, what do you predict the median house price will be in 2030?

e What year saw the greatest increase in house prices?

f What year saw the greatest decrease in house prices? How much did prices change?

9 Owners of *The Happy Pet* boarding house for pets whose owners are out of town are looking to expand their business. A business analyst has asked them to keep track of the last 400 pets staying at their boarding house to determine what kind of pets they will need to accommodate in their planned expansion. This information is displayed in the Pareto chart to the right.



a What percentage of the last 400 pets at the boarding house were dogs? How many dogs was this?

b Rabbits and guinea pigs can stay in the same type of cage. What percentage of the last 400 pets staying were rabbits or guinea pigs?

c To maximise business profits, the owners decide to concentrate on the three most common pets. What percentage of the last 400 pets were one of these three?

d What percentage of pets fell into one of the three least common categories?

e Comment on the size of the 'others' category.

f What other matters should the owners take into account, besides the numbers of pets looking for boarding?

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

DEVELOPMENT

10 A school is investigating reasons why students arrive late to class. Students who are late are asked to state a reason. The reasons given by the last 100 students are recorded in the table to the right.

Reasons late to class	Frequency
Didn't hear the bell	20
Held back in last class	27
Cancelled music lesson	10
Lost bag	5
Late back from lunch	3
Summons by senior teacher	2
Medical	3
Distance from last class	20
Other	10

a Construct a table with the categories ordered by decreasing frequency. Add a cumulative frequency column, which will also be the cumulative frequency percentage because there were 100 students in the survey.

b Construct a Pareto chart of the data. Use a scale of 1 cm per 10 units on each vertical scale.

c Explain why the cumulative frequency polygon of a Pareto chart will always be concave down.

d What percentage of the reasons are included in the first three categories?

e Comment on how the school could work to reduce the first three causes of tardiness.

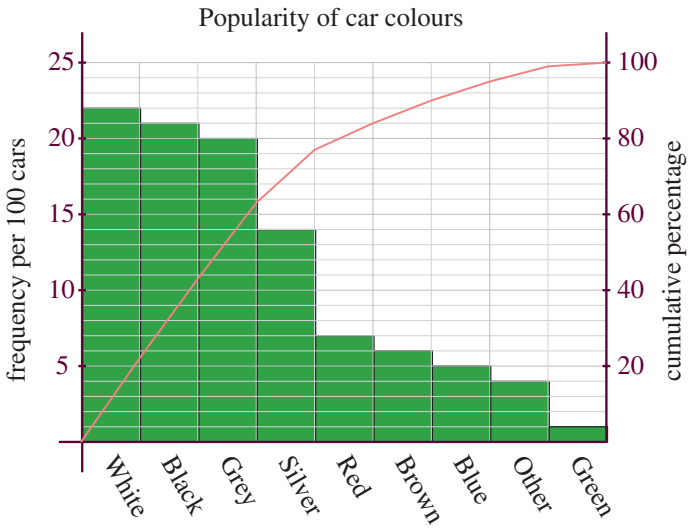
11 The colours of cars on the road are recorded in the following Pareto chart.

a What percentage of cars are brown?

b What percentage of cars are of one of the three most common colours, white, black or grey?

c How many cars are not one of the seven most popular colours?

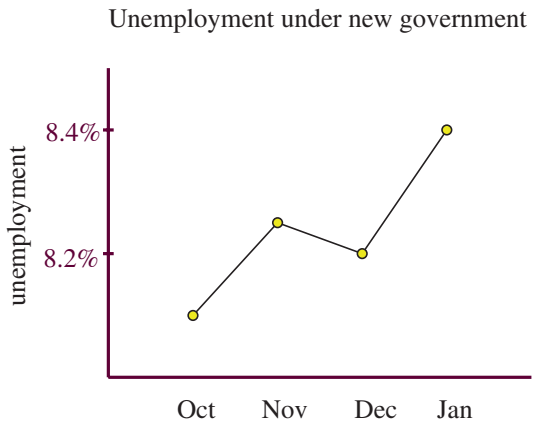
d This Pareto chart uses different scales on the two axes — is this confusing?



ISBN 978-1-108-76630-2

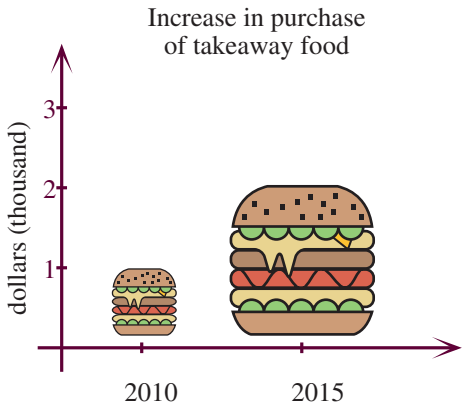
12 Statistics can easily be misinterpreted or deliberately used to mislead.

a



The graph copied above was published in a newspaper. Can you suggest why it might be a misleading use of statistics and graphing?

b

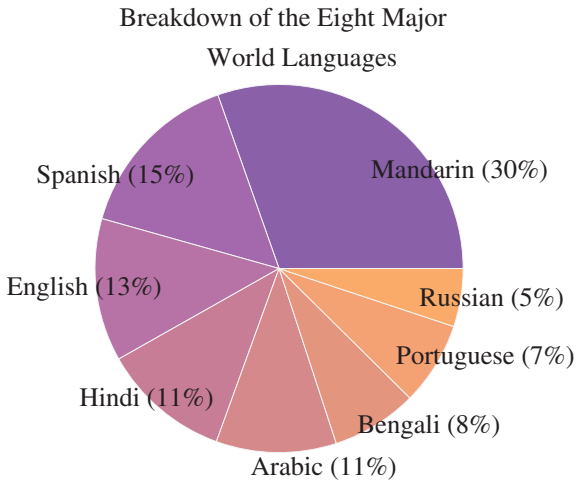


A study has been commissioned into the consumption of fast food. The graph to the right shows one of the results of this study. Discuss why this graph may be misleading if used in a local newspaper or television advertisement.

c A survey was designed to collect data to investigate the following scenarios. Comment on problems with the design of the experiment.

- **i** A study is carried out to determine people's musical tastes. People are asked to fill in an online survey, and the results are then collated.
- **ii** To investigate the growth in medical costs, a community group accesses data from their local hospital. The growth in total expenses at the hospital over time is displayed in a line graph.

13 In 2019, around 40% of the world's 7.7 billion population were first-language speakers of one of eight different languages. The sector chart to the right shows the breakdown of first-language speakers as a percentage of the top eight.



a What percentage of the 40% speak one of the three most common languages as a first language?

b How many people speak one of these eight as their first language?

c How many people in the world speak Mandarin as their first language?

d Around what percentage of the world's population speak English as their first language?

e Is this chart useful to a school deciding what languages to offer or a student deciding what language to learn?

14 The mean temperature for each month in Sydney (Observatory Hill) and Dubbo (airport) at 3pm is recorded in degrees Celsius on a radial chart.

For example, to read the mean temperatures for February, look at the radius marked 'Feb' — the mean temperature in Dubbo is about 30°C, and in Sydney it is about 25°C. The mean temperatures in January are on the adjacent radius, and each pair of values are joined by an unbroken interval for Sydney, and a broken interval for Dubbo.

ISBN 978-1-108-76630-2

Using the 3pm temperature as a measure of the temperature in Sydney and Dubbo:

a What is the 3pm temperature in Dubbo in **i** July, **ii** December?

b What is the 3pm temperature in Sydney in **i** August, **ii** March?

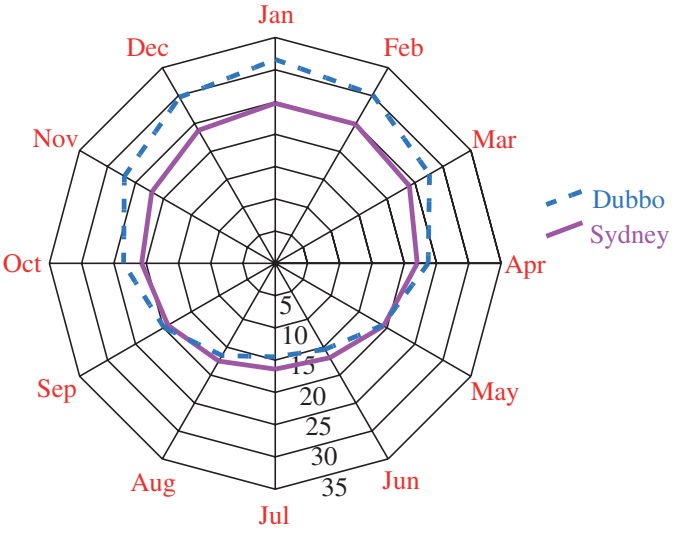
c What is the maximum 3pm mean temperature difference between the two locations, and in what month does it occur?

d In which months is the mean 3pm temperature in Sydney and Dubbo the same?

e In which months is Dubbo at least 5 degrees hotter than Sydney?

f Are there any months where Dubbo is colder than Sydney?

Sydney-Dubbo Mean 3pm Temperature



g Is this a good style of chart to display the data? Would there be a better type of chart to use?

h Why do you think that the designer of this chart chose to use dotted and solid lines, rather than just the cyan and magenta colours, to distinguish the two temperature lines?

15 A small class is working with pupils who have difficulty with English. After students had been in the class for some time, their examination results were recorded in the stacked bar chart below. The examination consisted of three sections: an interpretative exercise, an essay on a novel studied in class, and a creative writing exercise. Each section was awarded a mark out of 10.

a What was the examination out of in total?

b What were the highest and lowest scores, as a percentage?

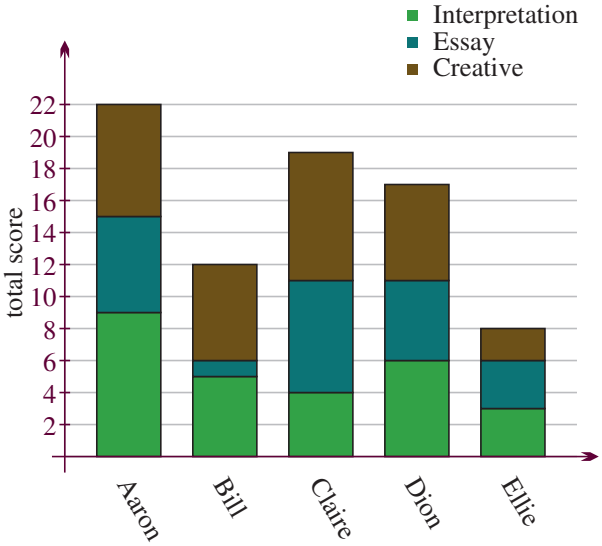
c Identify any of the three sections for which a pupil may need additional help.

d What percentage did Claire receive in Interpretation?

e What percentage did Dion receive in Essay?

f Students who receive 55% overall and at least 50% in each section leave this class. Who will be leaving the class following this examination?

English Examination Results



Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

ENRICHMENT

16 The following two-way table shows the highest non-school qualification received by Australians aged between 15–64, with a break-down by age. The entries are percentages.

	15–24	25–34	35–44	45–54	55–64	Total 15–64
Postgraduate	0.4	6.4	6.2	5.3	4.7	4.6
Graduate diploma/cert	0.1	1.9	2.5	3.2	3.1	2.1
Bachelor degree	7.4	26.8	21.2	15.1	13.3	17.0
Adv diploma/diploma	4.2	9.6	11.1	11.7	9.1	9.1
Certificate I-IV	14.0	21.7	24.1	23.3	22.9	21.1
Other	2.1	3.0	2.7	2.5	2.8	2.6
None	71.8	30.6	32.2	38.9	44.1	43.5
Total (percent)	100%	100%	100%	100%	100%	100%
Total (thousands)	3122.5	3215.9	3118.1	2969.3	2422.3	14848.1

a *There is a greater percentage of people in the 25–34 age group with postgraduate degrees, compared with the 55–64 aged group. This suggests that more people are gaining postgraduate degrees in more recent times.* Comment on whether this is a reasonable interpretation of the data.

b What it the probability that an Australian chosen at random from the 45–54 age group has a post-school qualification?

c What it the probability that an Australian aged 15–64 chosen at random from the group with a post-school qualification lies in the age group 45–54?

ISBN 978-1-108-76630-2

15B Grouped data and histograms

The main purpose of organising data into tables and graphs is to see the data as a whole. When a table has too many rows or columns, or a graph has too much detail, such an overview is much more difficult. The usual approach in such situations is to *group* the data. This reduces the number of rows or columns on the tables, and reduces the amount of clutter on the graphs.

Grouping data

Here are the heights of 100 people in centimetres, from a file detailing individuals of all ages from the !Kung people of the Kalahari desert. The underlying random variable here is continuous (assuming that heights are real numbers), but height cannot be measured correct to more than a few significant figures.

151.765	139.7	136.525	156.845	145.415	163.83	149.225	168.91	147.955	165.1
154.305	151.13	144.78	149.9	150.495	163.195	157.48	121.92	105.41	86.36
161.29	156.21	129.54	109.22	146.4	148.59	147.32	137.16	125.73	114.3
147.955	161.925	146.05	146.05	142.875	142.875	147.955	160.655	151.765	171.45
147.32	147.955	144.78	121.92	128.905	97.79	154.305	143.51	146.7	157.48
127	110.49	97.79	165.735	152.4	141.605	158.8	155.575	164.465	151.765
161.29	154.305	145.415	145.415	152.4	163.83	144.145	129.54	129.54	153.67
142.875	146.05	167.005	91.44	165.735	149.86	147.955	137.795	154.94	161.925
147.955	113.665	159.385	148.59	136.525	158.115	144.78	156.845	179.07	118.745
170.18	146.05	147.32	113.03	162.56	133.985	152.4	160.02	149.86	142.875

The data seem to be given correct to 0.005 cm, which seems less than one can reliably measure, and the trailing zeroes that we normally insert are missing — always question the credibility of raw data. Perhaps heights were recorded in inches, then converted to centimetres. We have grouped the data in 10 cm intervals because that results in 10 classes, which is a good number for seeing the big picture. Here is the table of frequencies and cumulative frequencies.

interval	class centre	frequency	cumulative frequency
80–90	85	1	1
90–100	95	3	4
100–110	105	2	6
110–120	115	5	11
120–130	125	8	19
130–140	135	6	25
140–150	145	34	59
150–160	155	22	81
160–170	165	16	97
170–180	175	3	100

ISBN 978-1-108-76630-2

The *class centre* on each row is the midpoint of the interval used in the grouping. The table could just as well have been written with rows instead of columns, and rows have been used later in the calculation of the mean and variance.

This makes the distribution of heights reasonably clear. A frequency distribution table based on the raw data, however, would be practically useless, because the frequency of almost every score is just 1 (and the histograms drawn below would also be useless).

Grouping data is a form of rounding. It is useful because it allows us to see the big picture, but it always involves ignoring information, and the summary statistics for the grouped data will only be an approximation of the summary statistics of the raw data. Never discard the original data.

For example, the median height is the average of the 50th and 51st heights. For the grouped data, both these heights are 145 cm (taking the class centre as the measurement), whereas if we work with the original data, the median is 147.955. Similarly, the range of the grouped data is $175 - 85 = 90$, but the range of the raw data is $179.07 - 86.36 = 92.71$.

When the underlying random variable is continuous, any data are already grouped by the rounding that all measurement involves. When those measurements involve several significant figures, as they do here, further grouping is usually required when displaying the data.

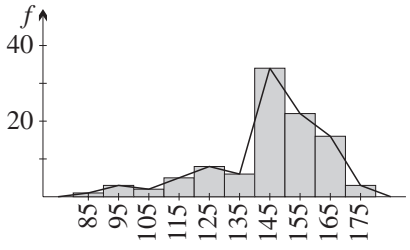
6 GROUPING DATA

- Numeric data, whether discrete or continuous, may be *grouped* so that the resulting tables and graphs give a clearer overview of the data.
- The grouping involves *intervals of equal width* and *class centres*.
- Grouping involves ignoring information. This may or may not be an issue.
- Data on a boundary should be treated consistently, and the treatment noted.

With a continuous variable, there may be data on the boundary, because data are always rounded. You can place these scores in the lower interval — this is consistent with the cumulative frequency convention in Section 15A. Or you can place them in the upper interval — this is also standard practice. But be consistent, and make a note about it if any boundary data actually occurred. This didn't happen with the data above.

Frequency histograms and frequency polygons

Whether or not the data have been grouped, a *frequency histogram* is the most basic way of displaying data in a graph. The diagram to the right shows the histogram of the grouped heights in the previous frequency table.



The *frequency polygon* has been added to the display. The two graphs can be drawn separately or together, and only one may be needed.

ISBN 978-1-108-76630-2

Some guidelines when drawing a frequency histogram

- For ungrouped data, each rectangle is centred on the value. For grouped data, each rectangle is centred on the class centre.
- The rectangles join up with no gaps.
- As a practical concern, too many columns in a histogram can make it difficult to interpret. Coarser grouping is the best solution here.
- The subintervals on the horizontal axis are often called *bins*.

Some guidelines when drawing a frequency polygon

- The plotted points are at the centre of the top of each rectangle.
- Join the plotted points with intervals.
- On the left, start the polygon on the horizontal axis, at the previous value or class centre.
- On the right, end the polygon on the horizontal axis, at the next value or class centre.

A question from the graph: Always ask questions about the data display.

- Why is the data so *skewed to the left*, with such low frequencies? Were children's heights included?

Histograms with discrete data

Histograms are designed for data from a continuous variable, which is the main reason why the rectangles should join up. When they are used for discrete data, be aware that the rectangles still have width, that they still join up (according to most conventions), and that they are centred on the values (or on the class centres for grouped data). These conventions will routinely involve numbers such as half-integers that are not possible values of the random variable.

Example 4 15B

Section 15A prepared a frequency table for 40 spelling test marks in Year 7.

Mark x	1	2	3	4	5	6	7	8	9	10
Frequency f	2	4	2	1	6	8	7	6	2	2

a Group the data by pairing the marks 1–2, 3–4, ...
b Draw a histogram and frequency polygon for the original data.
c Draw a histogram and frequency polygon for the grouped data.
d Comment on what the various displays have shown.

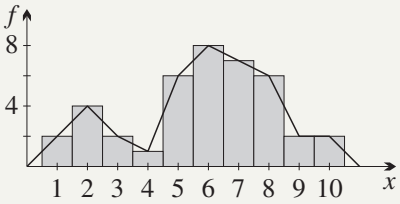
SOLUTION

a

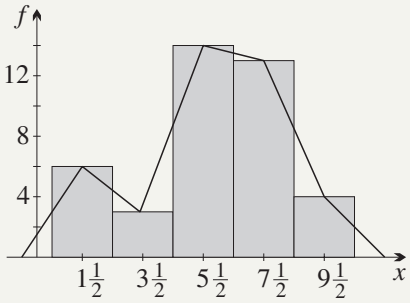
Interval	1–2	3–4	5–6	7–8	9–10
Class centre x	$1\frac{1}{2}$	$3\frac{1}{2}$	$5\frac{1}{2}$	$7\frac{1}{2}$	$9\frac{1}{2}$
Frequency f	6	3	14	13	4

ISBN 978-1-108-76630-2

b



c

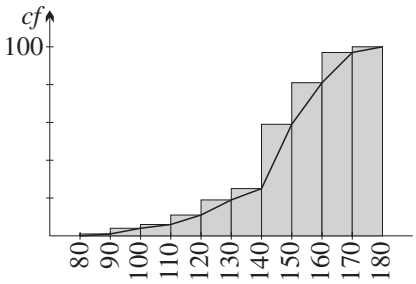


d The histogram of the grouped data perhaps makes it a little clearer that a significant group have difficulties either with spelling or with tests.

Note: The frequency polygon starts and finishes on the horizontal axis at the previous or next value or class centre. For the original data in part **b**, it starts at 0 and finishes at 11. For the grouped data, it starts at $-\frac{1}{2}$ and finishes at $11\frac{1}{2}$.

Cumulative frequency histograms and polygons (ogives)

A *cumulative frequency histogram* is drawn using the same procedures as for the earlier frequency histogram. The *cumulative frequency polygon*, also called an *ogive*, is drawn slightly differently corresponding to its cumulative nature.



We have drawn the two graphs together for the grouped table of heights at the start of this section, but again, each can be drawn separately.

Some guidelines when drawing a cumulative frequency histogram

- The rectangles of the frequency histogram are piled on top of each other to form the cumulative frequency histogram.
- The height of the last rectangle is the total size of the sample.

Some guidelines when drawing a cumulative frequency polygon

- The polygon starts at zero at the bottom left-hand corner of the first rectangle, when no scores have yet been accumulated.
- It passes through the top right-hand corner of each rectangle because it plots the scores less than or equal to the upper bound of the class interval.
- It finishes at the top right-hand corner of the last rectangle, and its height there equals the total size of the sample.

Example 5 15B

Section 15A also prepared a cumulative frequency table for the spelling test marks.

Mark x	1	2	3	4	5	6	7	8	9	10
Frequency f	2	4	2	1	6	8	7	6	2	2
Cumulative	2	6	8	9	15	23	30	36	38	40

ISBN 978-1-108-76630-2

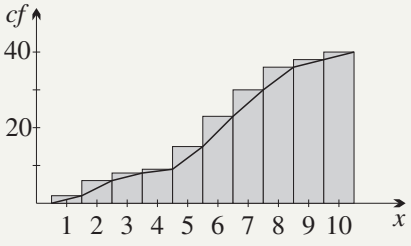
a Group the data by pairing the marks, adding the cumulative frequency.
b Draw a cumulative frequency histogram and ogive for the data.
c Draw a cumulative frequency histogram and ogive for the grouped data.
d Use the cumulative frequency tables to calculate the median using the original data and the grouped data, and compare them.

SOLUTION

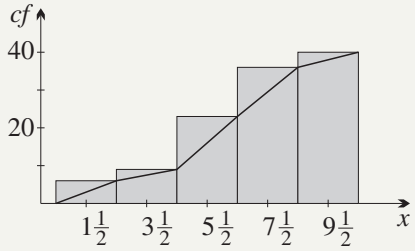
a

Interval	1–2	3–4	5–6	7–8	9–10
Class centre x	$1\frac{1}{2}$	$3\frac{1}{2}$	$5\frac{1}{2}$	$7\frac{1}{2}$	$9\frac{1}{2}$
Frequency f	6	3	14	13	4
Cumulative	6	9	23	36	40

b



c



d We calculated before that the median was the average of the 20th and 21st scores, which the cumulative frequencies tell us are both 6, so the median is 6.

From the grouped data, the class centres of the 20th and 21st scores are both $5\frac{1}{2}$, so the median is $5\frac{1}{2}$. Such discrepancies are normal after grouping.

7 HISTOGRAMS AND POLYGONS

- The rectangles of the *frequency histogram* and the *cumulative frequency histogram* join up. For ungrouped data, they are centred on the value, and for grouped data, they are centred on the class centre.
- The *frequency polygon* passes through the centres of the rectangles.
- On the left and right, the frequency polygon starts and finishes on the horizontal axis, centred on the previous or next value or class interval.
- The *cumulative frequency polygon* or *ogive* starts at the bottom left corner of the first rectangle.
- The ogive passes through the right-hand top corner of each rectangle.

The mean

In Year 11, we calculated the mean and variance of a sample in Section 13D, and we used relative frequencies in those calculations because our attention was on estimating probabilities from relative frequencies. This chapter, however, is about data, so in reviewing mean and variance, we will instead use formulae based only on the frequencies. The review also takes grouped data into account.

ISBN 978-1-108-76630-2

Recall that we use the symbols $\bar{x}$ and s for the mean and standard deviation when we are dealing with a sample, that is, with data, as we are in this chapter. When dealing with a population or a theoretical distribution, we use the symbol μ or $\mathrm{E}(X)$ for the mean, and σ for the standard deviation.

Here again is the grouped frequency table of heights, arranged this time in rows, but it could as well be in columns.

class centre x	85	95	105	115	125	135	145	155	165	175	Sum
frequency f	1	3	2	5	8	6	34	22	16	3	100
$x \times f$	85	285	210	575	1000	810	4930	3410	2640	525	14470

The sum of the scores is the sum of the products $x \times f =$ score $\times$ frequency, except that with grouped data we use the class centres. This sum is 14470. The number of scores is the sum of the frequencies, which is 100. Hence

$$\text{mean} = \frac{\sum xf}{n} = \frac{14470}{100} = 144.70 \text{ cm.} \qquad \left(\sum xf \text{ means add all the products } xf.\right)$$

The authors also calculated the mean of the heights without grouping and obtained 145.352 cm. Thus the grouping reduced the mean by about $6\frac{1}{2}$ mm.

In Section 13D (Year 11) we wrote the mean as a weighted mean of the scores, weighting each score by its relative frequency using the formula $\bar{x} = \sum xf_r$, where f_r is the relative frequency. The relative frequencies are obtained by dividing each frequency by the number n of scores, that is, $f_r = \frac{f}{n}$, so the formula used above and the earlier formula are the same. To prove this, start with the earlier formula,

$$\bar{x} = \sum xf_r = \sum \left(x \times \frac{f}{n}\right) = \frac{\sum xf}{n}.$$

The variance and standard deviation

As explained in the Year 11 book, the variance and standard deviation are measures of spread, meaning that they measure how spread out the data are away from the centre. The standard deviation is the square root of the variance. Thus the variance has symbol s^2 for a sample and σ^2 or $\mathrm{Var}(X)$ for a population or a theoretical distribution. In Section 13D of the Year 11 book, we developed two formulae for the variance of data,

$$s^2 = \sum (x - \bar{x})^2 f_r \qquad \text{and} \qquad s^2 = \sum x^2 f_r - \bar{x}^2.$$

Substituting $f_r = \frac{f}{n}$ for the relative frequency, these two formulae become

$$s^2 = \sum (x - \bar{x})^2 f_r \qquad \text{and} \qquad s^2 = \sum x^2 f_r - \bar{x}^2$$
$$= \frac{\sum (x - \bar{x})^2 f}{n} \qquad\qquad = \frac{\sum x^2 f}{n} - \bar{x}^2.$$

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

It is rare with data that the mean $\bar{x}$ is a round number, so the second form is the recommended form for calculation. The first form, however, makes it clear that the variance is a measure of spread — we are looking at deviations from the mean, squaring them so that they are all positive, then taking their weighted mean, weighted according to the frequencies.

8 MEAN AND STANDARD DEVIATION OF A SAMPLE

Suppose that data have been organised into a frequency table with scores x and frequencies f.

- The mean is $\bar{x} = \dfrac{\sum xf}{n}$ (where $\sum$ says take the sum over the distribution).
- The variance is $s^2 = \dfrac{\sum x^2 f}{n} - \bar{x}^2$.
- The standard deviation is the square root of the variance, and has the same units as the scores.
- With grouped data, use the class centres rather than the scores.

The actual definition of the variance is $s^2 = \dfrac{\sum (x - \bar{x})^2 f}{n}$. This formula is usually less suitable for calculation, but it makes it clear that we are taking the average of the squares of the deviations from the mean.

Here are the calculations for the variance and standard deviations of the heights.

x	85	95	105	115	125	135	145	155	165	175	Sum
f	1	3	2	5	8	6	34	22	16	3	100
xf	85	285	210	575	1000	810	4930	3410	2640	525	14470
x^2f	7225	27075	22050	66125	125000	109350	714850	528550	435600	91875	2127700

Thus $\bar{x} = \dfrac{\sum xf}{n}$

$= \dfrac{14\,470}{100}$

$= 144.7$ cm,

and $s^2 = \dfrac{\sum x^2 f}{n} - \bar{x}^2$

$= \dfrac{2\,127\,700}{100} - 144.7^2$

$= 338.91,$

$s \doteqdot 18.41$ cm.

The authors used technology with the raw scores, and obtained $s^2 = 325.5407$ and $s \doteqdot 18.04$ cm. Grouping has produced results that are slightly different.

ISBN 978-1-108-76630-2

Example 6 15B

a Find the mean, variance and standard deviation of the Year 7 spelling test marks in worked Example 4.
b Calculate them again using the grouped data.

SOLUTION

a

x	1	2	3	4	5	6	7	8	9	10	Sum
f	2	4	2	1	6	8	7	6	2	2	40
$x \times f$	2	8	6	4	30	48	49	48	18	20	233
$x^2 \times f$	2	16	18	16	150	288	343	384	162	200	1579

Hence $\bar{x} = \dfrac{\sum xf}{n}$

$= \dfrac{233}{40}$

$= 5.825,$

and $s^2 = \dfrac{\sum x^2 f}{n} - \bar{x}^2$

$= \dfrac{1579}{40} - 5.825^2$

$= 5.544375,$

$s \doteqdot 2.355.$

b

Interval	1–2	3–4	5–6	7–8	9–10	Sum
Class centre x	1.5	3.5	5.5	7.5	9.5	—
Frequency f	6	3	14	13	4	40
$x \times f$	9	10.5	77	97.5	38	232
$x^2 \times f$	13.5	36.75	423.5	731.25	361	1566

Hence $\bar{x} = \dfrac{\sum xf}{n}$

$= \dfrac{232}{40}$

$= 5.8,$

and $s^2 = \dfrac{\sum x^2 f}{n} - \bar{x}^2$

$= \dfrac{1566}{40} - 5.8^2$

$= 5.51,$

$s \doteqdot 2.347.$

Again, the differences between the results of parts **a** and **b** arise from grouping.

A correction factor for the sample variance

These qualifications may not be required. We mentioned at the end of Section 13D in the Year 11 book that when we know the theoretical or the population mean μ, and we are sampling to find the variance, there is no problem with the formulae for the sample variance. When, however, we are sampling both to find the mean and to find the variance, then the sample mean will drift very slightly towards the sample results, with the effect that the sample variance will tend to be slightly smaller than it should be.

The standard solution is to multiply the sample variance by a correction factor $\dfrac{n}{n-1}$, where n is the size of the sample. Thus in the example using 100 heights, we were using a sample mean rather than a theoretical or population mean, so the correction factor is $\frac{100}{99}$. Using the correction factor would yield

ISBN 978-1-108-76630-2

$\bar{x} = 144.7$ cm (as before) $\quad s^2 = 338.91 \times \frac{100}{99} \div 342.3333 \quad s = \sqrt{s^2} \div 18.50$ cm.

The larger the size n of the sample, the less difference the correction makes.

On the calculator, a button $\boxed{\sigma_n}$ or something equivalent does not apply this correction, and seems to be all that is required in this course. A button labelled $\boxed{\sigma_{n-1}}$ or equivalent applies the correction factor.

Currently in Excel 365, the function `STDEV.S` (S for 'sample') applies the correction, and the function `STDEV.P` (P for 'population') does not apply the correction, but earlier versions did things differently.

Like so many things in statistics, the distinction between a sample and a population is not straightforward. The Year 7 spelling test marks may be regarded as the record of spelling on that day from every member of the cohort — then the marks are a population. They may also be regarded as one set of estimates of an underlying random variable, 'spelling ability of each child', to be augmented by next week's spelling test and more in later weeks — then the marks are a sample. Scaling software, which massages results into aggregates and positions and ranks, tends to regard marks as a population. A classroom teacher, who is watching students learn and develop and is very aware that the choice and design of test questions are arbitrary, tends to regard marks as a sample.

A possible project

Systematic testing of the validity of this correction factor by taking a large number of samples from a known population could be developed into a project.

Perhaps theoretical probability distributions could also be considered, perhaps both discrete and continuous (as developed in Chapter 16).

Exercise 15B

FOUNDATION

1 a Copy and complete the following table to determine the mean and standard deviation for the data. (The mean is a whole number, so calculations using this formula will be straightforward.)

x	3	5	6	7	8	9	10	Sum
f	1	1	1	3	2	1	1	
$x \times f$								
$(x - \bar{x})^2$								
$(x - \bar{x})^2 f$								

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

$$\text{Variance} = s^2 = \frac{\sum (x - \bar{x})^2 f}{n} = \ldots$$

b Repeat the calculation using the alternative formula for the variance.

x	3	5	6	7	8	9	10	Sum
f	1	1	1	3	2	1	1	
$x \times f$								
$x^2 \times f$								

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

$$\text{Variance} = s^2 = \frac{\sum x^2 f}{n} - \bar{x}^2 = \ldots$$

ISBN 978-1-108-76630-2

2 Use a table as in Question 1 to calculate manually the mean and standard deviation of the following datasets. Use the two forms for the variance in different parts — the means here are all whole numbers. Give your answers correct to two decimal places.

a 12, 14, 16, 17, 19, 21, 22, 23

b 2, 3, 3, 3, 6, 6, 7, 8, 8, 8, 9, 9, 10, 10, 13

c 40, 49, 50, 50, 51, 54, 57, 57, 57, 60, 65, 70

d 7, 8, 9, 9, 10, 10, 10, 10, 11, 11, 11, 11, 11, 12, 12, 12, 12, 13, 13, 14, 15

3 Use your calculator to find the mean and standard deviation of each dataset.

a 3, 7, 9, 10, 3, 4, 6, 8, 13, 6, 5, 12

b 4, 4, 4, 4, 5, 5, 7, 8, 8, 8

c 3.2, 3.6, 1.3, 2.4, 1.9, 4.1, 3.5, 4.1, 3.9, 2.3

d 34, 45, 23, 56, 34, 53, 23, 43, 37, 55, 52, 41, 43, 51, 57, 39

DEVELOPMENT

4 A census was carried out on the houses in Short Street to determine the number x of people in each household. The results are recorded in the frequency table below.

The population in this question is all the houses in the street. Because the data here are determined by a census of the whole population, statisticians use the symbol μ for the mean of the population (called the *population mean*) and the symbol σ for the standard deviation of the population (called the *population standard deviation*).

x	0	1	2	3	4	5	6	7	8
f	1	5	6	7	8	3	3	0	1

a How many houses are there in Short Street?

b Calculate the mean μ and standard deviation σ of the data.

c Group the data into the classes 0–2, 3–5 and 6–8 and construct a grouped frequency table.

d Calculate the mean and standard deviation of this grouped data.

e Why do your results from part **b** and **d** differ?

5 Xiomi recorded her time to get to work each day. Her results in minutes were:

22 30 23.5 27 25 21.5 39 30 32.5 33 35.5 37 42 22 23.5 27 29.5 23 34

a Write the data out in order, and determine the median.

b Group the data into classes by completing the following table.

class	20–24	24–28	28–32	32–36	36–40	40–44
class centre						
frequency						
cumulative						

c Find the median of the grouped data. Does it agree with your answer to part **a**?

d Draw a frequency histogram and polygon on the same chart.

e Draw a cumulative frequency histogram and polygon on the same chart.

ISBN 978-1-108-76630-2

6 In a class experiment, students measured their heights. The results in centimetres were:

155 152 165 162 170 168 165 162 166 154 158 159
163 166 164 164 159 157 163 154 166 158 159 163

a Display the data in a frequency table.

b Calculate the median of the dataset.

c Why would it not be helpful to graph the data without first grouping it into classes?

d Group the data into the intervals 150–154, 154–158, 158–162, 162–166, 166–170 and display your results in a grouped frequency table. Include any scores on a boundaries in the lower group, thus x is in the group 150–154 if $150 < x \leq 154$.

e Calculate the median of the dataset from this grouped frequency table.

f Display your grouped data on a histogram with a frequency polygon joining the centres. Construct a cumulative frequency histogram and ogive — remember that the ogive starts at the bottom left-corner of the first rectangle and passes through the right-hand top corner of each rectangle.

g Trace the line at frequency 12 (50%) until it meets the ogive, and check whether this agrees with your answer for the median of the grouped data in part **e**.

h Construct a cumulative frequency histogram and ogive of the *ungrouped* data.

i Compare your grouped and ungrouped cumulative histograms in parts **f** and **h**. How similar are the graphs? Contrast the differences between the histogram of the grouped data in part **f** and what you would expect the histogram of the ungrouped data to look like (not drawn).

j Confirm that the line at frequency 12 meets your ungrouped ogive to give the same median as in part **g**.

ENRICHMENT

7 The Australian Bureau of Statistics (ABS) surveys important medical and physical information for the Australian population. According to their 1995 survey, the mean weight of a male over 18 was 82 kg, with a standard deviation of 13.6 kg.

The data were gathered from a sample of the whole population, but the quoted standard deviation has been calculated using the population standard deviation formula. Thus the *population variance* should be multiplied by the *correction factor* $\frac{n}{n-1}$, as discussed in the notes at the end of this section, to give the sample variance. This corrects for the drift of the calculated variance towards the sample results and away from the true population standard deviation.

a What would be the corrected sample standard deviation, assuming that the sample only surveyed:

i 10 people, **ii** 100 people, **iii** 1000 people?

b Actually, the ABS survey involved 10000 people. What percentage change would the correction factor make to the standard deviation? Give your answer correct to three decimal places.

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

15C Quartiles and interquartile range

For numeric data, the spread of the data around the median can also be identified by the quartiles and the interquartile range.

Upper and lower quartiles

Write the scores in increasing order. The lower quartile, the median and the upper quartile attempt to divide this list into four equal parts.

An odd number of scores: Reliable Appliances sell toasters. Here are the numbers of toasters that they sold in each of 15 successive weeks.

19 16 18 15 16 19 17 21 16 16 20 18 30 19 21

First we write them out in a list in increasing order,

15 16 16 16 16 17 18 18 19 19 19 20 21 21 30

↑

The number of scores in the list is 15, which is odd. The median Q_2 is the 8th score 18. Now divide the list into two sublists of 7, with the median in the middle,

15 16 16 16 16 17 18 18 19 19 19 20 21 21 30

↑ ↑

The lower quartile Q_1 is the median of the left-hand list, which is 16, and the upper quartile Q_3 is the median of the right-hand list, which is 20. In summary:

$Q_1 = 16$ and $Q_2 = 18$ and $Q_3 = 20$.

An even number of scores: On the 16th week they sold 21 toasters, making 16 scores, which is even. The list can now be written out in two equal sublists,

15 16 16 16 16 17 18 18 19 19 19 20 21 21 21 30

↑ ↑ ↑

The median Q_2 is the average of the 8th and 9th scores, which is $18\frac{1}{2}$. The lower quartile Q_1 is the median of the left-hand list, which is 16, and the upper quartile Q_3 is the median of the right-hand list, which is $20\frac{1}{2}$. In summary:

$Q_1 = 16$ and $Q_2 = 18\frac{1}{2}$ and $Q_3 = 20\frac{1}{2}$.

Interquartile range: The difference $Q_3 - Q_1$ between the upper and lower quartiles is called the *interquartile range* or IQR. It is the range of the middle 50% of the marks. Thus in the two examples above, the interquartile ranges are

$\text{IQR} = Q_3 - Q_1 = 20 - 16 = 4$ and $\text{IQR} = Q_3 - Q_1 = 20\frac{1}{2} - 16 = 4\frac{1}{2}$.

9 QUARTILES AND INTERQUARTILE RANGE

Suppose that a set of scores is arranged in increasing order.

An odd number of scores:

- Omit the middle score, thus separating the list into two sublists of equal size.
- The *lower* or *first quartile* Q_1 is the median of the left-hand list,
- The *upper* or *third quartile* Q_3 is the median of the right-hand list.

An even number of scores:

- Separate the list into two sublists of equal size.
- The *lower* or *first quartile* Q_1 is the median of the left-hand list,
- The *upper* or *third quartile* Q_3 is the median of the right-hand list.

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

The interquartile range — a measure of spread:

- The *interquartile range* or IQR is the difference $Q_3 - Q_1$.
- The interquartile range is a measure of spread.

Quartiles, like medians, are easily calculated from the cumulative frequency table. They can also be calculated for theoretical distributions, as in worked Example 7.

The five-number summary

A data distribution can be usefully summarised in a *five-number summary*, which can then be displayed in a box-and-whisker plot. This summary presents the median, the quartiles, and the two extreme scores, that is,

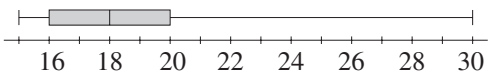
- the minimum score (sometimes written as Q_0)
- the lower quartile Q_1
- the median Q_2
- the upper quartile Q_3
- the maximum score (sometimes written as Q_4).

Thus the five-number summaries of the two sets of weekly toaster-sale scores are:
15 weekly toaster-sale scores: 15, 16, 18, 20, 30
16 weekly toaster-sale scores: 15, 16, $18\frac{1}{2}$, $20\frac{1}{2}$, 30

Notice that the range is the difference between the first and last numbers, and the interquartile range is the difference between the second and second-last numbers. The symbols Q_0 and Q_4 are convenient, but are not standard notation.

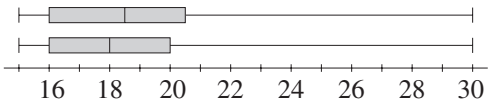
Box-and-whisker plots (box plots)

To the right is the *box-and-whisker plot* (also called *box plot*) of the 15 weekly toaster-sales scores. It displays the five-number summary in a clear diagram.



- The *box* extends from the lower quartile $Q_1 = 16$ to the upper quartile $Q_3 = 20$. Its length is the IQR.
- The vertical line within the box is the median $Q_2 = 18$.
- The *whiskers* extend left to the least score 15, and right to the greatest score 30, showing the range.

The second diagram shows the box plot of the 15 toaster-sale scores underneath, and above it the box plot of the subsequent 16 toaster-sale scores. It is a *parallel box plot*.



The addition of the one extra score 17 has increased the median and the upper quartile. The point here is to see immediately any significant differences in the overall picture.

Outliers

When scientists do experiments, they often end up with data that they secretly wish that they had not collected — perhaps they were not expecting these results, or the data do not fit their theories, or the data are 'clearly' the result of an experimental error, or they were 'wrongly recorded' by a research assistant, or they just look strange and the scientist doesn't know what to do about them. These pieces of data are scores that lie a long way from most of the data collected and they consequently muck up the patterns that the other data create.

ISBN 978-1-108-76630-2

Such pieces of data are called *outliers*, and the inclusion or exclusion of these outliers from datasets causes serious arguments wherever statistics is used. And time and time again, outliers have been an indication of an inadequate theory that needed to be reformulated.

There are no generally accepted criteria for outliers, just a few contradictory rules that people argue about. In this course, we shall usually take a criterion based on quartiles and the interquartile range IQR. We usually take an outlier to be a score that lies

more than 1.5 × IQR below Q_1 or more than 1.5 × IQR above Q_3.

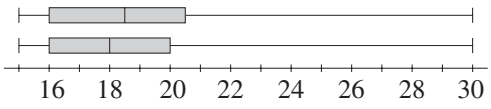
This criterion is very simple and can usually be calculated mentally.

- One problem with this test is that the gap between the suspect outlier and the next score or scores is also an important criterion, and *analysis of such gaps is missing from this test*.
- A second problem is that when there are very large datasets, we expect from the normal laws of probability to have scores many IQRs from the quartiles, but *the size of the sample is missing from this test*.

In the end, nothing can replace careful attention to the scores themselves. In most circumstances, outliers should be left in the dataset, but probably should be displayed differently and commented on.

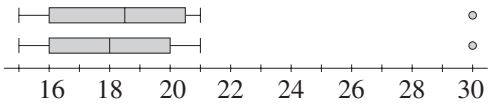
Outliers in box-and-whisker plots

Box-and-whisker plots are easily adapted to show outliers. You will have seen in both the sets of toaster-sales data that one week 30 toasters were sold, whereas the next highest score is 21.



The score of 30 is an outlier by two criteria — it is well separated from the other scores, and (in the case of the 15 scores) it is 10 above the upper quartile, which is 2.5 times the interquartile range.

We have therefore redrawn the parallel box plots to the right, with the right-hand whisker stopping at 20, and a circle placed at 30 as a code for the outlier. Outliers are often indicated on a box plot in this or a similar way.



Explaining outliers is most important — there was a toaster sale that week.

Example 7 15C

a Throw a die until a six occurs, and record the number n of throws needed. Do this 50 times, or use the results of class members, or simulate it using random numbers or a spreadsheet.

b Construct a frequency table and cumulative frequency table.

c Find the median, the quartiles, and the interquartile range.

d Draw a box plot and discuss any outliers.

e Let X be the number of tosses required to get a six. Explain why

$$P(X = 1) = \tfrac{1}{6}, \quad P(X = 2) = \tfrac{5}{6} \times \tfrac{1}{6}, \quad P(X = 3) = \left(\tfrac{5}{6}\right)^2 \times \tfrac{1}{6}, \quad \ldots,$$

and use GP theory to prove that the limiting sum of these probabilities is 1.

f Copy and complete the following cumulative discrete probability table, giving each value correct to three decimal places.

n	1	2	3	4	5	6	7	8	9	10	...
$P(X = n)$	0.167										
$P(X \leq n)$	0.167										

ISBN 978-1-108-76630-2

g Hence find the theoretical mean, quartiles and interquartile ranges. What values of n are classified as outliers according to the IQR criterion? Then sketch a box plot of the theoretical results.

h Explain why both the box plot of the data and the box plot of the theoretical distribution are unsymmetric.

SOLUTION

a These results were obtained by simulation using random numbers based on part **e**. Note the gaps in the last two results.

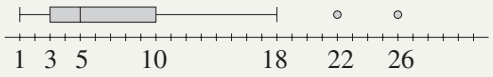
b

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	22	26
f	4	5	9	4	5	2	3	1	2	3	1	2	1	1	1	2	1	1	1	1
Cml	4	9	18	22	27	29	32	33	35	38	39	41	42	43	44	46	47	48	49	50

c The median Q_2 is the average of the 25th and 26th scores, which is 5. The lower quartile Q_1 is the 13th score, which is 3. The upper quartile Q_3 is the 38th score, which is 10.

d The IQR is $10 - 3 = 7$, so

$$Q_3 + 1.5 \times \text{IQR} = 10 + 10\tfrac{1}{2} = 20\tfrac{1}{2}.$$



The IQR criterion for outliers classifies the last two scores 22 and 26 as outliers. This is also a common-sense classification, because these last two scores are well separated from the other scores.

e Using standard probability techniques from Chapter 12 of the Year 11 book,

$$P(X = 1) = \tfrac{1}{6} = \left(\tfrac{5}{6}\right)^0 \times \tfrac{1}{6},$$
$$P(X = 2) = P(\text{TH}) = \tfrac{5}{6} \times \tfrac{1}{6} = \left(\tfrac{5}{6}\right)^1 \times \tfrac{1}{6},$$
$$P(X = 3) = P(\text{TTH}) = \tfrac{5}{6} \times \tfrac{5}{6} \times \tfrac{1}{6} = \left(\tfrac{5}{6}\right)^2 \times \tfrac{1}{6},$$
$$\ldots$$

This is a GP with first term $a = \frac{1}{6}$ and ratio $r = \frac{5}{6}$,

so $$S_\infty = \frac{a}{1 - r} = \tfrac{1}{6} \div \left(1 - \tfrac{5}{6}\right) = 1.$$

f

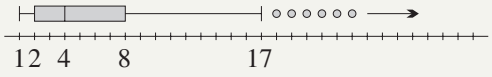
n	1	2	3	4	5	6	7	8	9	10
$P(X = n)$	0.167	0.139	0.116	0.096	0.080	0.067	0.056	0.047	0.039	0.032
$P(X \leq n)$	0.167	0.306	0.421	0.518	0.598	0.665	0.721	0.767	0.806	0.838

g The median is $Q_2 = 4$, because it is the first score whose cumulative probability is at least $\frac{1}{2}$.

The lower quartile is $Q_1 = 2$ because it is the first score whose cumulative probability is $\frac{1}{4}$.

The upper quartile is $Q_3 = 8$ because it is the first score whose cumulative probability is $\frac{3}{4}$.

Hence the IQR is $8 - 2 = 6$. Thus the IQR criterion classifies as an outlier every score greater than $Q_3 + 1.5 \times \text{IQR} = 8 + 9 = 17$.



h The frequencies or probabilities are bunched up on the left and spread out or *skewed* on the right.

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

Median and quartiles vs mean and standard deviation

We now have two families of summary statistics:

- Mean, variance and standard deviation.
- Median, quartiles and interquartile range.

Earlier in this section, we used 15 weeks of toaster sales and found that

$$Q_1 = 16, \qquad Q_2 = 18, \qquad Q_3 = 20, \qquad \text{IQR} = 4.$$

After further calculation, the mean is 18.73 and the standard deviation is 3.53.

Now suppose that we replace the outlier 30 by 21, the highest of the other scores. The mean, quartiles and interquartile range do not change, but the mean changes to 18.13, and the standard deviation changes dramatically to 2.00. That is, the standard deviation may be very sensitive to outliers, the mean less so.

If Reliable Appliances is tallying up its cash flow and profits, they would use mean and standard deviation. If they were looking at their marketing and want to study toaster purchases outside exceptional situations such as sales, they would use the median and the quartiles, which are *robust* to outliers.

House prices are another much-discussed example. The prices of homes are stretched out or *skewed* on the right by very expensive homes, so that the median is a more useful measure of prices of ordinary homes than the mean. The upper quartile is also not affected by those very expensive homes, so that interquartile range may be a better measure of the spread than the standard deviation.

Summary statistics review

The summary statistics discussed in Sections 15A–15C were:

10 SUMMARY STATISTICS

Measures of location

Mode, median, mean

Measures of spread

Range, interquartile range, variance, standard deviation,

The five-number summary

minimum, first quartile Q_1, median Q_2, third quartile Q_3, maximum

A *box-and-whisker plot* is constructed from the five-number summary.

We have described data several times as being *skewed*. Recall from earlier years that data are skewed in the direction of the tail, not the peak.

- *Skewed to the right*, or *positively skewed*, means that there is bigger tail on the right-hand side.
- *Skewed to the left*, or *negatively skewed*, means that there is bigger tail on the left-hand side.

ISBN 978-1-108-76630-2

Exercise 15C

FOUNDATION

1 For each dataset, calculate the three measures of central tendency (mean, median and mode), and calculate the range.

a 4, 8, 5, 2, 9, 12, 8

b 12, 23, 18, 30, 24, 29, 19, 22, 25, 12

c 7, 6, 2, 5, 7, 3, 4, 5, 7, 6

d 54, 62, 73, 57, 61, 61, 54, 66, 73

2 Find the middle quartiles Q_1, Q_2, Q_3 and the interquartile range IQR $= Q_3 - Q_1$ for each dataset.

a 3, 7, 9, 13, 14, 17, 20

b 8, 12, 13, 17, 20, 24, 27, 31

c 4, 7, 8, 9, 11, 14, 17, 19, 20

d 2, 5, 7, 10, 13, 17

e 2, 4, 5, 7, 12, 13, 14

f 8, 10, 12, 15, 17, 21, 22

g 3, 4, 6, 7, 9, 11, 13, 14, 18

h 9, 12, 13, 15, 18, 21

3 Find the IQR of each unordered dataset.

a 15, 10, 12, 19, 1, 17, 13, 6, 2

b 1, 11, 14, 9, 0, 4

c 12, 7, 9, 11, 13, 2, 9

d 6, 3, 2, 12, 0, 6, 8, 4

e 7, 11, 7, 5, 10, 7

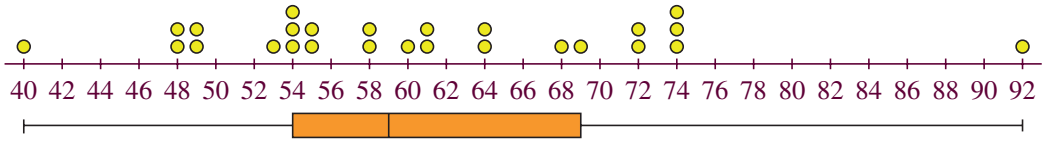
f 2, 9, 5, 4, 5, 9, 12

g 8, 3, 4, 1, 12, 2, 4, 11

h 10, 4, 5, 18, 11, 13, 2, 9, 7

DEVELOPMENT

4



The combined box plot and dot diagram above shows the exam scores for a small cohort of 26 students.

a Use your intuition to identify any *outliers*, thinking here of outliers as scores that are a long way from the rest of the data.

b Write down the five-number summary statistics: the minimum value, the lower quartiles Q_1, the median Q_2, the upper quartile Q_3, and the maximum value.

c The IQR criterion identifies *outliers* as those values less than $Q_1 - 1.5 \times$ IQR or more than $Q_3 + 1.5 \times$ IQR. Use this criterion to identify any outliers.

d Do your answers to part **a** and **c** agree?

e Recalculate the three quartiles Q_1, Q_2, Q_3 if we:

i omit 40 only, **ii** omit 92 only, **iii** omit both 40 and 92.

f Do the quartiles and the IQR change much when outliers are removed?

g Calculate the mean and standard deviation of the dataset correct to one decimal place:

i with all values, **ii** without 40,

iii without 92, **iv** without 40 and 92.

h What is the change in the standard deviation in part **g iv** as a result of removing the outlying values, as a percentage of the standard deviation in part **g i**?

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

5 **a** For each dataset, calculate the interquartile range and identify any outliers using the IQR criterion. Then draw a combined dot plot and box-and-whisker plot.

i 3, 5, 6, 7, 7, 9, 18

ii 9, 10, 10, 12, 12, 13, 18

iii 9, 10, 10, 12, 12, 13, 18, 19

iv 1, 3, 7, 9, 9, 9, 10, 10, 11, 12

v 1, 3, 8, 9, 9, 9, 10, 10, 11, 12

vi 5, 7, 7, 8, 8, 8, 8, 9

vii 5, 7, 7, 8, 8, 8, 8, 9, 9

viii 1, 1, 2, 2, 3, 10, 12, 12, 12, 12, 13, 13, 13, 13, 13, 14, 14, 15, 16, 22, 22, 23, 24, 25

b By reference to your answers in part **a**, comment on how well this definition of outliers seems to agree with your intuitive notion of an outlier as an extreme value.

c The dataset in part **a i** contained an outlier. Redraw the box-and-whisker plot for part **a i** with the whiskers excluding the outlier, and with the outlier shown as a single small circle on your diagram.

6 **a** Draw up a frequency table and hence determine the mean and standard deviation for the dataset

16, 12, 13, 14, 14, 14, 12, 16, 15, 24

b Construct a dot plot. Do any of the values seem to be *outliers*, that is isolated points, a long way from the rest of the data?

c Confirm that the IQR criterion of an outlier agrees with your instinct in part **b**.

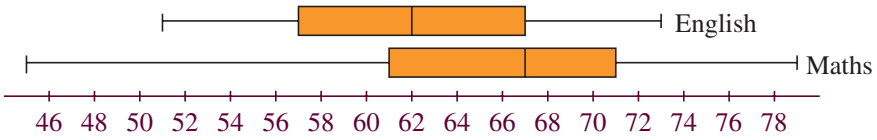
d Remove the outlier you should have found in parts **b**–**c**, and recalculate the mean and standard deviation.

e Has the outlier had a big effect on the mean and standard deviation? Use your tabular calculations to explain why this might be so.

f Has the outlier had a big influence on the median or IQR?

g The interquartile range and the standard deviation are both measures of spread. When might the interquartile range be a better measure than the standard deviation?

7 The results of a class in an English task and a mathematics task are shown in the parallel box plot below.



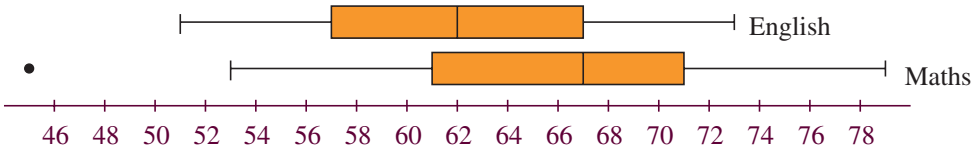
a Emily's result was in the bottom half of the English cohort. What can be said about her mark?

b What percentage of the students fall below 67 in mathematics?

c Comment on the spread of the two distributions by comparing the medians of the two sets of scores, and then by comparing the entire distributions.

d Xavier gets 66 in both English and mathematics. Is this a more impressive result in English or in mathematics, and why?

e Closer analysis shows that there is actually an outlier in the mathematics results. The box plots have been redrawn, and the outlier is shown separately with a closed circle, as in the diagram below:



i What was the outlier score in mathematics?

ii Compare again the spread of the bottom 25% of the students in mathematics and English, this time with the outlier excluded.

ISBN 978-1-108-76630-2

8 An English class completes a writing and a speaking task. The results are displayed in the back-to-back stem-and-leaf plot below.

Writing task		Speaking task
5	3	7
	4	
6 6 5 2	5	1 3 4 7 8 8
9 9 8 8 7 7 4 4	6	3 5 5 6 6 7 8
5 5 4 4 2 1	7	1 1 3 5 7
	8	
1	9	3

a Genjo got 35 in the writing task. What was his score in the speaking task?

b For the writing task:

i calculate the mean, median and range,

ii calculate the interquartile range and determine any outliers.

c For the speaking task:

i calculate the mean, median and range,

ii calculate the interquartile range and determine any outliers.

d Which set of results was more impressive?

ENRICHMENT

9 [An investigation that could become a project] In this section we have given two tests for outliers — the IQR criterion, and graphing the data and applying common sense with particular attention to any gaps. The discussion of outliers is an important subject and books have been written on the subject. Interested students may like to investigate this further. Questions may include the following:

a Some statisticians label scores that are below $Q1 - 3 \times \text{IQR}$ or above $Q3 + 3 \times \text{IQR}$ as *extreme outliers*. Scores below $Q1 - 1.5 \times \text{IQR}$ or above $Q3 + 1.5 \times \text{IQR}$ that are not extreme outliers are called *mild outliers*. What questions in this exercise have included extreme outliers? Generate some data with both and note the difference on a dot plot and a box-and-whisker plot.

b What is the effect of having two or more outliers in a distribution — can they mask each other's existence from an IQR criterion test? Generate datasets to explore this. Include the possibility of multiple outliers at both ends of the dataset, or only on one end.

c Outliers can also be defined by measuring the number of standard deviations from the mean. Calculate the number of standard deviations that the outliers in this exercise are from the mean, and decide if this could be developed into a reasonable criterion. Start with the datasets in Question **5**.

ISBN 978-1-108-76630-2

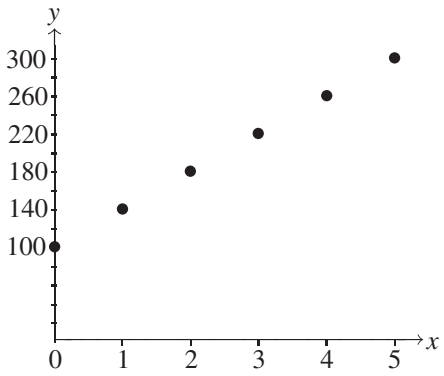
15D Bivariate data

For bivariate data, two further summary statistics are used — correlation, and the line of best fit. This section takes an intuitive approach to correlation and line of best fit, using displays and drawing the line of best fit by eye.

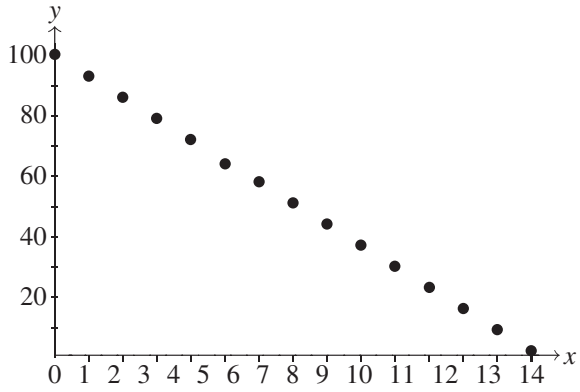
Section 15E will introduce formulae for correlation and the least squares version of the line of best fit. The last section, Section 15F, will use technology to analyse and display bivariate data.

Variables can be correlated without being related by a function

This course mostly concerns functions, where a variable y is completely determined by a variable x. Here are two linear functions, one with positive gradient, and one with negative gradient.

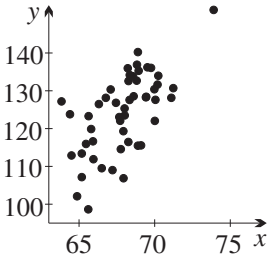


An electrician charges \$100 to visit a home, then \$40 for each power point. His total fee y for installing x power points in a home is $y = 100 + 40x$.



One hundred old cars were dumped in a park, and the council is removing 7 per day. The number y of cars remaining after x days is $y = 100 - 7x$.

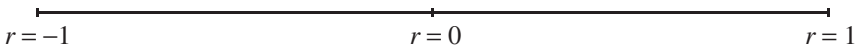
Now think about the relationship between the heights and weights of people. The *scatterplot* to the right plots the height x inches and weights y pounds of a group of people, with x as the *independent variable* and y as the *dependent variable*.



The cluster of dots is spread out because people of the same height don't all have the same weight, that is, y is not completely determined by x. And yet there seems to be more than a random relationship between the heights and weight of the people in the group, because it looks as if taller people tend to be heavier.

Correlation is the word for such a statistical relationship. In this section we will judge correlation by eye after the points are plotted on a scatterplot, then in Sections 15E–15F we will introduce a summary statistic called *Pearson's correlation coefficient* r. When y is determined by x, as in the two graphs above,

- $r = 1$, if they lie on a straight line and increase together,
- $r = -1$, if they lie on a straight line with one increasing, the other decreasing.
- $r = 0$, if there is no linear relationship between the two variables.



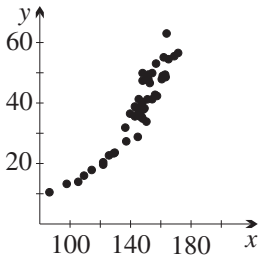
Correlations of 1 and -1 are called *perfect correlation*. We will perform some calculations of Pearson's correlation coefficient by hand in Section 15E, and by technology in Section 15F.

ISBN 978-1-108-76630-2

The choice of which variable is independent and which is dependent is sometimes obvious, but is sometimes rather arbitrary, and in either case the choice may well be a matter for a scientist rather than a mathematician. A choice must be made, however, for the methods of these sections to work.

Heights and weights — a positive correlation

The raw data of the heights of people in Section 15B came also with the weights of those people.

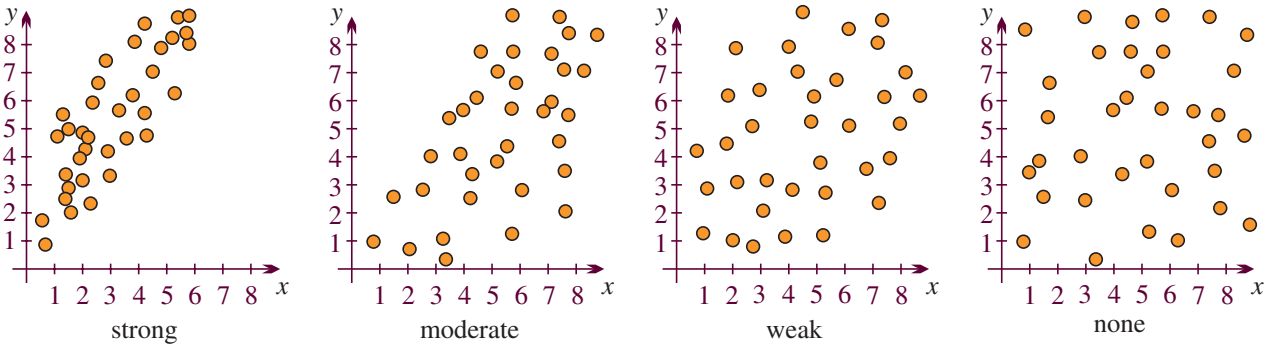


Here is the *scatterplot* of the first 50 pairs of measurements, with the height x cm taken as the independent variable on the horizontal axis, and the weight y kg taken as the dependent variable on the vertical axis.

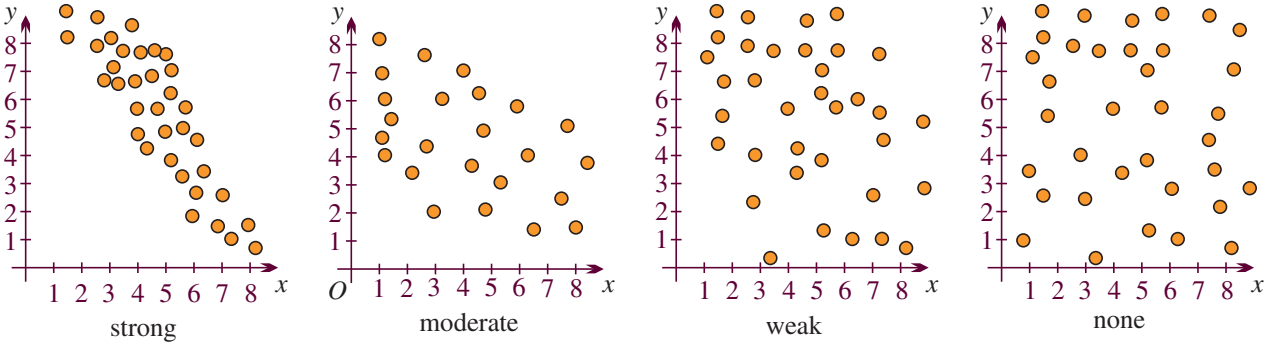
Each dot in the chart represents the height and weight of one person. Notice that the horizontal scale starts at 80 kg, not at zero.

This gives a visual demonstration that the weight in that group is very closely related to the height. We can classify this visually as a *strong correlation* — that the actual value of the correlation here is about 0.928, which is regarded as very strong (the correlation of the previous scatterplot is 0.64 — still strong, according to many published guides). The correlation is also *positive*, meaning that both variables increase together and that the slope of the cluster is positive.

Here is a rough guide to positive correlations — it can only be a rough guide:



And here is a rough guide to negative correlations:



Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

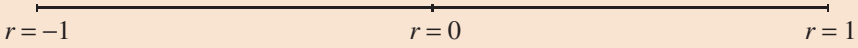
Non-linear correlation

Data are always more complicated than whatever is said about them. The cluster of dots in the heights–weights scatterplot above has a definite curve in it. Perhaps it should be tested not against a straight line, which is all that we will do in this chapter, but against a curve. Should that curve be quadratic, or exponential? Perhaps it should be cubic, because the volume of similar figures is proportional to the cube of the height. Such reasoning shows how prediction and causation are always involved in any discussion about data.

11 CORRELATION

Bivariate data means data in the form of ordered pairs. Suppose that we have identified an *independent variable* x and a *dependent variable* y in bivariate data.

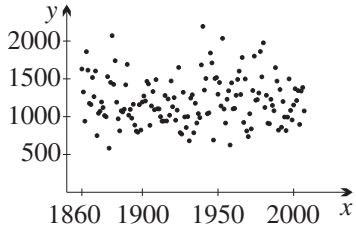
- A *scatterplot* graphs all these pairs on the one coordinate plane.
- *Linear correlation* occurs when the dots in the scatterplot tend to cluster in a shape vaguely like a line.
- Linear correlation is usually measured by *Pearson's correlation coefficient*, or simply the *correlation*, which is a real number r in the interval $-1 \leq r \leq 1$.



- *Non-linear correlation* occurs when the dots in the scatterplot tend to cluster in a shape vaguely like some other curve.

Rainfall over the years — no correlation

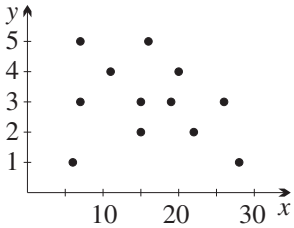
The Bureau of Meteorology has data for rainfall in Sydney for every day of every year since about 1858. We can download and massage that data to find the annual rainfall y in millimetres each year x from 1860–2007 — the scatterplot is drawn to the right.



There is no (linear) correlation. Pearson's correlation coefficient is -0.014, which is virtually zero. But look at the dots! You can see the drought years and the flood years.

Callers put on hold — a negative correlation

Customers of a large technology company often ring about problems, and are put on hold for short or long periods. At the end of their call they are asked to rank the company 1 (very dissatisfied) to 5 (very satisfied). Some preliminary test data on the waiting time x minutes and the rank y were taken, with the following results.



x	7	15	22	11	20	15	7	28	6	16	26	19
y	5	2	2	4	4	3	3	1	1	5	3	3

The correlation here is negative because the cluster slopes backwards, and it can be roughly characterised as weak (the calculated correlation is about -0.26).

The point $(6, 1)$ could be identified as an *outlier*. It has a big effect. Place your finger over this single dot. The correlation now looks moderate (the calculated value is about -0.57). Correlation is very sensitive to outliers when there are few values — both the visual impression and the calculated value.

ISBN 978-1-108-76630-2

12 POSITIVE, NEGATIVE AND ZERO CORRELATION

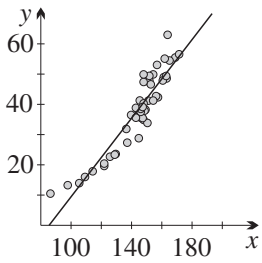
- A positive slope in the cluster corresponds to $0 < r \leq 1$, and is called *positive correlation*. A correlation of $r = 1$ means *perfect positive correlation*, that is, y is a linear function of x with some positive gradient.
- A negative slope in the cluster corresponds to $-1 \leq r < 0$, and is called *negative correlation*. A correlation of $r = -1$ means *perfect negative correlation*, that is, y is a linear function of x with some negative gradient.
- $r = 0$ means that there is no linear correlation between the variables.

One qualification about zero correlation is needed. A horizontal line of points, or a cluster in the vague shape of a horizontal line, both have zero correlation.

The line of best fit

The purpose of correlation is to test whether the cluster of results are suggesting a line, sloping forwards or backwards. The line that they best cluster around is called the *line of best fit* or the *regression line*. The most common way of calculating it is to find the line that minimises the squares of the vertical distances from the points to the line, and the resulting line is called the *least squares regression line*. In Section 15E (a Challenge section) we will use formulae to calculate the gradient and y-intercept of this line of best fit, and in Section 15F we will use technology, but for now, we will estimate it by eye.

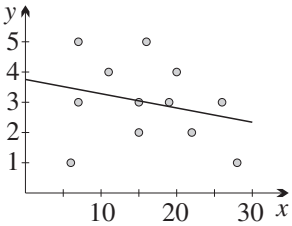
Here again is the scatterplot of the heights and weights of 50 individuals. We have drawn on it by eye a line of best fit. It has gradient about 0.65, and x-intercept about 85. Its equation is therefore about



$$y - 0 = 0.65(x - 85)$$
$$y = 0.65x - 55.25$$

Using the formulae in the next section, the gradient and y-intercept of the least-squares regression line are about 0.649 and -55.52. Again, be careful because the horizontal scale starts at 80 kg, not zero.

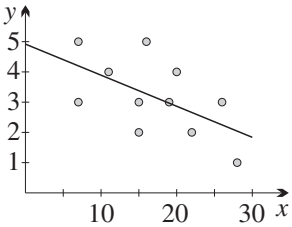
The scatterplot of the callers on hold shows only weak correlation, so it is difficult to draw a line of best fit by eye. By the formulae, the gradient is about -0.047, and the y-intercept is about 3.75, giving the line



$$y = -0.037x + 3.75$$

that we have drawn on the graph.

In the second scatterplot, the outlier has been omitted (this is not the recommended procedure). The correlation is now moderate, and this makes it easier to draw the line by eye. The formulae in the next section tell us that the gradient is about -0.10, and the y-intercept is about 4.92. The resulting line



$$y = -0.1x + 4.92$$

has been drawn on the graph.

Why was the outlier there? Perhaps someone rang simply to cancel the service because of all sorts of other problems, and the short waiting time had no effect on his rank of 1. Should it be ignored? That is up to management and what questions they are asking.

ISBN 978-1-108-76630-2

Double, triple and multiple points

None of our diagrams or datasets so far have repeated points, but in practice repeated points often occur. There are different conventions — use ever larger circles, use a code of circles, squares and crosses, place numbers inside the circle, ... Use whatever is convenient.

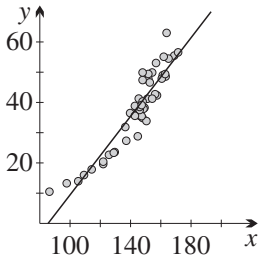
It is important to be aware of repeated points when judging correlation and line of best fit by eye, otherwise one's judgement will be right out. See Question 4 of Exercise 15D.

13 REGRESSION — THE LINE OF BEST FIT

- Given bivariate data, we can calculate the *line of best fit* or *regression line*.
- When correlation can be seen clearly, we can draw the line of best fit by eye. It is important then to identify any multiple points in the scatter graph.
- Outliers may have a significant effect on correlation and the line of best fit.

Interpolation, extrapolation, prediction and causation

Interpolation mean *predicting further results within the range of the variables in the data*. That is reasonably straight forward once the line of best fit has been drawn. Interpolation is justified, provided that we are convinced that the sample we are working from is not biased.



Extrapolation mean *predicting further results outside the range of the variables in the data*. That can present a real problem, because the situation is often quite different outside the range of sample values. For example, given the scatterplot and line of best fit of 50 heights and weights, a person of height 85 cm would be predicted to have zero weight, and a baby of height 40 cm would float away!

Even when data have extremely high correlation — high enough for the relationship to be regarded as a function — extrapolation is dangerous. Newton's laws of motion cannot be extrapolated to speeds approaching the speed of light because of relativity theory, and cannot be extrapolated to very tiny particles because of quantum mechanics.

Causation is best left to scientists. If events A and B are correlated, there are four possibilities — A causes B, B causes A, A and B have a common cause C, and the correlation is a fluke. Events can have multiple causes, particularly in medicine and weather. Many phenomena are chaotic, such as the weather and eddies in flowing water, making prediction impossible and rendering the idea of causation extremely complicated.

14 INTERPOLATION, EXTRAPOLATION, PREDICTION AND CAUSATION

- Provided that the data are reliable, the line of best fit can reasonably be used for *interpolation*, meaning prediction of results within the range of the variables.
- It can also be used for *extrapolation*, meaning prediction of results outside the range of the variables in the data, but this requires caution and common sense because the results of extrapolation are often very misleading.
- Questions of causation are probably best left to scientists.

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

Exercise 15D

FOUNDATION

1 In the following relationships, identify the most reasonable choice for: **i** the independent variable, **ii** the dependent variable. Is there any uncertainty in your answer?

a height and weight,

b the area of a circle and its radius,

c the weight of meat purchased and the price,

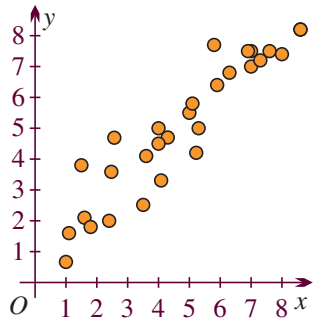
d the outcome for a player at Wimbledon Tennis championships and the player's previous world rank,

e outside temperature and household power consumption,

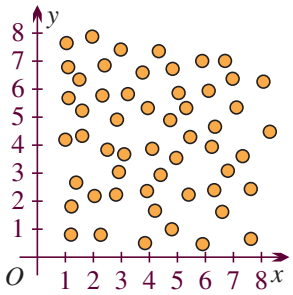
f pronumerals x and y for the many-to-one function $y = f(x)$.

2 By eye, decide if each of these graphs is an example of strong, moderate, weak or no linear correlation. Where there is a correlation, note whether it is a positive or negative correlation.

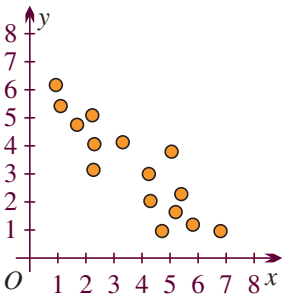
a



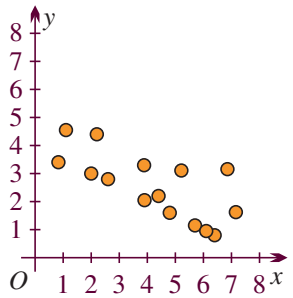
b



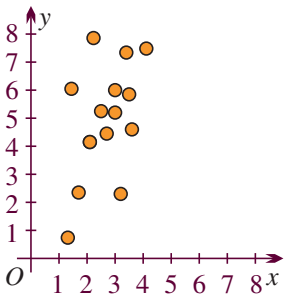
c



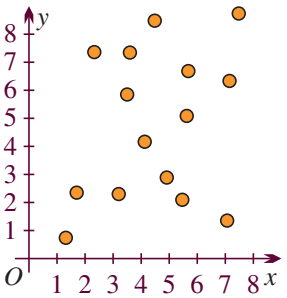
d



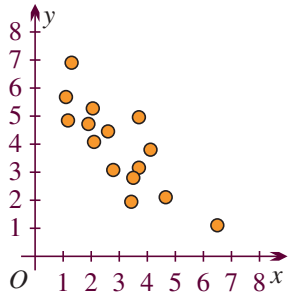
e



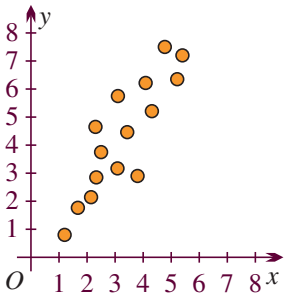
f



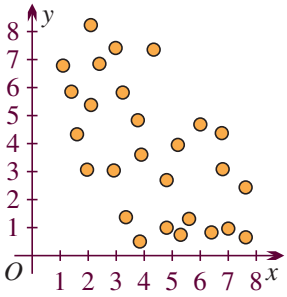
g



h



i



3 Plot each dataset on a separate diagram and draw a line of best fit by eye. Write down the equation of the line of best fit that you have drawn.

a $(0, 3)$ $(1, 2)$ $(2, 4)$ $(3, 3)$ $(4, 5)$ $(5, 5)$ $(6, 7)$ $(9, 7)$

b $(0, 30)$ $(1, 35)$ $(2, 45)$ $(3, 40)$ $(4, 55)$ $(5, 55)$ $(6, 70)$ $(7, 65)$

c $(0, 16)$ $(2, 14)$ $(3, 8)$ $(4, 10)$ $(6, 7)$ $(7, 4)$ $(9, 4)$ $(12, 0)$

d $(0, 6)$ $(1, 8)$ $(2, 7)$ $(3, 5)$ $(4, 4)$ $(5, 5)$ $(6, 4)$ $(7, 1)$

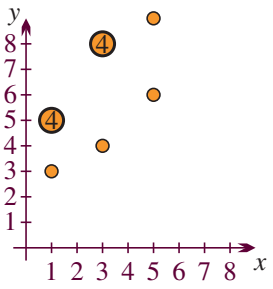
Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

4 The following datasets include repeated points, with the frequency indicated by a number on the plot. For each question, allowing for the extra weighting of the repeated points:

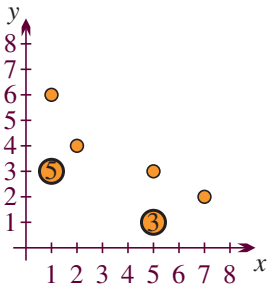
i Estimate the strength of the correlation

ii Copy the diagram and draw in a line of best fit by eye,.

a

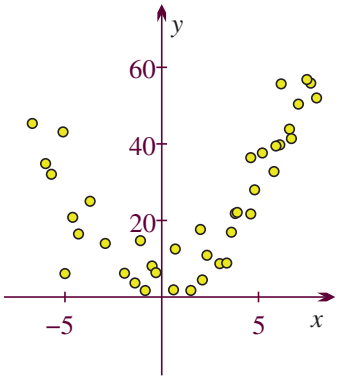


b

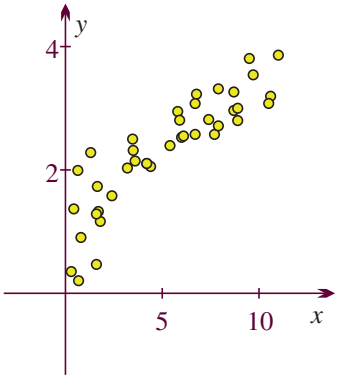


5 Not every dataset shows a linear relationship — some data are best modelled by a quadratic curve (a parabola), a circle (or semi-circle), a hyperbola, a square root, an exponential, or some other such curve. By eye, suggest what type of curve might model each dataset.

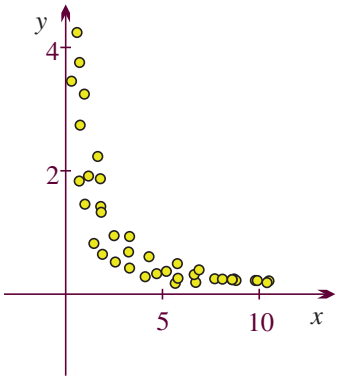
a



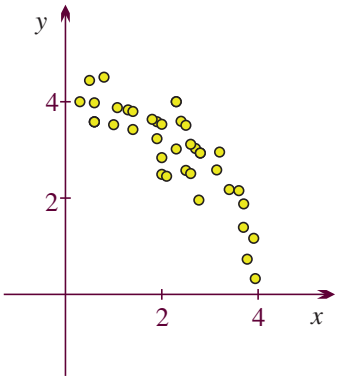
b



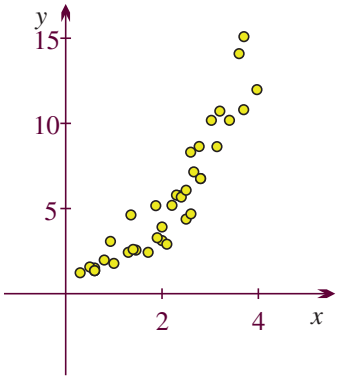
c



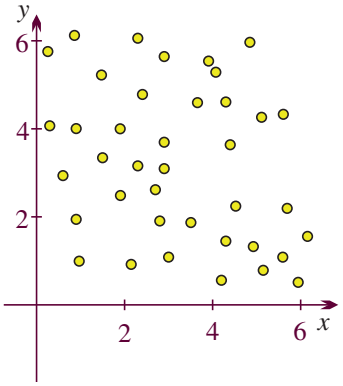
d



e



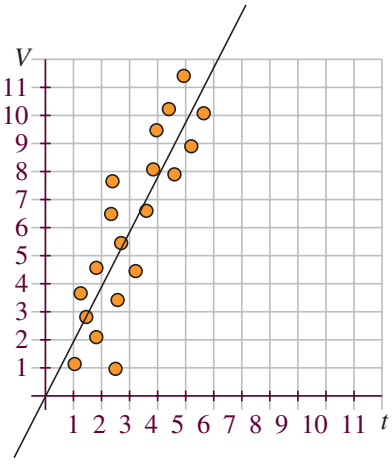
f



Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

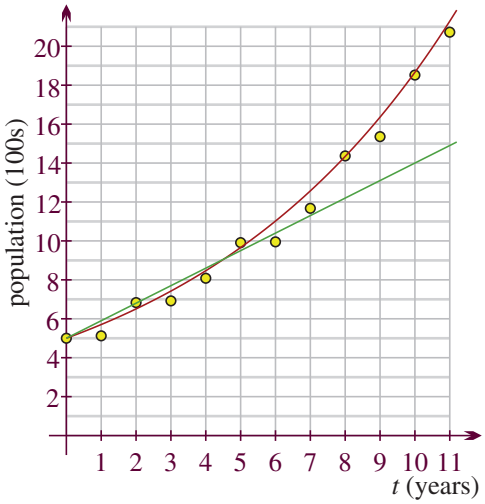
DEVELOPMENT

6 Yasuf has conducted an experiment. He recorded the volume in litres of water that flowed through a pipe in a given time of between 1 and 6 minutes. Then he repeated this procedure many times over a period of several hours in the middle of the day. His results are shown in the scatterplot to the right. Yasuf has also drawn a line of best fit through the data.



- **a** Use interpolation, by reading off the graph, to estimate the volume of water flowing through the pipe in: **i** 3 minutes, **ii** 5 minutes.
- **b** Estimate the equation of the line of best fit.
- **c** What is the value of the V-intercept, and why would you expect that value?
- **d** What is the physical meaning of the gradient?
- **e** Should Yasuf have drawn the line down through the origin into the third quadrant?
- **f** Why do you think the correlation in this experiment is not perfect, that is, why don't all the points lie on a straight line?
- **g** Use the equation of the line of best fit determined by Yasuf to extrapolate the amount of water flowing through the pipe in half an hour. Is this extrapolation reasonable?
- **h** How long would be required for the pipe to disgorge 45 litres of water?
- **i** Yasuf wishes to estimate how much water will flow through the pipe in one 24-hour day. Explain why extrapolating from these results may not be valid.

7 Population in the town of Hammonsville has been growing strongly over the last few years. When town planners first took a census in 2010, the population of people living in the town was 500. The population over the next 11 years is recorded in the scatterplot to the right. Planners have also attempted to fit the data with various curves to model future population growth.



- **a** What was the population five years later in 2015?
- **b** After seven years the planners began to draw a scatterplot, and they added a line of best fit in order to estimate population trends.

  Place your fingers over the last four points on the scatterplot so that you see only the 8-point scatter graph that the planners had after 7 years.

  - **i** Find the equation of the line (population P in 100s as a function of time t years since 2010). Use the fact that the line passes through (0, 5) and (10, 14).
  - **ii** Does this line look a good fit for the seven-year period?
  - **iii** In 2017 predictors used this model to extrapolate the population after a further two years (that is, when $t = 9$). What was the error in their prediction, compared with the plotted population in that year?

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

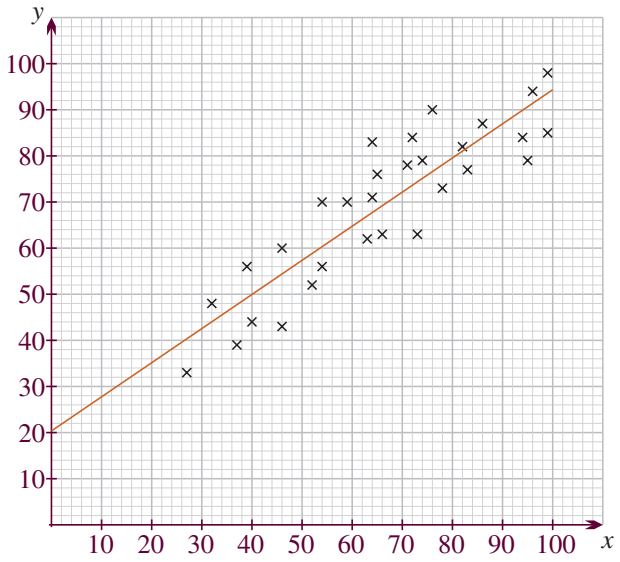
c Experts noticed that their model became an increasingly poor predictor as time went on, and instead attempted to fit the data with an exponential curve $P = 5 \times 2^{0.19t}$, where P is the population t years after 2010.

i Check the accuracy of this model by calculating its population prediction in 2019.

ii Would you expect this model to be accurate for the next 10 years?

d What does this question suggest about the general viability of using a line to fit data (that is, a line of best fit)?

8 Student percentage marks for assessment 1 (x) and assessment 2 (y) have been compared on a scatterplot, and a line of best fit has been drawn. There are no repeated points.



a What was the top mark in each assessment? Were they obtained by the same student?

b What was the bottom mark in each assessment? Were they obtained by the same student?

c Which assessment was more difficult? Give reasons for your claim.

d Some students were absent for one or the other of the assessment tasks, and their missing scores must be estimated. All scores are to be recorded as a whole number. Use interpolation on the line of best fit to estimate the score in the other assessment of a student who received:

i 40 in assessment 1, **ii** 60 in assessment 1, **iii** 80 in assessment 2,
iv 40 in assessment 2, **v** 15 in assessment 2.

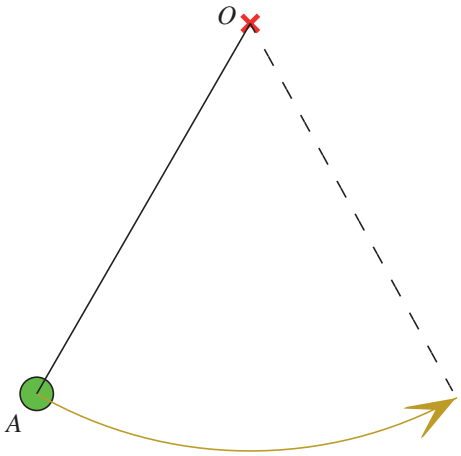
e Read off the coordinates of the points on the line at $x = 0$ and $x = 100$. Hence find the equation of the line of best fit.

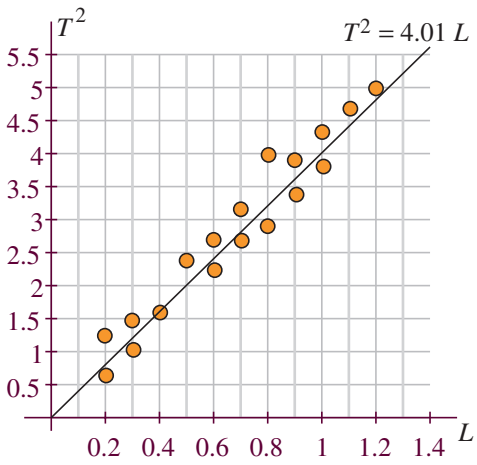
f Is this an accurate method of estimating a student's missing assessment score?

ISBN 978-1-108-76630-2

ENRICHMENT

9 A class is carrying out an experiment. A weight is attached by a string to a fixed point O. It is drawn aside and released, allowing it to swing and then return to its original position. The time taken to return is called the *period* of the pendulum. The experiment requires students to measure the periods T minutes for different string lengths L metres. Theory suggests that the square of the period is related to the length of the string. The class's results are shown in the scatterplot below, graphing T^2 on the vertical axis and L on the horizontal axis.





a For a given length of string, what is the maximum difference between the square of the period predicted by the linear model and the square of the measured period?

b This particular maximum difference is significantly larger than for other points, and it appears to be an outlier in the data. Comment on possible causes.

c The class has claimed great accuracy in measuring the times for a period, as can be seen from the strong correlation. What methods may they have used to achieve this accuracy?

d Scientist have developed a theoretical model relating T and L. The model predicts that $T = 2\pi\sqrt{\frac{L}{g}}$, where $g \doteqdot 9.8$ m/s^2. Does this agree with these experimental results?

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

15E Formulae for correlation and regression

You must be able use technology to calculate Pearson's correlation coefficient and the line of best fit from given data. The next section goes into some detail about those procedures.

This section presents the actual formulae for the line of best fit. They are rather elaborate, and calculations using them take more time and paper than the earlier calculations for mean and standard deviation. Nevertheless, calculating at least a few examples by hand can prevent statistics becoming a 'black-box' where the user of the results has no real idea what is happening. In a mathematics course, understanding is key.

These formulae could be regarded as Enrichment. The very short exercise has just one purpose — familiarity with the formulae.

The formula for Pearson's correlation coefficient

The standard measure of correlation is *Pearson's correlation coefficient* r. It tests only for linear correlation, that is, it gives a measure of how close the data are to being on a line of non-zero gradient,

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

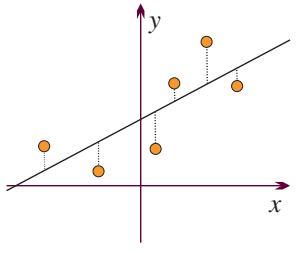
We will not develop this formula, but these remarks should help to understand its significance and how to go about calculating it.

- We must first calculate the means $\bar{x}$ and $\bar{y}$ of the x-values and of the y-values. The point $(\bar{x}, \bar{y})$ will lie in the middle of the cluster on the scatterplot.
- We need all the *deviations from the mean*. These are the deviations $x - \bar{x}$ of the x-values from their mean $\bar{x}$, and the deviations $y - \bar{y}$ of the y-values from their mean $\bar{y}$.
- For the numerator, we take each product $(x - \bar{x})(y - \bar{y})$. This is the key object, because if x and y both lie on the same side of their means, the product is positive, and if they lie on opposite sides, the product is negative. Adding them up gives some sense of whether the variables are working together, or working contrary to each other and cancelling out.
- The denominator is necessary to normalise the quantity and make it a ratio. In particular, notice that the units of x cancel out and the units of y cancel out, so that the resulting quotient r is a pure number.
- The denominator is closely related to the formulae for the standard deviation of x and of y. In fact, the formula for r can be rewritten using the standard deviation, and standard deviation calculations can be re-used here.
- Pearson's correlation coefficient is unaffected by units and gradient (apart from the sign of the gradient). If we change metres to centimetres, or multiply all the y-values by $+7$, there is no change in r. That is, only the clustering and the sign of the gradient are relevant.

Formulae for the regression line

The standard method of finding the regression line is to find the line that minimises the sum of the squares of the vertical distances from each plot to the line — this line is called the *least squares regression line*. Again, we omit any derivation, and simply state that the line is

$$y = mx + b, \text{ where } m = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} \text{ and } b = \bar{y} - m\bar{x}.$$



ISBN 978-1-108-76630-2

- The numerator of the gradient m is the same as the numerator of r.
- The denominator of m has already been calculated when calculating r.
- The value of the y-intercept b ensures that the line passes through $(\bar{x}, \bar{y})$.

Thus once the calculations required for r have been done, the calculation of the line of best fit is very quick.

Calculations using the callers on hold

The example in Section 15D of the waiting times of callers on hold was deliberately engineered so that the means were whole numbers. Otherwise an alternative form of the formula would need to be developed, which is not appropriate in this course, or machine calculation would be necessary.

The sums of the first and second lines of the table allow the means $\bar{x}$ and $\bar{y}$ of the x-values and y-values to be calculated. These means are needed in the third and fourth lines to calculate the deviations.

													Sum
x	7	15	22	11	20	15	7	28	6	16	26	19	192
y	5	2	2	4	4	3	3	1	1	5	3	3	36
$x - \bar{x}$	−9	−1	6	−5	4	−1	−9	12	−10	0	10	3	0
$y - \bar{y}$	2	−1	−1	1	1	0	0	−2	−2	2	0	0	0
$(x - \bar{x})^2$	81	1	36	25	16	1	81	144	100	0	100	9	594
$(y - \bar{y})^2$	4	1	1	1	1	0	0	4	4	4	0	0	20
$(x - \bar{x})(y - \bar{y})$	−18	1	−6	−5	4	0	0	−24	20	0	0	0	−28

First, $\bar{x} = \dfrac{\sum x}{n}$

$= \dfrac{192}{12}$

$= 16,$

and $\bar{y} = \dfrac{\sum y}{n}$

$= \dfrac{36}{12}$

$= 3.$

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

$= \dfrac{-28}{\sqrt{594 \times 20}}$

$\doteq -0.25689.$

For regression, $m = \dfrac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$

$= \dfrac{-28}{594}$

$\doteq -0.04714,$

and $b = \bar{y} - m\bar{x}$

$= 3 + \dfrac{28 \times 16}{594}$

$\doteq 3.75421.$

Thus the correlation is -0.26, and the line of best fit is $y = -0.047x + 3.754$.

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

15 FORMULAE FOR CORRELATION AND REGRESSION

Let $\bar{x}$ and $\bar{y}$ be the means of the x-values and y-values of a set of bivariate data.

- Pearson's correlation coefficient r is given by

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

- The least-squares regression line is $y = mx + b$, where

$$m = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} \quad \text{and} \quad b = \bar{y} - m\bar{x}.$$

Other forms of these formulae combine them with the formula for variance.

Classifying correlations

The verbal descriptions used in Section 15D for the strength of correlation are visual impressions of the scatterplot. There is no agreed relationship between those verbal descriptions and the values of the correlation coefficient calculated in this section and the next. The same criteria may not be appropriate for different disciplines or for different experiments within a discipline.

The authors regard the following as reasonably helpful suggestions. For positive correlations (and similarly for negative correlations),

Correlation	0.6–1.0	0.4–0.6	0.1–0.4	0.0–0.1
Description	strong	moderate	weak	virtually none

There are no rules — have the scatterplot at hand, think about the experiment, and be aware of any outliers.

According to this classification, the caller waiting times and ranks, with a correlation of about -0.26, show weak negative correlation. With the outlier (6, 1) removed, the correlation is about -0.57, which is moderate negative correlation.

The correlations of the set of eight scatterplots in Section 15D are:

Top row: 0.82, 0.59, 0.35, 0.07 Bottom row: -0.87, -0.5, -0.37, -0.09

Exercise 15E

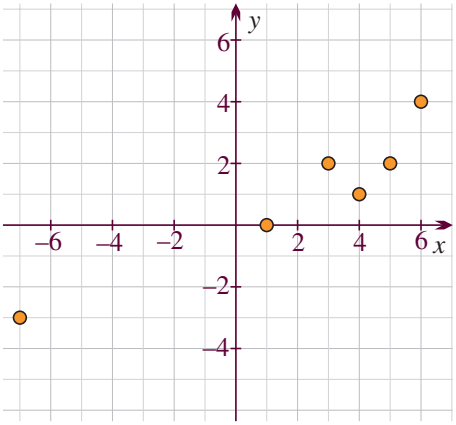
Note: If this exercise is attempted, the best approach is perhaps to do the calculations once or twice to gain some familiarity, and then to combine calculations by hand with calculations by technology. Further opportunities to use technology for such calculations are given in Questions 1–2 of Exercise 15F.

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

1 A student is given the task of calculating the line of best fit for the small dataset:

$(-7, -3)$ $(1, 0)$ $(3, 2)$ $(4, 1)$ $(5, 2)$ $(6, 4)$

The data are shown on the scatterplot to the right.



a Does the correlation of the data appear linear, and if so, does the correlation appear strong, weak or moderate?

b Copy the scatterplot into your book. By eye, estimate and draw the line of best fit for the data.

c Copy the following table into your book. Complete the sum for the first two rows, and hence calculate the means $\bar{x}$ and $\bar{y}$.

							Sum
x	-7	1	3	4	5	6	
y	-3	0	2	1	2	4	
$x - \bar{x}$							
$y - \bar{y}$							
$(x - \bar{x})^2$							
$(y - \bar{y})^2$							
$(x - \bar{x})(y - \bar{y})$							

d Mark the point $(\bar{x}, \bar{y})$. Does it fall on your estimated line of best fit?

e Complete the last five rows of the table.

f Use the formula

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

to find Pearson's correlation coefficient for the data, correct to two decimal places.

g Is r close to 1 or to -1? This would indicate that the line is a good fit for the data.

h Use the formula to calculate the gradient m of the least squares line of best fit,

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

i Use the formula $b = \bar{y} - m\bar{x}$ to calculate the y-intercept of the line of best fit,

j Write down the line of best fit calculated from these formula, rounding m and b each correct to one decimal place. If this line differs from your estimate of the line of best fit in part **b**, add this line to your diagram.

2 Repeat the previous question for the following datasets.

a $(-2, 0)$ $(0, 0)$ $(1, 1)$ $(3, 1)$ $(4, 2)$ $(6, 2)$

b $(-3, -4)$ $(-2, -3)$ $(0, 1)$ $(2, 3)$ $(3, 4)$ $(6, 8)$

c $(-4, 7)$ $(-2, 6)$ $(-1, 1)$ $(0, -1)$ $(1, -3)$ $(2, 1)$ $(4, -4)$

d $(-2, 6)$ $(-1, 3)$ $(0, 4)$ $(1, 2)$ $(2, 0)$ $(6, -3)$

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

15F Using technology with bivariate data

It is clear from Section 15E that calculations by hand are laborious. The right tools are statistics calculators, spreadsheets and statistical packages, all of which may be online. This section has several purposes, loosely grouped around technology.

- Small datasets suited to a statistics calculator to find r and line of best fit.
- Larger datasets for investigations using spreadsheets and other software.
- Investigation activities such as surveys to generate data for analysis.
- Investigations allowing the reader to search out raw data from the internet.

Waiting times of callers on hold — correlation and regression

The previous section calculated Pearson's correlation coefficient for the waiting times of callers and their ranking of the company (data in Section 15D). If you have a *calculator* that is capable of performing the calculations, then work out now how to use it to get the correlation and the line of best as calculated in Section 15E. The various calculators differ from one another, and we have not chosen not to give detailed instructions for any particular model. Find the instructions and read them.

For any extended work, however, a *spreadsheet* is the best tool to use (in the absence of specialised statistical software). Excel is widely used, and we have worked the example here using Excel. This example, and the questions in Exercise 15F, should be easily adapted to other technology, including online versions.

Excel has many different versions, and its functions are complicated. Use the help file or search online for guidance. In particular:

- The functions in Excel for Windows keep changing over time. All the functions used here were different some years ago.
- The Mac versions of Excel have their own peculiarities. For example, 'Fill down' and 'Fill right' are well-known difficulties on a Mac.

In particular, if you cannot find the search box in the fourth dotpoint below, search for 'Formula Builder' in the help file.

Here are the steps in finding the correlation and regression line using the most recent version of Excel 365. The dataset is drawn from the 'Waiting time of callers' just above Box 12 in Section 15D.

The data: Type the data into Excel from the table in Section 15D.

- Type 'Time x' into cell `A1` and 'Rank y' into cell `B1` — these are the headers.
- Then type the 12 data pairs into the cells `A2:A13` and `B2:B13`.

The means: Type 'Mean x' and 'Mean y' into cells `D1:E1`.

- Place the cursor into cell `D2` and type `=` into it. This initial character `=` is the code for Excel to interpret what follows in a cell as a function.
- You will notice that in the top left below the word `File` there is now a text box with a down arrow. Click on the arrow, then click on 'More functions'.
- In the resulting search box, enter 'Average'. Select `AVERAGE` and read its description. Perhaps also click the link to the help file, and perhaps compare this function with the function `AVERAGEA`.
- Double-clicking will bring up a dialogue. You only need Number1, and you can either enter the cells with the x-values as `A2:A13`, or select them in the spreadsheet so that Excel enters the cell labels. The result should be the mean 16.

ISBN 978-1-108-76630-2

- With cell D2 selected, look at the text box above the row-and-column array. It should give the formula in the cell, which is =AVERAGE(A2:A13).
- Do the same for the mean of the *y*-values in cell E2 — the mean is 3. But Excel is cleverer than this! Instead, select cells D2:E2, and press Ctrl+R to Fill right. You can check that the formula in cell D2 has been copied to cell E2, except that column A in the formula has been changed automatically to column B. You can also see immediately what cells have been referenced by selecting cell E2 and pressing the F2 key.

The standard deviations: Type 'SD *x*' and 'SD *y*' into cells D4:E4.

- Then repeat the steps to insert the means into cells D5:E5, except search for 'standard deviation', and select the function STDEV.P.
- (As we remarked in an extension note at the end of Section 15B, it may be more correct, because this is a sample, to use the sample standard deviation STDEV.S rather than the population standard deviation STDEV.P.)

The correlation: Type 'Correlation' into cell D7.

- Insert the correlation into cell D8 as before — type = and click on the down arrow, select 'More functions', but search for 'correlation' and select PEARSON.
- Enter the *x*-values into the top box and the *y*-values into the second box.

The line of best fit: Type 'Regression' into cell D10, 'Gradient' into cell D11 and 'Intercept' into cell E11.

- To insert the gradient of the regression line into cell D12, search for 'Regression' and select SLOPE. Be careful here because things are reversed! Enter the *y*-values into the top box, and the *x*-values into the second box.
- To insert the *y*-intercept of the regression line into the cell E12, search for 'Regression' and select INTERCEPT. As before, enter the *y*-values into the top box, and the *x*-values into the second box.

	A	B	C	D	E
1	Time *x*	Rank *y*		Mean *x*	Mean *y*
2	7	5		16	3
3	15	2			
4	22	2		SD *x*	SD *y*
5	11	4		7.035624	1.290994
6	20	4			
7	15	3		Correlation	
8	7	3		-0.25689	
9	28	1			
10	6	1		Regression	
11	16	5		Gradient	Intercept
12	26	3		-0.04714	3.754209
13	19	3			

Waiting times of callers on hold — the scatterplot and regression line

Excel can draw a scatterplot of the data with the regression line inserted.

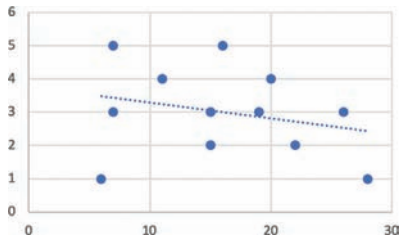
ISBN 978-1-108-76630-2

The scatterplot: First select the 24 cells `A2:B2` down to `A13:B13` that contain all the data.

- Click on the 'Insert' tab at the top of Excel. Find the 'Charts' group, and click 'Recommended charts'. Then click on the 'All charts' tab.
- You will see all sorts of charts there (including box-and-whisker that we discussed in Section 15C), but the one we want is 'X Y Scatter'. Click on it, place the cursor on the second of the two charts that just shows the 12 dots all in one colour, and double-click.
- You now have a scatterplot placed onto your spreadsheet. You can move it and resize it in the usual ways. When you double-click on the chart, a menu appears on the far right allowing other changes to be made.

The line of best fit: Click on the chart again, and a new tab 'Design' appears on the very top of the Excel window.

- Click on the 'Design' tab, find the group 'Chart layouts' (on the far left), and select 'Add chart element'.
- Go to 'Trendline' and select 'Linear'. Now the least squares regression line appears, calculated as in Section 15E.



Saving the chart: You can export the chart to other software such as Word or image-processing software.

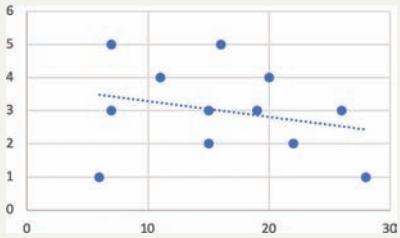
- Click on the chart, then right-click on the top border, then click 'Copy'.
- Paste the file where you want it.

Example 8 15F

Repeat all the steps above for the call waiting times with the outlier removed. Alternatively, your spreadsheet should allow you to remove one data point.

SOLUTION

- The means are about 16.91 for the x-values and 3.18 for the y-values.
- The standard deviations are about 6.64 for the x-values and 1.19 for the y-values.
- The correlation is about -0.57.
- The regression line is about $y = -0.1x + 4.92$.
- Excel's scatterplot with the regression line is copied to the right.



Internet data — investigations and possible projects

Exercise 15F is mostly concerned with processing data from the internet using technology. The methods are the same, whether there are 12 or 12 000 pairs of data.

Internet data, however, can be very rough. Always look carefully through the data for obvious anomalies. For example, in the heights and weights data introduced in Section 15D, many entries had the height field or the weight field missing, or both, and there were a few entries that seemed to have become corrupted — these are two of the few good reasons for omitting outliers.

The first few questions use technology for correlation and regression. Then the exercise is intended to provide investigations of various types. Most questions can easily be extended to projects by asking further questions and using and comparing several sources of data.

ISBN 978-1-108-76630-2

Exercise 15F

INVESTIGATION

Calculation of the gradient and the y-intercept of the line of best fit, and the calculation of Pearson's correlation coefficient, are best done using technology. These calculations may be done using a calculator with a regression analysis mode (also sometimes labelled '2 Var Stats' or '$A + Bx$'), with a spreadsheet program (such as Excel), with statistical software (such as the software package R), or with an online program designed to analyse such data. A number of such free online programs are available if you type *online line of regression* into a search engine.

The datasets in the first four questions are small, to enable practice with the technology before attempting to analyse some more realistic data. Further practice with technology is provided by the datasets in Exercise 15E.

1 Use technology to calculate Pearson's correlation coefficient and the equation of the line of best fit for each dataset.

a (1, 1.7) (2, 1.9) (3, 3.7) (4, 4) (5, 4.5) (6, 6.7)

b (−2, 3) (0, 2) (2, 2.3) (3, 3.1) (4, 4) (5, 5.7) (6, 6.1)

c (3.6, 3.2) (5.9, 3.9) (7, 1) (4.4, 5.2) (3.4, 6.2) (2.2, 5.7) (5, 2.3) (1.2, 8.2)

d (6.3, 0.7) (3.6, 4.9) (4, 2.6) (5.4, 1.2) (9, 2.3) (1.9, 1.8) (1.4, 7.4) (0.4, 3.6)

e (2.1, 3) (4.7, 6.9) (2.8, 4) (3.3, 5.5) (1.3, 3.3) (2.7, 4.8) (4.9, 7.8) (1.7, 2) (3.8, 6.2) (0.5, 1.8) (3.3, 6.1) (4.4, 6.6)

2 The following datasets each include an *outlier point*. Recall that for our purposes, outliers are points that are a large vertical distance from the line of best fit in relation to the other points. For each dataset below:

i Repeat the previous question and also look carefully at a scatterplot of the data.

ii Note any points that appear to be outliers.

iii Remove the outliers and calculate the correlation and line of best fit again.

a (0.9, 5.2) (6.7, 8.8) (3.9, 1.1) (1.8, 6.7) (4.6, 8.7) (0.8, 3.9)

b (2.9, 3.7) (1.4, 5.6) (4.4, 2.6) (4.3, 5.6) (2.5, 5.1) (6.4, 1.3)

c (4.7, 8.3) (2.4, 2) (5.3, 7.1) (1.3, 2.9) (3.6, 6.4) (5.5, 9.5) (1.8, 0.5) (1.3, 2.9) (6.5, 10.5) (3.6, 2.2) (6.1, 8.9) (2.6, 4.9)

3 Explain why in Question **2**, Pearson's correlation coefficient and the line of best fit seemed less affected by the outlier in part **c** than in the other datasets.

4 This question presents two datasets, each with a repeated point, which can have a significant and possibly overlooked effect on the line of best fit.

Dataset 1: (1, 3) (3, 3) (3, 5) (5, 5) and repeated point (5, 7) with frequency 5

Dataset 2: (5, 7) (6, 8) (4, 7) (1, 4) (3, 4) (4, 4) (1, 2) (8, 7) (0, 2) (5, 5) and repeated point (3, 6) with frequency 5

a For each dataset, calculate the equation of the line of best fit, and Pearson's correlation coefficient:

i with the dataset as given,

ii if the repeated point is only included once in the dataset.

b Comment on the strength of correlation in these examples.

c Comment on the effect of the repeated point on the equations of the line of best fit.

ISBN 978-1-108-76630-2

5 [Pareto charts]

Each of these suggested investigations involves a survey or collection of data, which is then collated into a Pareto chart.

- A preliminary class discussion or sample survey with a small number of respondents may be necessary to decide on categories for the data and chart.
- You will need to consider how to design your survey or data collection so that it is random — that you are, for example, collecting data in multiple locations, or conducting your school survey with respondents from a range of year groups.
- An online survey (search for *online questionnaire*) may be a good way to get a range of data, if the target group all have access to the online survey.
- How many respondents do you think you need to get accurate results?

a In Exercise 15A, a question explored the most common car colours and showed this information on a Pareto chart.

i Design your own experiment where you record the colour of the first 100 cars that pass the school. Let several groups do the experiment simultaneously and check agreement. You may need to discuss the colour categories carefully — it can be a matter of opinion if a car is silver or grey.

ii Do your results agree with the results shown on the chart in Exercise 15A?

b Investigate causes of customer dissatisfaction with the school canteen.

i First discuss in class suitable categories, such as prices, variety and opening hours. A limited number of categories is best.

ii Survey people to find their major cause of dissatisfaction — only one category can be chosen. In a good survey, it is important to survey a range of interest groups, such as all year groups and the teachers.

iii Alternatively, do a survey on causes of customer *satisfaction*.

c Investigate the reasons students are late to school.

i Decide on a number of categories in a class discussion, such as, 'slept in', 'traffic', 'unwell'.

ii Design your questionnaire and show your results on a Pareto chart.

d Do the people at your school use recycling? Investigate the reasons why they don't. This could be a question about recycling at school, at home, or across the spectrum of their lives.

i Decide on categories. Some examples are: unaware of environmental value of recycling, too inconvenient, no recycling bins at school, not aware how to recycle (because of issues such as bin labelling or lack of information on how to recycle old electronic items).

ii Complete your results and draw up a Pareto chart.

e Pareto charts are regarded as one of the *seven basic tools of quality control*, because they are frequently used to investigate questions about quality control, customer satisfaction, and so forth. Investigate what the other six tools are, and see if you can apply a number of them to a quality issue in your school or environment.

6 [Contingency tables]

Contingency tables investigate the interrelationship between different variables in a complex dataset. They work best where there are a limited number of categories (such as male/female and blond/red/brown/black hair).

a Survey a number of students to find their favourite style of music, recording also whether they are male or female.

i Display your results on a contingency table.

ii Investigate whether the favourite style of music appears to differ between males and females in your sample.

ISBN 978-1-108-76630-2

b Consider other surveys that might be explored using contingency tables. Some examples: favourite music and year group, gender and time spent playing video games (group these in categories such 1 hour per week and 2−3 hours per week), gender and time spent on homework, favourite winter sport and favourite summer sport, favourite social media and age.

7 [Scatterplots]
Many of these investigations generate large amounts of data. Excel or another spreadsheet program could be used to generate a scatterplot and to draw the line of best fit.

a Investigate the ages of students at your school and their heights.
- **i** Plot their age in months on the horizontal axis and their height in centimetres on the vertical axis. Do they appear to be correlated? Can you draw a reasonable line of best fit through the data?
- **ii** If your program does not allow you to shift the intersection of the axes, it may be clearer to plot their age in months above 11 years and their height in centimetres above 130 cm.
- **iii** Find the equation of the line of best fit and Pearson's correlation coefficient using technology.

b Measure the lengths of students' forearms and their heights.
- **i** Draw a scatterplot of height x and forearm length y.
- **ii** Construct the line of best fit, and use Pearson's coefficient to decide if there is a good linear correlation between the quantities.
- **iii** Vary this experiment to: leg length and height, stride length and height, foot length and height, arm span and height, hand and forearm length, length of thumb and middle finger, height of student and their father (or mother), circumference of head and height.

Large datasets: The following questions involve the analysis of large sets of data. There are many such datasets available on the internet. For your convenience, some of datasets for these questions may be downloaded from the Cambridge-GO website for this textbook.

No solutions are provided, and it is recommended that the solutions to this exercise should be discussed in class.

8 There is frequent mention in the media of rising sea levels over recent years, because of melting ice sheets and glaciers, and because of the expansion of seawater as it warms. This question uses satellite data provided by NASA to investigate the rise in sea levels since 1995. Further records are also available on the website if you wish to pursue historical records from coastal tide gauge records.

Source: https://climate.nasa.gov/vital-signs/sea-level/

a Download the data files provided on the Cambridge-GO website. Alternatively, press the button marked 'DOWNLOAD DATA' on the NASA webpage given above, and massage the data into a spreadsheet.

The text file on the Cambridge-GO website with extension .txt is provided should you wish to investigate more closely what the various columns in the provided spread-sheet mean and how the data are collected. The file GMSL.csv is provided as a comma-separated spreadsheet that can be read by any spreadsheet program. The file GMSL.xlsx is provided for those running Excel.

b Open GMSL.xlsx. Copy column `C` and column `F` to a new sheet at columns `A` and `B`. Column `C` is the date since 1993, in year and fraction of a year, and column `F` is the sea level measure known as Global Mean Sea Level (GMSL) with a 20-year reference mean taken as 0. Because our concern is the change in sea level, the zero reference point is not important.

c Construct a scatterplot of the data, with date on the horizontal axis and sea level on the vertical level.

d Find the line of best fit.

e Get Excel to display the R^2 value (which is the square of Pearson's correlation). How good is the fit?

f Get Excel to display the equation of the line of best fit. What is the meaning of the gradient here? Is the y-intercept significant?

ISBN 978-1-108-76630-2

g It is probably more meaningful at a glance to adjust the vertical axis to be *change in sea level since the mean height in 1993* (when our data starts).

- **i** Calculate the mean of the sea level values in 1993, storing this in cell `E1`.
- **ii** Add a new column `C`, defined by `=B1-$E$1`, and fill this value down to the rest of the cells in column `C`.
- **iii** Construct a new scatterplot from columns `A` and `C`.

h For further investigation and calculation:

- **i** Eighty percent of the Maldives is less than one metre above sea level. How long will it take if this trend continues for that eighty percent to be under water?
- **ii** Find the height above sea level of your current location, and estimate when it will be under water, if the trend continues.

9 An interesting investigation would be to repeat this question using instead the data from Fort Denison in Sydney, http://www.bom.gov.au/oceanography/projects/ntc/monthly/. The data at this URL go back to 1914, and is regarded worldwide as providing one of the most reliable set of measurements of past sea level. The data are presented quite differently, so you will need to adapt your methods and your questions.

10 Economists make use of linear correlation and regression to forecast a number of economic indicators. This question examines data from the Australian Bureau of Statistics (http://stat.data.abs.gov.au/) on the gross operating profits of the Australian mining industry, collected from 1994–2018.

a Download the spreadsheet CompanyProfits.xlsx from the Cambridge-GO website. Open the tab labelled Data1 and copy Columns `A` and `B` to a new spreadsheet. Delete rows `1–12`, leaving the data from 1995–2018. (The data from 1994 are incomplete, so we shall begin in 1995).

b In cell `C1` enter the formula `=(A1-DATE(1995,1,1))/365+1995`. This converts the date in cell `A1` to a year-and-decimal-fraction-of-a-year format, so that 1/Mar/1995 should convert to 1995.162 because it is 1/6 $\div$ 0.162 of the way through 1995.

c Fill the formula in `C1` down to the rest of the column, to convert all the dates to this more useful format.

d Create a scatterplot with column `C` on the horizontal axis, and column `B` on the vertical axis. It may be easier in your version of Excel if you first copy column `B` to column `D`, so that the data are in the expected order — first x-values, then y-values. Create the scatterplot of columns `C` and `D`.

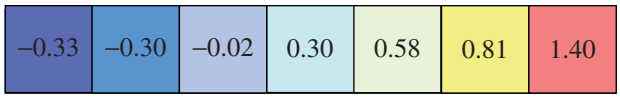
e Determine the R^2 value and the formula for the line of best fit. How well are the data correlated to the straight line model?

f What are the correct units for the company profits on the vertical axis?

g The data do not fit perfectly on a straight line, but is it still a useful model for economic prediction?

11 Astronomers have discovered that for a certain large class of stars, brightness is well correlated with colour. These stars are on the so-called *main sequence*, which omits the very massive red giants and the relatively light white dwarves. Astronomers measure brightness by both apparent magnitude and absolute magnitude, the second of which is adjusted so that the measure of brightness does not depend on a star's distance from Earth. Colour is measured on the BV colour scale, which gives each colour a number. (The initials BV comes from the way the colour is determined by the use of Blue and Visible light filters.)

ISBN 978-1-108-76630-2



BV Colour Indices

a Download the spreadsheet starseq.xlsx from the Cambridge-GO website. This dataset is also provided in a .csv format, which may be useful for those not using Excel.

b Copy column `C` to column `J`. In cell `K2` enter the formula `=B2+5*LOG(D2*10)`. This formula converts the star's apparent magnitude to its absolute magnitude. Fill cell `K2` down to the rest of the cells in column `K`, down to `K6221`. Add the heading *Absolute Magnitude* in `K1`. (You can find out more about this conversion formula if you search for *Convert Apparent Magnitude absolute magnitude* in a web browser).

c Create a scatterplot of the data in columns `J` and `K`.

d Because of the large number of points, the default size that Excel uses to display a point is too large. Click on a point and change the pointer option to a smaller size.

e Adjust the scale on your scatterplot so that the bulk of the data is visible on the plot, say horizontally from -0.5 to 2 and vertically -5 to 20. Excel will allow you to double-click on each axis and set the range of the axis.

f Traditionally, the vertical axis should show the axis flipped, with the negative numbers above the positive. Excel includes an option on the axis to display *Values in reverse order*.

g This dataset will NOT give a correlation coefficient close to -1, because the data include stars off the main sequence.

i Copy columns `J-K` to `O-P`, Replace `P2` by the formula

```
=IF(ABS(B2$+5*LOG($D2*10)-7.5*O2)>3,NA(),B2$+5*LOG($D2*10)).
```

This code is designed to remove all points (stars) vertically well separated from the apparent line of best fit $y = 7.5x$, which seems to model the main sequence. The values that have been eliminated are marked #NA, meaning the value is ignored.

ii Draw a scatterplot of these new data and check if the result gives a good correlation. Remember to adjust the point size and the default axes range, and to flip the direction of the vertical axis.

iii Is it valid to eliminate much of our data in this way?

h Compare your resulting plots with those found online under a search for Hertzsprung–Russell Diagram.

i [Extension] Investigate further the formula relating apparent and absolute magnitude, the BV colour scale, the parsec and arcsec scale for measuring distance, and the 'correct' way to choose stars on the main sequence, by mass.

j [Extension] The magnitude data may be separated into different columns on the basis of the colour index. When Excel constructs the scatterplot, data from separate columns may be coloured independently, illustrating the differing star colours on the main sequence.

12 Weather is notoriously difficult to model, but it is such an important phenomenon that much effort has been applied to modelling its behaviour. The Bureau of Meterology keeps historical and recent data on Australian weather in the data section of its website

http://www.bom.gov.au/climate/data/

Investigate correlation between data on temperature and rainfall for May (or some other month) over a number of years. Choose other variables on the BOM website for similar investigations of correlation.

ISBN 978-1-108-76630-2

Chapter 15 Review

Review activity

- Create your own summary of this chapter on paper or in a digital document.

Chapter 15 Multiple-choice quiz

- This automatically-marked quiz is accessed in the Interactive Textbook. A printable PDF worksheet version is also available there.

Chapter review exercise

1 Write down the mean, median, mode and range of each dataset.

a 4 7 9 2 4 5 8 6 1 4

b 16 17 14 13 18 15 16 15 11

2 A market gardener is testing a new fertiliser on his coriander. Plant beds are set aside to test the new fertiliser. One set of beds have the new fertiliser, while others continue to use the old fertiliser. The total average yield in kilograms in each set of beds is recorded.

Old fertiliser: 1.8 1.4 1.6 2.1 1.9 2.3 1.8 2.1 1.9 1.8

New fertiliser: 2.1 1.6 1.7 2.2 1.9 2.2 1.9 2.2 2.0 1.8

a Find the median and the two quartiles Q_1 and Q_3 for the two datasets.

b Construct a parallel box-and-whisker diagram for the data.

c Do you have any recommendations for the market gardener?

3 **a** Draw up a dot plot of the data in the following list:

4 9 2 6 7 8 2 2 13 5 7 8 6 4 3 6 9 7 4 15 4 6 7 8 9 7 6

b Examine your plot and decide if would you call any of the data points outliers.

c Calculate the interquartile range and determine if the IQR criterion agrees with your assessment in part **b**.

d Draw up a box-and-whisker plot, showing any outliers by a separate dot outside the whiskers.

4 An athlete records his times in seconds running 100 m over a number of events and trials.

11.22 11.43 11.17 11.58 12.10 12.53 11.45 12.04
11.29 13.04 11.31 11.67 12.45 12.14 12.24 11.46

a Calculate the mean and standard deviation for the data using your calculator.

b Construct a frequency table for the data grouped into classes

10.8–11.2, 11.2–11.6, 11.6–12.0, 12.0–12.4, 12.4–12.8, 12.8–13.2.

c Calculate the mean and standard deviation for the grouped data. How does your answer in part **a** compare?

d Draw a frequency histogram for the data.

e Explain why you probably wouldn't want to use a wider class width, such as 0.5.

f Construct a cumulative frequency histogram and ogive for the grouped data.

g Use your cumulative frequency polygon (ogive) to estimate the median for the dataset.

Review

ISBN 978-1-108-76630-2

Review

5 A restaurant wishes to streamline the rate at which it gets food out to its customers at their tables. It divides the evening service into two sittings — the first sitting (arrive 6:00 pm–leave 8:00 pm) and the second sitting (after 8:00 pm). Data are gathered over the course of one night on whether a customer orders an entrée or not.

	first sitting	second sitting	Total
order entrée	45	42	
no entrée	38	28	
Total			

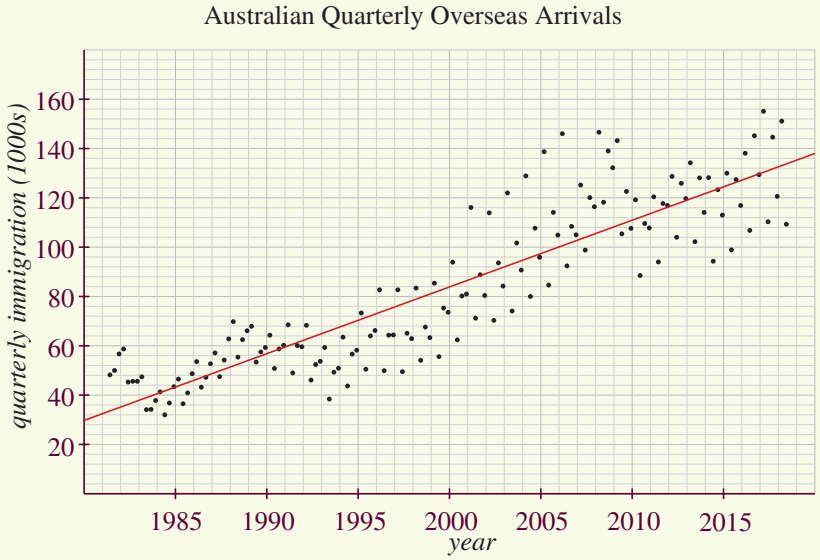
a Copy and complete the table.
b How many people attended the restaurant that night?
c What is the probability that a customer ordered an entrée?
d What percentage of the customers attended the first sitting?
e The manager claims that someone attending the first session is more likely to order an entrée than someone attending the second sitting. Is this correct? Explain.
f It is discovered at the end of the evening that one customer absconded without paying her bill. Given that she ordered an entrée, what is the probability that she was in the first session?
g The next day the restaurant is expecting an exceptionally busy evening for the second sitting, with 90 customers placing a booking. Estimate how many entrées will be ordered.
h Can you suggest why there might be a difference in the ordering habits of those attending the restaurant?
i Why might this survey contain useful information for a restaurant?

6 These small datasets are suitable for calculation by hand or by technology.
i Plot the points and estimate a line of best fit;
ii Calculate Pearson's correlation coefficient and the gradient and y-intercept of the line of best fit, correct to two decimal places.

a (1, 5) (2, 4) (6, 2) (5, 3) (3, 4) (7, 1) (4, 2)
b (7, 5) (0, 1) (2, 2) (6, 4) (3, 3) (5, 3) (1, 2) (4, 4)

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2

7 The scatterplot below shows the immigration of people coming to live in Australia, measured in 1000s. Measurements are made quarterly (March, June, September, December).



a How many people came to live in Australia in the first quarter of 2011?

b Estimate the arrivals in the four quarters of 2000.

c What was the total number of annual arrivals and the average quarterly arrival in 2000?

d Why can it be misleading to examine only the arrivals in one quarter?

e Read the estimate predicted by the line of best fit in the first quarter of 2000.

f What information is missing to make the data useful to someone investigating immigration into Australia?

g The line of best fit shown on the diagram is $y = 2.7x - 5328.8$, correct to one decimal place, where x is the year and y is the quarterly immigration in 1000 s.

- **i** Estimate the immigration rate for the first quarter of 2000 using this formula with $x = 2000.16$ (March 2000).
- **ii** Compare your estimate in part **i** with your estimate from part **e** and explain any discrepancy between these results.
- **iii** Repeat your calculation from part **i** using the formula $y = 2.70633x - 5328.8$.
- **iv** Estimate the immigration rate for the year 2030, by using this formula with $x = 2030$ and then multiplying by 4.
- **v** Use your estimate from part **c** to estimate further the percentage increase in immigration over the 31 years from 2000 to 2030 inclusive.

Mathematics Extension 1 Year 12
Cambridge Maths Stage 6
ISBN 978-1-108-76630-2