Collins

A-level Mathematics

Year 1 and AS
Student Book

Helen Ball
Kath Hipkiss
Michael Kent
Chris Pearce

CONTENTS

Introduction v

1 – Algebra and functions 1: Manipulating algebraic expressions 1

1.1 Manipulating polynomials algebraically 3
1.2 Expanding multiple binomials 6
1.3 The binomial expansion 11
1.4 Factorisation 15
1.5 Algebraic division 16
1.6 Laws of indices 21
1.7 Manipulating surds 24
1.8 Rationalising the denominator 26
Summary of key points 28
Exam-style questions 1 29

2 – Algebra and functions 2: Equations and inequalities 31

2.1 Quadratic functions 32
2.2 The discriminant of a quadratic function 35
2.3 Completing the square 37
2.4 Solving quadratic equations 39
2.5 Solving simultaneous equations 42
2.6 Solving linear and quadratic simultaneous equations 46
2.7 Solving linear inequalities 47
2.8 Solving quadratic inequalities 52
Summary of key points 56
Exam-style questions 2 56

3 – Algebra and functions 3: Sketching curves 58

3.1 Sketching curves of quadratic functions 59
3.2 Sketching curves of cubic functions 63
3.3 Sketching curves of quartic functions 67
3.4 Sketching curves of reciprocal functions 72
3.5 Intersection points 79
3.6 Proportional relationships 82
3.7 Translations 86
3.8 Stretches 91
Summary of key points 97
Exam-style questions 3 98

4 – Coordinate geometry 1: Equations of straight lines 100

4.1 Writing the equation of a straight line in the form $ax + by + c = 0$ 101
4.2 Finding the equation of a straight line using the formula $y - y_1 = m(x - x_1)$ 103
4.3 Finding the gradient of the straight line between two points 105
4.4 Finding the equation of a straight line using the formula $\frac{y - y_1}{y_2 - y_1} = \frac{x - x_1}{x_2 - x_1}$ 109
4.5 Parallel and perpendicular lines 112
4.6 Straight line models 118
Summary of key points 124
Exam-style questions 4 124

5 – Coordinate geometry 2: Circles 126

5.1 Equations of circles 127
5.2 Angles in a semicircle 134
5.3 Perpendicular from the centre to a chord 138
5.4 Radius perpendicular to the tangent 144
Summary of key points 151
Exam-style questions 5 151

6 – Trigonometry 154

6.1 Sine and cosine 155
6.2 The sine rule and the cosine rule 158
6.3 Trigonometric graphs 166
6.4 The tangent function 169
6.5 Solving trigonometric equations 172
6.6 A useful formula 174
Summary of key points 177
Exam-style questions 6 177

7 – Exponentials and logarithms 179

7.1 The function a^x 180
7.2 Logarithms 184
7.3 The equation $a^x = b$ 189
7.4 Logarithmic graphs 191
7.5 The number e 196
7.6 Natural logarithms 201
7.7 Exponential growth and decay 204
Summary of key points 209
Exam-style questions 7 209

8 – Differentiation 212

8.1 The gradient of a curve 213
8.2 The gradient of a quadratic curve 217
8.3 Differentiation of x^2 and x^3 222
8.4 Differentiation of a polynomial 225
8.5 Differentiation of x^n 228
8.6 Stationary points and the second derivative 232
8.7 Tangents and normals 237
Summary of key points 240
Exam-style questions 8 240

9 – Integration 243
9.1 Indefinite integrals 244
9.2 The area under a curve 249
Summary of key points 258
Exam-style questions 9 258

10 – Vectors 260
10.1 Definition of a vector 261
10.2 Adding and subtracting vectors 266
10.3 Vector geometry 269
10.4 Position vectors 278
Summary of key points 283
Exam-style questions 10 283

11 – Proof 285
11.1 Proof by deduction 286
11.2 Proof by exhaustion 289
11.3 Disproof by counter example 294
Summary of key points 298
Exam-style questions 11 298

12 – Data presentation and interpretation 299
12.1 Measures of central tendency and spread 301
12.2 Variance and standard deviation 308
12.3 Displaying and interpreting data 317
Summary of key points 339
Exam-style questions 12 339

13 – Probability and statistical distributions 342
13.1 Calculating and representing probability 343
13.2 Discrete and continuous distributions 350
13.3 The binomial distribution 359
Summary of key points 365
Exam-style questions 13 365

14 – Statistical sampling and hypothesis testing 368
14.1 Populations and samples 369
14.2 Hypothesis testing 374
Summary of key points 388
Exam-style questions 14 388

15 – Kinematics 391
15.1 The language of kinematics 392
15.2 Equations of constant acceleration 395
15.3 Vertical motion 402
15.4 Displacement–time and velocity–time graphs 408
15.5 Variable acceleration 416
Summary of key points 422
Exam-style questions 15 422

16 – Forces 425
16.1 Forces 426
16.2 Newton's laws of motion 429
16.3 Vertical motion 438
16.4 Connected particles 441
16.5 Pulleys 449
Summary of key points 455
Exam-style questions 16 455

Exam-style extension questions 460

Answers* 487
1 Algebra and functions 1: Manipulating algebraic expressions 487
2 Algebra and functions 2: Equations and inequalities 490
3 Algebra and functions 3: Sketching curves 497
4 Coordinate geometry 1: Equations of straight lines 511
5 Coordinate geometry 2: Circles 515
6 Trigonometry 518
7 Exponentials and logarithms 521
8 Differentiation 527
9 Integration 533
10 Vectors 535
11 Proof 540
12 Data presentation and interpretation 544
13 Probability and statistical distributions 549
14 Statistical sampling and hypothesis testing 553
15 Kinematics 557
16 Forces 562

Exam-style extension answers 566

Formulae 581

Glossary 583

Index 587

* Short answers are given in this book, with full worked solutions for all exercises, large data set activities, exam-style questions and extension questions available to teachers by emailing education@harpercollins.co.uk

12 DATA PRESENTATION AND INTERPRETATION

Gender inequality in pay remains high on the agenda in most countries. A map showing the pay gaps is produced by Eurostat, the equivalent of the UK's Office for National Statistics. The figures show that pay levels throughout the whole of Europe differ hugely, with some areas seeing women earning more than men on average, while other areas see men earning more than their female counterparts. Here, 'average' is used to describe the mean earnings.

Data presentation allows you to analyse the data related to the gender pay gap. Using various measures of displacement and location, you can calculate the range and mean pay, and select the best measure to represent your findings about wages. It is also helpful to be able to deal with grouped data, both in calculations and graphically. A graphical representation is often used as it is visual, easier to interpret and allows for comparisons to be made between data sets, making it easier to discover any anomalies. It is also important to be able to find out if a result fits a trend or if it is an outlier, because outliers might skew results and make them unrepresentative.

Another important aspect of data interpretation is making inferences from your findings and using these to come to a conclusion. You can use data to report that differences in the mean wages favour males in certain areas within the UK. This is a powerful conclusion as it leads the reader to understand that men earn more.

LEARNING OBJECTIVES

You will learn how to:

- use discrete, continuous, grouped or ungrouped data
- use measures of central tendency: mean, median, mode
- use measures of variation: variance, standard deviation, range and interpercentile ranges
- use linear interpolation to calculate percentiles from grouped data
- use the statistic $S_{xx} = \sum(x - \bar{x})^2 = \sum x^2 - \frac{\left(\sum x\right)^2}{n}$
- use standard deviation $\sigma = \sqrt{\frac{S_{xx}}{n}}$
- understand and use linear coding
- interpret and draw histograms, frequency polygons, box and whisker plots (including outliers) and cumulative frequency diagrams
- use and interpret the equation of a regression line and understand how to make predictions within the range of values of the explanatory variable and understand the dangers of extrapolation
- describe correlation using terms such as positive, negative, zero, strong and weak.

TOPIC LINKS

In this chapter you will use the skills and techniques that you developed in **Chapter 1 Algebra and functions 1: Manipulating algebraic expressions**. Learning about displaying and interpreting data will help you to analyse data critically, including that in any large data set. You will work with the large data set provided by your chosen exam board in this chapter and in **Chapter 14 Statistical sampling and hypothesis testing**. Knowing how to display data will help in **Book 2, Chapter 13 Statistical distributions**.

PRIOR KNOWLEDGE

You should already know how to:

- interpret and construct tables, charts and diagrams, including frequency tables, bar charts, pie charts and pictograms for categorical data, vertical line charts for ungrouped discrete numerical data, tables and line graphs, and know their appropriate use
- construct and interpret diagrams for grouped discrete data and continuous data (i.e. histograms with equal and unequal class intervals and cumulative frequency graphs), and know their appropriate use
- interpret, analyse and compare the distributions of data sets
- use appropriate graphical representation involving discrete, continuous and grouped data, including box plots
- use appropriate measures of central tendency and spread
- apply statistics to describe a population
- use and interpret scatter graphs of bivariate data, recognise correlation and know that it does not indicate causation
- draw estimated lines of best fit, make predictions, interpolate and extrapolate apparent trends while knowing the dangers of doing this.

You should be able to complete the following questions correctly:

1 A test was given to 50 students and the following marks were awarded:

13	22	41	36	32	26	31	41	31	41
41	14	26	41	41	26	39	39	45	45
34	23	36	23	47	23	47	40	41	29
15	39	36	41	27	46	16	40	19	31
12	27	39	27	28	41	47	28	41	30

a Is this data qualitative or quantitative? If it is quantitative, is it discrete or continuous?

b Calculate the mean, median, mode and range of the data.

c Display the data and make a statement about your findings.

12.1 Measures of central tendency and spread

Types of data

Information obtained from various sources is called **data**. There are two distinct types of data – qualitative and quantitative.

Qualitative data are usually descriptive data, given as categories – such as hair colour, car type or favourite chocolate bar. No numerical value means no numerical meaning can be calculated. Another name for qualitative data is **categorical data**.

Quantitative data, or **numerical data**, are given in numerical form and can be further split down in two categories – discrete and continuous. If all possible values can be listed, the data is **discrete**. Examples of discrete data are shoe size, clothes size and number of marks in a test. **Continuous** data are called continuous because they can be represented at any point on a scale. For instance, height has meaning at all points between any values, e.g. a student's height could be measured as 1.7 m, or 1.67 m, or even 1.668 m. Continuous data can be shown on a number line, and all points on the line have meaning and are different, whereas discrete data can only have a particular selection of values.

KEY INFORMATION

Qualitative data are not numerical, but categorical.

Quantitative data, or numerical data, can be subdivided into discrete and continuous.

Discrete data may only take separate values, for example whole numbers.

Continuous data can be shown on a number line, and all points on the line have meaning and are different from the others.

Example 1

Are the following sets of data qualitative or quantitative, and if quantitative, is the data discrete or continuous?

A Eye colours of students in a class: {2 hazel, 7 blue, 6 green, etc}

B Temperatures of water in an experiment: {34.56, 45.61, 47.87, 56.19, etc}

C Totals of numbers shown on two dice in a board game: {2, 6, 7, 7, 8, 9, 12, etc}

D Numbers of spectators at football matches: {12 134, 2586, 6782, 35 765, etc}

E Favourite types of cereal: {cornflakes, oatflakes, rice crisps, etc}

F Times in seconds between 'blips' of a Geiger counter in a physics experiment: {0.23, 1.23, 3.03, 0.21, 4.51, etc}

G Scores out of 50 in a maths test: {20, 24, 43, 45, 49, etc}

H Sizes of epithelial cells: {1.2×10^{-5} m, 1.21×10^{-5} m, etc}

I Shoe sizes in a class: {6, 7, 7, 7, 8, 9, 10, 10, 11, etc}

Solution

A and E are qualitative data; all the others are quantitative.

For example, different eye colours are separate categories without numerical values, so they are qualitative data.

C, D, G, I – discrete

B, F, H – continuous

Discrete data can only take certain values. Continuous data can take any value within a range.

Measures of central tendency

Measures of central tendency – otherwise referred to as averages or measures of location – are often the first tools for comparing data sets and interpreting data. In your GCSE course you will have encountered three measures of central tendency – the median, the mode and the mean. At AS-level you need to be confident in deciding which measure is the most appropriate to use to answer a specific question.

Median

Data is arranged in numerical order and the **median** is the item of data in the middle. When there is an odd number of data items, the median is simply the middle number. However, when there is an even number of items, the median lies between two middle values, and you use the mean of these two values for the median. Use the formula $\frac{n+1}{2}$ to find the position of the median, where n is the number of data items.

You should use the median for quantitative data, particularly when there are extreme values (values that are far above or below most of the other data) that may skew the outcome.

Mode

The **mode** is the most commonly occurring item of data. It is the item with the highest **frequency**. So for the data set {1, 3, 5, 5, 7}, the mode is 5, as this item appears twice. There may be more than one mode, if more than one item has the highest frequency – for instance {1, 2, 2, 5, 5, 7} has modal values of 2 and 5.

You should use the mode with qualitative data (car models, etc) or with quantitative data (numbers) with a clearly defined mode (or which are bi-modal). The mode is not much use if the distribution is evenly spread, as any conclusions based on the mode will not be meaningful.

Mean

When people are discussing the average, they are usually referring to the **mean**. This is the sum of all of the items of data divided by the number of items of data.

The formula is normally written as $\bar{x} = \frac{\Sigma x}{n}$.

- $\bar{x}$ stands for the mean and is pronounced 'x bar'.
- The Greek letter sigma (Σ) means 'the sum of'. It gives the total of all the data.
- The number of data items is n.

You use the mean for quantitative data (numbers). As the mean uses all the data, it gives a true measure, but it can be affected by extreme values (outliers).

TECHNOLOGY

You can use calculators to calculate many useful statistics. Put your calculator in statistics mode and it may show a table for you to enter your raw data. Use the following set of data to calculate the mean:

2, 4, 6, 8, 10

When a salary increase is being negotiated, the management may well have a different opinion to the majority of workers. The following figures were collected to compare the salaries in a small fast-food company:

£3500, £3500, £3500, £3500, £4500, £4500, £4500, £8000, £10 000, £10 000, £10 000, £12 000, £12 000, £18 000, £30 000

Who do you think earns what? The median salary is £8000, the modal salary is £3500 and the mean salary is £9166.67. These figures can be used in a variety of ways, but which is the most appropriate measure? If you were the manager, you might quote the mean of £9166.67, but fewer than half of the employees earn this amount.

Workers leading the pay negotiations who want to criticise the current wage structure may choose to quote the mode (£3500), as this is the lowest average. This would highlight issues in the wage structure.

The mean takes account of the numerical value of every item of data. It is higher because of the effect of the £30 000 salary, which is an extremely large value in comparison to the others. The median and mode are not affected by extreme values.

The mean is a good measure to use as you use all your data to work it out. It can, however, be affected by extreme values and by distributions of data which are not symmetrical. The median is not affected by extremes so is a good measure to use if you have extremes in your data, or if you have data which isn't symmetrical. The mode can be used with all types of data, but some data sets can have more than one mode, which isn't helpful at all.

KEY INFORMATION

The median is the middle value when the data items are placed in numerical order.

The mode is the most common or frequent item of data.

The mean ($\bar{x}$) is found by adding the data values together and dividing by the number of values: $\bar{x} = \frac{\sum x}{n}$.

Using the large data set 12.1

a Calculate the median and the mean of one category in one location. Comment on your findings.

b Repeat for the same category but in another location. Does the location affect the mean and the median? Comment critically on your findings.

c State any assumptions you have made whilst using your chosen category.

Measures of spread

Measures of spread show how spread out, or scattered, data items are. In your GCSE course you will have encountered two measures of spread – the range and the interquartile range.

Range

The simplest measure of spread is the difference between the highest and smallest data items, known as the **range**. This is straightforward to calculate, but can be highly sensitive to extreme values.

Range = highest value – lowest value

Interquartile range

A more trustworthy measure of spread is the range in the middle half of the data. The upper **quartile** is the median of the upper half of the data, and the lower quartile is the median of the lower half of the data.

Quartiles split the data into quarters. If the data items are arranged in numerical order, the lower quartile is one quarter of the way through the data and the upper quartile is three quarters of the way through.

See **Section 12.3** for examples of the interquartile range in the context of box and whisker plots.

The **interquartile range** measures the range of the middle 50% of the data.

In **Section 12.2** you will encounter another measure of spread – the standard deviation.

Using frequency tables to calculate measures of central tendency and spread

When data contains items that are repeated it is easier to record them using a frequency table. You used frequency tables for categorical data in your GCSE course – now you will use them to deal with numerical data. When you use frequency tables you can calculate measures of spread more quickly.

Example 2

Here are the scores when a four-sided spinner was spun repeatedly:

$$\{3, 4, 3, 3, 1, 2, 1, 2, 2, 2, 2, 3, 4, 2, 3, 4, 2, 2, 1, 3\}$$

The scores are recorded in the frequency table. Calculate the mean of the data set.

Score	Frequency
1	3
2	8
3	6
4	3

Solution

The mean of the data shown in the frequency table above can be written as $\bar{x} = \frac{\sum fx}{\sum f}$.

The frequencies are added to find the total number of data items. Representing the score by x and its frequency by f, the calculation of the mean starts by multiplying each score by its corresponding frequency to find the total value of all the scores.

Score, x	Frequency, f	fx
1	3	$1 \times 3 = 3$
2	8	$2 \times 8 = 16$
3	6	$3 \times 6 = 18$
4	3	$4 \times 3 = 12$
Total	$\sum f = 20$	$\sum fx = 49$

The total score is 49 and there are 20 data items. Dividing the total by the number of data items gives the mean.

Add up the two columns, f and fx, and use their totals to calculate the mean.

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{49}{20} = 2.45$$

Finding the median using a frequency table is straightforward as the data is already ordered. There are 20 numbers in **Example 2**, so if the data were written out in a line the median would lie between the 10th and 11th values (20 + 1 = 21, then divide by 2 to give 10.5). The next step is to find the class that contains the 10.5th value.

Score, x	Frequency, f	Cumulative frequency
1	3	3 (not enough – cumulative frequency needs to be between 10 and 11)
2	8	3 + 8 = 11 (must be this score as the combined frequencies are more than 10.5)

The median is within the data associated with a score of 2. Therefore the median score is 2.

You can identify the mode as it is the score with the highest frequency. From the data in **Example 2** the highest frequency is 8, so the mode has to be 2.

Discrete grouped data

Grouped frequency tables are used when data is widely spread. The downside to this is that raw data is lost, since now you will only know the frequency of each grouping. Consider the following data on gross salaries for 427 professions, drawn from the Office for National Statistics' 2010 salary tables (*Annual Survey of Hours and Earnings*).

Salary, £	Frequency
0–9999	21
10 000–19 999	129
20 000–29 999	172
30 000–39 999	74
40 000–49 999	24
50 000–59 999	4
60 000–69 999	1
70 000–79 999	1
80 000–89 999	0
90 000–99 999	1

Immediately, you can see that the modal wage is between £20 000 and £29 999, as this grouping has the highest frequency, and that salaries above £60 000 are unusual.

TECHNOLOGY

Repeat the above calculation using a frequency table on your calculator. You may have to go into Set-up on your calculator and select 'Frequency on' before you start.

KEY INFORMATION

For discrete data, the groupings should not overlap.

Continuous grouped data

The PTA have given some money for 20 hockey players from a Year 7 school team to get their new team uniforms. Their heights in metres are:

1.45	1.48	1.46	1.52	1.46	1.61	1.60	1.51	1.55	1.56
1.61	1.64	1.53	1.51	1.48	1.70	1.70	1.62	1.45	1.50

Height, *h* (metres)	Frequency, *f*
$1.45 \leqslant h < 1.50$	6
$1.50 \leqslant h < 1.55$	5
$1.55 \leqslant h < 1.60$	2
$1.60 \leqslant h < 1.65$	5
$1.65 \leqslant h < 1.70$	0
$1.70 \leqslant h < 1.75$	2
Total	20

KEY INFORMATION

Measurements are frequently recorded to the nearest unit. So a height of 1.65 m could actually be in the range $1.645 \leqslant h < 1.655$.

Estimating the mean from grouped data

The reason this is now an estimate is because you are estimating the middle value of each class by assuming that the data is evenly distributed throughout each interval.

Salary (£)	Frequency, *f*	Estimate of *x*	*fx*
0–9999	21	5000	105 000
10 000–19 999	129	15 000	1 935 000
20 000–29 999	172	25 000	4 300 000
30 000–39 999	74	35 000	2 590 000
40 000–49 999	24	45 000	1 080 000
50 000–59 999	4	55 000	220 000
60 000–69 999	1	65 000	65 000
70 000–79 999	1	75 000	75 000
80 000–89 999	0	85 000	0
90 000–99 999	1	95 000	95 000
Total	$\sum f = 427$		$\sum fx = 10\,465\,000$

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{10\,465\,000}{427} = £24\,508.20\ (2\text{ d.p.})$$

Exercise 12.1A

Answers page 544

1 State whether the following data are discrete or continuous:

a daily rainfall in Lincoln

b monthly texts you send on your mobile

c the number of burgers sold in a fast-food restaurant

d the duration of a marathon

e the ages of the teachers in your school.

2 Classify the following as qualitative or quantitative, discrete or continuous:

a gender

b height

c GCSE grades in maths

d examination score in maths

e waist size

f whether people are car owners or not

g weekly self-study time.

CM 3 Write down two quantitative variables about your class. Identify each variable as discrete data or continuous data.

CM 4 The numbers of visits to a library made by 20 children in one year are recorded below.

0, 2, 6, 7, 5, 9, 12, 43, 1, 0, 45, 2, 7, 12, 9, 9, 32, 11, 36, 13

a What is the modal number of visits?

b What is the median number of visits?

c What is the mean number of visits?

d Comment on the best measure of central tendency to use and why.

5 Fiona records the amount of rainfall, in mm, at her home, each day for a week. The results are:

2.8, 5.6, 2.3, 9.4, 0.0, 0.5, 1.8

Fiona then records the amount of rainfall, x mm, at her house for the following 21 days. Her results are:

$$\sum x = 84.6\,\text{mm}$$

a Calculate the mean rainfall over the 28 days.

b Fiona realises she has transposed two of her figures. The number 9.4 should be 4.9 and the 0.5 should be 5.0. She corrects these figures. What effect will this have on the mean?

CM 6 An employee has to pass through 8 sets of traffic lights on her way to work. For 100 days she records how many sets of lights she gets stopped at. Here are her results:

Number of times stopped	Number of journeys
$\leqslant 1$	3
2	5
3	11
4	21
5	22
6	17
7	14
8	7

Find the median, mode and mean number of times stopped at traffic lights and comment on the best measure of central tendency to represent the data.

7 During a biological experiment, fish of various breeds were measured, to the nearest cm, one year after being released into a pond. The lengths were recorded in the following table:

Length, x (cm)	Frequency
$7.5 \leqslant x < 10$	30
$10 \leqslant x < 15$	70
$15 \leqslant x < 20$	100
$20 \leqslant x < 30$	80
$30 \leqslant x < 35$	40

Calculate an estimate of the mean length of one of these fish.

8 The following data was collected but some information was missed out. Complete the missing parts and confirm the estimate of the mean.

Height (cm)	Frequency	
$100 < h \leqslant 120$	5	
$120 < h \leqslant 140$	4	
	12	1800
$160 < h \leqslant 180$	13	
$180 < h \leqslant 200$	8	
		6600

Estimate of the mean = 157.14

12.2 Variance and standard deviation

One limitation with quartiles and the interquartile range is that they do not take all the data items into account. Variance and standard deviation are measures that involve determining the spread of *all* the data items.

Consider a small set of data: $\{1, 2, 3, 4, 5\}$

The mean of this data is 3.

The **deviation** is the difference between each data item and the mean, usually notated as $x - \bar{x}$.

The set of deviations for this set of data is:

$1 - 3 = -2$

$2 - 3 = -1$

$3 - 3 = 0$

$4 - 3 = 1$

$5 - 3 = 2$

Adding up the deviations:

$-2 + -1 + 0 + 1 + 2 = 0$

Stop and think Why do the deviations add up to zero?

The deviations, $x - \bar{x}$, give a measure of spread. Combining the deviations, by squaring each deviation and then adding them together, gives the sum of their squares, notated as S_{xx}.

The sum of the squares is usually written

$$S_{xx} = \sum(x - \bar{x})^2 = \sum x^2 - n\bar{x}^2$$

The second formula may be easier to work with if you are given raw data.

To derive this formula:

$$\begin{aligned} S_{xx} &= \sum(x - \bar{x})^2 \\ &= \sum\left(x^2 - 2x\bar{x} + \bar{x}^2\right) \\ &= \sum x^2 - \sum 2x\bar{x} + \sum \bar{x}^2 \\ &= \sum x^2 - 2\bar{x}\sum x + \sum \bar{x}^2 \\ &= \sum x^2 - 2n\bar{x}^2 + n\bar{x}^2 \\ &= \sum x^2 - n\bar{x}^2 \end{aligned}$$

Taking another small set of data, {1, 4, 5, 6, 14}:

$$\begin{aligned} S_{xx} &= (1 - 6)^2 + (4 - 6)^2 + (5 - 6)^2 + (6 - 6)^2 + (14 - 6)^2 \\ &= -5^2 + -2^2 + -1^2 + 0^2 + 8^2 = 94 \end{aligned}$$

The mean of the sum of the squares is a measure of spread called the **variance**, $\frac{\sum(x - \bar{x})^2}{n}$.

The square root of the variance is called the **standard deviation**, $\sqrt{\frac{\sum(x - \bar{x})^2}{n}}$, which is usually denoted by the symbol σ.

Therefore the variance for this set of data is $\frac{94}{5} = 18.8$ and the standard deviation is $\sqrt{18.8} = 4.34$.

You can divide by n or $(n - 1)$ when calculating either the variance of a population or an estimate for the population from sample data. Either divisor will be accepted unless a question specifically requests an unbiased estimate of a population variance, in which case you would use $(n - 1)$.

An easier formula, which involves doing fewer subtractions, is:

$$\sigma = \sqrt{\frac{\sum x^2}{n} - \bar{x}^2}$$

Standard deviation is especially useful when comparing different sets of data and when analysing the position of an item of data in a population. An advantage of standard deviation is that it uses all of the data. A disadvantage is that it takes longer to calculate than other measures of spread.

Example 3

Calculate the variance and standard deviation for this small data set: {2, 3, 5, 8, 13}.

Solution

Calculate the mean.

$$\bar{x} = \frac{2+3+5+8+13}{5} = 6.2$$

Determine the deviations from the mean.

$$= 6.2 - 2, 6.2 - 3, 6.2 - 5, 6.2 - 8, 6.2 - 13$$

$$= 4.2, 3.2, 1.2, -1.8, -6.8$$

Square these deviations.

$$(x - \bar{x})^2 = 17.64, 10.24, 1.44, 3.24, 46.24$$

Find the mean of these squared deviations, or the variance.

$$\Sigma(x - \bar{x})^2 = 17.64 + 10.24 + 1.44 + 3.24 + 46.24 = 78.8$$

$$= \frac{78.8}{5} = 15.76$$

Find the standard deviation by taking the square root of the variance.

$$\sqrt{\frac{\Sigma(x - \bar{x})^2}{n}} = \sqrt{15.76} = 3.97$$

Stop and think How could you use the alternative formula in the above example?

Calculating variance and standard deviation using frequency tables

Example 4

The table below shows the numbers of pieces of fruit eaten by sixth form students in one day. Find the standard deviation of the number of pieces of fruit eaten.

Number, x	Frequency, f
1	2
2	12
3	45
4	114
5	41
6	16
Total	230

Solution

First multiply each number of pieces of fruit by its corresponding frequency to give you fx, then sum to find out the total number of pieces of fruit, $\sum fx$.

Next, square each value (x^2) and then sum to find out the total $\sum x^2$.

Finally, multiply x^2 by the frequency to give fx^2, then sum to find out the total, $\sum fx^2$.

Number, x	Frequency, f	fx	x^2	fx^2
1	2	2	1	2
2	12	24	4	48
3	45	135	9	405
4	114	456	16	1824
5	41	205	25	1025
6	16	96	36	576
	230	918	91	3880

For discrete frequency distributions, the formulae are:

$$\sigma = \sqrt{\frac{\sum f(x-\bar{x})^2}{\sum f}} \quad \text{or} \quad \sigma = \sqrt{\frac{\sum fx^2}{\sum f} - \bar{x}^2}.$$

Σf, the sum of the frequencies, is used instead of n.

For grouped frequency distributions, you use the midpoint of the group for x.

$$\text{Mean} = \frac{918}{230}$$

$$\text{Variance} = \frac{3880}{230} - \left(\frac{918}{230}\right)^2 = 0.939$$

$$\text{Standard deviation} = \sqrt{0.939} = 0.969$$

$$\sigma = 0.969$$

TECHNOLOGY

Check that you can work out the mean and standard deviation for this example using the frequency table option on your calculator.

Exercise 12.2A

Answers page 544

1 Calculate the mean and standard deviation of the following.

a 2, 4, 6, 8, 10, 12, 14

b 50, 60, 70, 80, 90

c 12, 15, 18, 16, 7, 9, 14

Check your results using a calculator.

TECHNOLOGY

In all these questions you can use the statistical buttons on a calculator.

2 For each of the following sets of data, find the mean and standard deviation.

a 2, 2, 4, 4, 4, 5, 6, 6, 8, 9

b 13.1, 20.4, 17.4, 16.5, 21.0, 14.8, 12.6

CM **3** Here are the shoe sizes of a class of Year 7 students:

Size	3	4	5	6	7	8	9
Frequency	1	0	8	14	6	2	1

a Calculate the mean and standard deviation shoe size.

b What can you say about the mean shoe size compared with the median and modal sizes?

4 $\sum x = 27$, $\sum x^2 = 245$, $n = 3$

Find the mean and standard deviation of the data.

5 The lengths of time, t minutes, taken for a bus journey is recorded on 15 days and summarised by

$$\sum x = 102,\ \sum x^2 = 1181$$

Find the mean and variance of the times taken for this bus journey.

6 For a set of 20 data items, $\sum x = 12$ and $\sum x^2 = 144$.
Find the mean and standard deviation of the data.

CM **7** Mrs Moat has a choice of two routes to work. She times her journeys along each route on several occasions and the times in minutes are given below.

Town route	15	16	20	28	21
Country route	19	21	20	22	18

a Calculate the mean and standard deviation for each route.

b Which route would you recommend? Give a reason.

CM **8** A machine is supposed to produce ball bearings with a mean diameter of 2.0 mm.
A sample of eight ball bearings was taken from the production line and the diameters measured.
The results in millimetres were:

2.0, 2.1, 2.0, 1.8, 2.4, 2.3, 1.9, 2.1

a Calculate the mean and standard deviation of the diameters.

b Do you think the machine is set correctly?

Data cleaning

Any large data set is likely to be missing some data. If values are left blank, it is likely that the software package you use to work out statistics or draw graphs may not work, or may interpret missing values as zero. This will cause errors in your calculations, leading to misleading outcomes. Often, unusual data entries are classed as **outliers**. The process of dealing with missing and unusual values in data is called **data cleaning**.

Data cleaning involves identifying and then removing invalid data points from a data set. You can then calculate your statistics and draw your graphs using the remaining data. If data has been inputted as ‘n/a’, most packages will not understand this, so errors will appear in your calculations. You need to use your judgment to decide which points are valid and which are not.

The points to be cleaned are generally either missing data points or outliers. One way of detecting points which do not fit the trend is to plot the data and then inspect for points that lie far away from the trend.

The importance of reliable data in any statistical analysis cannot be over-emphasised. This is one reason why larger data sets are generally more reliable.

Using the large data set 12.2

The large data set has a lot of raw data so should be cleaned before you make any analysis. You need to be able to use spreadsheets to perform the calculations and produce the diagrams.

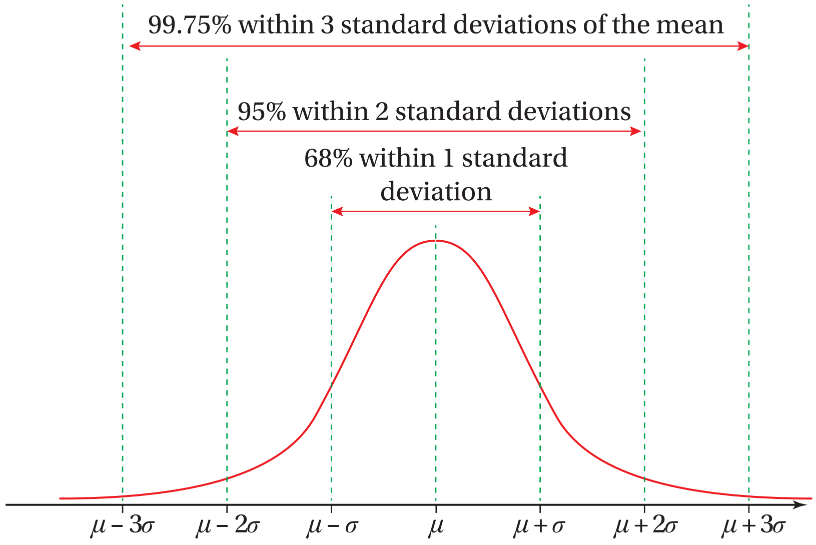
a Check through and clean the data for the category you used in **Using the large data set 12.1**, giving valid reasons for your omissions as you progress through the data.

b Does having clean data change the median and mean you calculated in **12.1**? Explain your reasons.

Standard deviation and outliers

Data sets may contain extreme values and you need to be able to deal with these. Many data sets are samples drawn from larger populations which are normally distributed. In cases like this, approximately:

- 68% of data items lie within 1 standard deviation of the mean
- 95% of data items lie within 2 standard deviations of the mean
- 99.75% of data items lie within 3 standard deviations of the mean.


99.75% within 3 standard deviations of the mean
95% within 2 standard deviations
68% within 1 standard deviation
$\mu-3\sigma$
$\mu-2\sigma$
$\mu-\sigma$
μ
$\mu+\sigma$
$\mu+2\sigma$
$\mu+3\sigma$

You will cover normal distribution in **Book 2, Chapter 13 Statistical distributions**. However, you should know that if a value is more than a certain number of standard deviations away from the mean it should be treated as an outlier. In your course, any value more than 2 standard deviations away from the mean will be treated as an outlier.

Example 5

The times of journeys to work, in minutes, were recorded by one person for two weeks. The results are:

51, 65, 48, 41, 56, 65, 45, 42, 41, 87

Determine whether any of the data items are outliers.

Solution

The mean is $\frac{\sum x}{n} = 54.1$ minutes.

The standard deviation is $\sqrt{\frac{\sum(x-\bar{x})^2}{n}} = 13.94$ minutes.

$$54.1 + 2(13.94) = 81.98 \text{ minutes}$$

$$54.1 - 2(13.94) = 26.22 \text{ minutes}$$

The longest time to get to work was 87 minutes. You should class this as an outlier because $81.98 < 87$ minutes.

Using the large data set 12.3

a Calculate the mean and standard deviation of one of the categories in your large data set.

b Describe how you would clean the data and how this may affect the mean and standard deviation.

c Clean the data for your category and evaluate any changes it has made to the mean and standard deviation.

Linear coding

Linear coding can be used to make certain calculations easier. It is used to simplify messy arithmetic in calculations and also to convert between different units.

The data set {12, 3, 8, 7, 11} has mean $\bar{x} = 8.2$. Using a calculator you can find the standard deviation, $\sigma = 3.19$ (2 d.p.).

Add 10 to all the data points: the data set is now {22, 13, 18, 17, 21}. You can work out the new mean and the new standard deviation using a calculator: $\bar{x} = 18.2$, $\sigma = 3.19$ (2 d.p.). The data points have all increased by 10 but the spread in the data has remained the same.

If the original data set is multiplied by 7, you would get {84, 21, 56, 49, 77}. Using a calculator you can see that the mean is $\bar{x} = 57.4$ and the standard deviation is $\sigma = 22.31$ (2 d.p.). This time the size and the spread in the data have increased.

The following table summarises what happens to the mean and standard deviation of x when various transformations are applied to the data set {2, −2, −8, 0, 2.3}.

	x	$x + 3$	$x - 2$	$10x$	$\frac{x}{2} + 2$
	2	5	0	20	3
	–2	1	–4	–20	1
	–8	–5	–10	–80	–2
	0.8	3.8	–1.2	8	2.4
	2.3	5.3	0.3	23	3.15
Mean	–1.0	2.0	–3.0	–9.8	1.5
Standard deviation	3.8	3.8	3.8	38.24	1.91

Stop and think What do you notice about the mean and standard deviation in each case?

If an number is added to or subtracted from a data set, the mean increases or decreases by that amount, but the standard devaition is not affected as the spread in the data remains the same.

If the data set is multiplied or divided by a number, the mean increases or decreases by that factor, as does the standard deviation.

Notice that the mean and standard deviation change by the magnitude of the x coefficient. The mean also changes by the number being added or subtracted from it, but the standard deviation remains unchanged when a number is simply added to or subtracted from x.

Five people were asked how far they worked from home. Here are their results:

x	24	28	30	33	35

$\bar{x} = 30$ and $\sigma = 3.8$

If you add 40 miles onto each person’s journey, the mean will increase but the standard deviation remains unchanged because the spread is still the same.

Y	64	68	70	73	75

$\bar{Y} = 70$ and $\sigma = 3.8$

There is no multiplier to affect the standard deviation.

Similarly, if you subtracted 20 miles from everyone’s journey the mean would decrease but the standard deviation would remain the same.

z	4	8	10	13	15

$\bar{z} = 10$ and $\sigma = 3.8$

There is no multiplier to affect the standard deviation.

If all the journey times doubled, the mean would double but so would the standard deviation, as the data would now be more spread out.

a	48	56	60	66	70

$\bar{a} = 60$ and $\sigma = 7.7$

The multiplier affects the standard deviation.

If the journey time doubled and then 20 more miles were added on, the mean would increase by a multiple of 2 plus 20 for everyones time:

b	68	76	80	86	90

$\bar{b} = 80$ and $\sigma = 7.7$

The multiplier affects the standard deviation, but the addition does not.

In general if x is coded to $Y = ax + b$:

Mean = $\bar{x}$	Coded mean = $a\bar{x} + b$
Standard deviation = σ	Coded standard deviation = $a\sigma$

Example 6

A data set has been coded using $Y = \frac{x}{12}$. The coded standard deviation is 1.41.

Find the standard deviation of the original data.

Solution

This example shows x being divided by 12. This will affect the standard deviation. The original data was coded, so to find the original you need to multiply by 12.

$$1.41 \times 12 = 16.92$$

Example 7

A data set has been coded using $Y = x - 2.7$. The coded standard deviation is 3.641.

Find the standard deviation of the original data.

Solution

This time a number is subtracted from the x values. This does not affect the standard deviation as the spread remains the same.

The original standard deviation was 3.641, as the standard deviation does not change.

Example 8

A data set has been coded using $Y = \frac{x}{7} + 1.79$. The coded standard deviation is 12.342.

Find the standard deviation of the original data.

Solution

1.79 has been added to the data which is then divided by 7. Adding 1.79 has no effect but the division by 7 does affect it. So you need to multiply this by 7.

The original standard deviation was 86.39.

Using the large data set 12.4

Using a category within your large data set, write a linear code and use this to find the mean and standard deviation of the data.

Exercise 12.2B

Answers page 545

1 The mark, x, scored by each student who sat an AS Mathematics exam is coded using

$$Y = 1.4x - 15$$

The coded marks have a mean of 60.2 and a standard deviation of 4.5.

Find the mean and the standard deviation of x.

2 A system is used in schools to predict students, A-level grades using their GCSE results. The GCSE score is g and the predicted A-level score is a. For students taking their maths GCSE in 2016, the coding equation was given by:

$$a = 3.1g - 9.53$$

In 2017 there are 97 students in their second year. Their GCSE scores are summarised as:

$$\Sigma g = 418.3 \text{ and } \Sigma g^2 = 2312.19$$

a Find the mean and standard deviation of the GCSE scores.

b Find the mean and standard deviation of the predicted A-level scores.

M **3** On her summer holiday, Farida recorded the temperature at noon each day for use in a statistics project. The values recorded, in degrees Fahrenheit, were as follows (correct to the nearest degree):

47, 59, 68, 62, 49, 67, 66, 73, 70, 68, 74, 84, 80, 72

a Find the mean and standard deviation of Farida's data.

b The formula for converting temperatures from f degrees Fahrenheit to c degrees Celsius is $c = \frac{5}{9}(f - 32)$
Use this formula to estimate the mean and standard deviation of the temperatures in degrees Celsius.

4 The mean weekly cheese consumption per household is 139 grams. The standard deviation is 5.7 grams. Assuming 1 ounce is approximately equal to 28.35 grams, calculate the mean and standard deviation in ounces.

12.3 Displaying and interpreting data

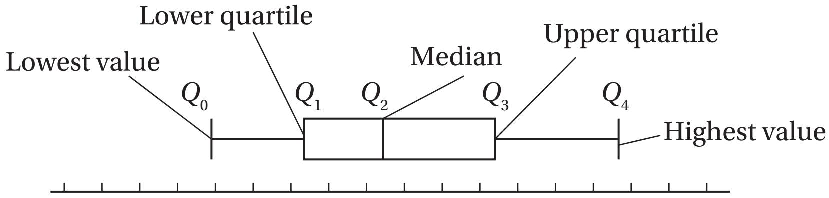
Different visual representations can be used to present and interpret data – including **box and whisker plots**, cumulative frequency diagrams, **histograms** and scatter diagrams. You will remember these types of representations from your GCSE course – at A-level you need to be able to make judgments about which method of presentation is most appropriate for particular data sets and requirements.

Other visual representations that you may come across are dot plots, which are similar to bar charts but with stacks of dots in lines to represent frequency, and frequency charts, which resemble histograms with equal width bars but the vertical axis is frequency.

Box and whisker plots

The median and quartiles can be displayed graphically by means of a box and whisker plot, sometimes just referred to as a box plot. This can be used to compare sets of data.

The diagram shows a generic box and whisker plot. A box is drawn between the lower and upper quartiles and a line is drawn in the box showing the position of the median. Whiskers extend to the lowest value and to the highest value. If an outlier is recorded but not used, this is displayed as a cross.



The lowest item of data is referred to as Q_0, the lower quartile as Q_1, the median as Q_2, the upper quartile as Q_3 and the highest data item as Q_4.

The range is the difference between the highest and smallest values in the data set.

$$\text{Range} = Q_4 - Q_0$$

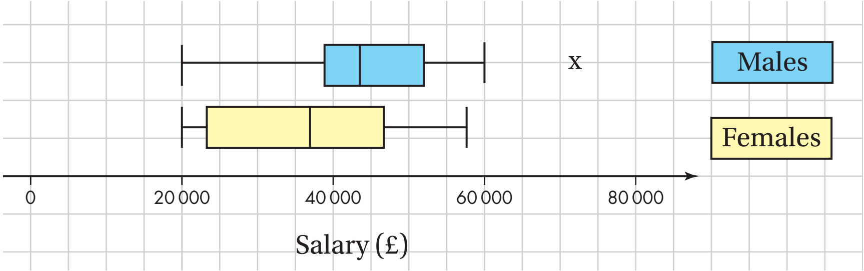
The mid-range is the value halfway between the upper and lower extreme values. It is easy to calculate but is only useful if the data is reasonably symmetrical and free from outliers.

$$\text{Mid-range} = \frac{(Q_4 + Q_0)}{2}$$

As you saw in **Section 12.1**, the upper quartile is the median of the upper half of the data, the lower quartile is the median of the lower half of the data and the interquartile range measures the range of the middle 50% of the data.

$$\text{Interquartile range} = Q_3 - Q_1$$

It is useful to be able to compare sets of data using their box and whisker plots. The following two box and whisker plots represent the annual salaries of 40-year-olds in 2010.



The ranges of salary are similar, shown by the distance between the whiskers. The males have a smaller interquartile range than the females, shown by the size of the boxes, which suggests that the majority of the pay is less spread out for males. The median and quartiles for males are higher than those for females, so on average males earn more than females at age 40. The cross on the plot for the males indicates an anomaly or rogue data item which doesn't seem to fit the trend – an outlier.

Identifying outliers using quartiles

The interquartile range is simply the difference between the quartiles, or $Q_3 - Q_1$. An outlier can be identified as follows (IQR stands for interquartile range):

- any data which are 1.5 × IQR below the lower quartile
- any data which are 1.5 × IQR above the upper quartile.

Example 9

The following data set lists the heights of employees.

1.45	1.48	1.46	1.52	1.46	1.61	1.60	1.51	1.55	1.56
1.61	1.64	1.53	1.51	1.48	1.70	1.70	1.62	1.45	1.50

Are there any outliers in this data set?

Solution

Q_1 = lower quartile = 1.48

Q_2 = median = 1.525

Q_3 = upper quartile = 1.61

The interquartile range (IQR) is 1.61 – 1.48 = 0.13 m.

$$1.5 \times \text{IQR} = 1.5 \times 0.13 = 0.195$$

1.5 × IQR below the lower quartile = 1.48 – 0.195 = 1.285 m, so there are no low outliers as everyone is taller than this.

1.5 × IQR above the upper quartile = 1.61 + 0.195 = 1.805 m, so there are no high outliers as everyone is shorter than this.

Cumulative frequency diagrams

These are used when the data is grouped. It is easy to find the median and quartiles using **cumulative frequency** diagrams.

Here is a grouped frequency table showing the time spent queuing for rides at a theme park:

Time, t (minutes)	$0 < t \leqslant 5$	$5 < t \leqslant 10$	$10 < t \leqslant 20$	$20 < t \leqslant 30$	$30 < t \leqslant 60$
Frequency	3	24	41	17	15

It shows, for example, that 41 people queued for between 10 and 20 minutes.

You can now work out the cumulative frequencies by adding up the frequencies as you go along.

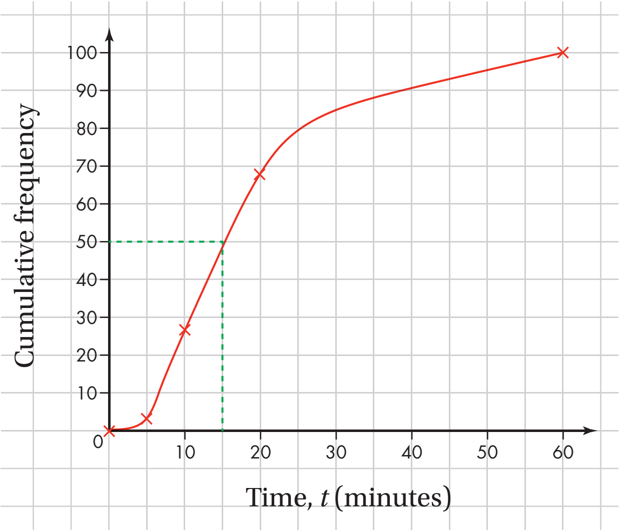
Time, t (minutes)	$0 < t \leqslant 5$	$0 < t \leqslant 10$	$0 < t \leqslant 20$	$0 < t \leqslant 30$	$0 < t \leqslant 60$
Frequency	3	24	41	17	15
Cumulative frequency	3	27	68	85	100

Notice that the time intervals have also changed as the frequencies have accumulated from the previous time.

You can now plot these cumulative frequencies on graph paper, making sure that you plot the cumulative frequencies at the upper bound of the interval, that is, at (5, 3), (10, 27), (20, 68), (30, 85) and (60, 100).

You should also plot the point (0, 0).

Join the points with a smooth curve (not straight line segments) to create the graph shown here:



Find the median by drawing a line from 50 (half of 100) on the cumulative frequency axis to the graph, then down to the time axis: here you can see that the median is 15 minutes.

TECHNOLOGY

You could do this quickly and accurately using a graph-drawing software package.

Estimating percentiles from grouped data

Percentiles give the value below which a given percentage of observations fall. They are often used in the reporting of scores. For example, if you achieved a score in the 40th percentile, this would mean your score is higher than 40% of the other scores.

You have already encountered some specific percentiles:

- the 25th percentile, known as the first quartile (Q_1)
- the 50th percentile, known as the median or second quartile (Q_2)
- the 75th percentile, known as the third quartile (Q_3).

If data is only available as a grouped frequency distribution, then it is not possible to find the exact values of the median, quartiles or other percentiles. However, it is possible to estimate values using linear **interpolation**.

Interpolation is the process of finding a value between two points.

To find the 50th percentile (the median) you need to know the total. There are 427 data items in the data set on gross salaries, so you need to find the 214th data item.

$\frac{427+1}{2}$ = 214th data item

Salary (£)	Frequency	Cumulative frequency
0–9999	21	21
10 000–19 999	129	150
20 000–29 999	172	322
30 000–39 999	74	396
40 000–49 999	24	420
50 000–59 999	4	424
60 000–69 999	1	425
70 000–79 999	1	426
80 000–89 999	0	426
90 000–99 999	1	427

To do this, create cumulative frequencies from the data in the table. This shows clearly that the 214th data item lies in the class interval 20 000–29 999. However, you want to estimate what the value of the 50th percentile is, not which class it is in.

To do this you need to find out where in the class the 214th value is. 214 – 150 = 64, so it is the 64th item in this class interval (which contains 172 items). As a fraction, it is $\frac{64}{172}$ within the class, and the class is 10 000 wide.

So in order to estimate the 50th percentile:

$$\frac{64}{172} \times 10\,000 = 3720.93$$

This value is the median within the class.

You must then add this value onto the lower bound of the class that the 50th percentile lies in.

£20 000 + £3720.93 = £23 270.93

If data is only available as a grouped frequency distribution, then it is not possible to find the median using exact values of the quartiles or other percentiles. However, it is possible to estimate values for these.

KEY INFORMATION

Use this formula to find a percentile:

$\left(\frac{\text{position in class}}{\text{class frequency}}\right) \times$ class width + lower boundary of class

Suppose you were asked to estimate the 45th percentile of the data set shown below.

Age of employees	Frequency
$16 \leqslant a < 25$	2
$25 \leqslant a < 30$	6
$30 \leqslant a < 40$	10
$40 \leqslant a < 50$	8
$50 \leqslant a < 65$	5
over 65	0

There are 31 data items, and $\frac{45}{100} \times 31 = 13.95$, so you need to find the 14th data item. It is helpful to add the cumulative frequencies to this table. This shows clearly that the 14th data item lies in the class interval $30 \leqslant w < 40$; it is the 6th item in this class interval (which contains 10 items).

Age of employees	Frequency	Age of employees	Cumulative frequency
$16 \leqslant a < 25$	2	$a \leqslant 25$	2
$25 \leqslant a < 30$	6	$a \leqslant 30$	8
$30 \leqslant a < 40$	10	$a \leqslant 40$	18
$40 \leqslant a < 50$	8	$a \leqslant 50$	26
$50 \leqslant a < 65$	5	$a \leqslant 65$	31
over 65	0	$a > 65$	31

$\frac{6}{10} \times 10 = 6$, so the 6th item is $30 + 6 = 36$.

When you are dealing with discrete data, you must ensure that cumulative frequencies less than or equal to a value, as it is cumulating the results together.

Example 10

The percentage marks scored in a driving theory test sat by 200 members of the public in one day were as follows:

Mark (%)	Frequency
1–10	3
11–20	11
21–30	13
31–40	18
41–50	26
51–60	33
61–70	45
71–80	34
81–90	11
91–100	6

Only 45% of the people who sat the test that day passed. Estimate the pass mark.

Solution

Mark (%)	Frequency	Mark, *m*	Cumulative frequency
1–10	3	$m < 10.5$	3
11–20	11	$m < 20.5$	14
21–30	13	$m < 30.5$	27
31–40	18	$m < 40.5$	45
41–50	26	$m < 50.5$	71
51–60	33	$m < 60.5$	104
61–70	45	$m < 70.5$	149
71–80	34	$m < 80.5$	183
81–90	11	$m < 90.5$	194
91–100	6	$m \leqslant 100$	200

The 45th percentile is needed. This is the 90th data item, which lies in the 51–60 class interval.

Although the data are discrete, they are taken to be continuous because a mark of 40.6% would be rounded to 41%, and so the class interval is taken to be $40.5 \leqslant m < 50.5$.

The 45th percentile is the 19th data item of the 33 data items in this class interval.

45th percentile $= 50.5 + \frac{19}{33} \times 10 = 56.26$

So, an estimate of the pass mark is 56%.

Exercise 12.3A

Answers page 545

1 This is a frequency table for the number of people in a household. Calculate the median number of people in a household.

Number in household	Number of households
1	15
2	20
3	22
4	23
5	11
6	4

2 A football coach measured the distance a random sample of 120 eleven-year-old children could kick a football. The lengths are summarised in the table.

Kick distance, l (m)	Number of children
$5 \leqslant l < 10$	5
$10 \leqslant l < 20$	53
$20 \leqslant l < 30$	29
$30 \leqslant l < 50$	15
$50 \leqslant l < 70$	11
$70 \leqslant l < 100$	7

a Display this data as a cumulative frequency graph.

b Use interpolation to estimate the distance of the kick from the 40th child.

c Calculate an estimate for the mean.

3 The number of aphids on a farmer's strawberry field were counted. The results are presented in the table.

Number of strawberry plants	Number of aphids
0–19	38
20–29	97
30–39	173
40–49	225
50–69	293
70–99	174

a Construct a cumulative frequency curve and find the median.

b Using linear interpolation, estimate the median.

c Find the estimate of the mean.

4 Here are the numbers of runs which Jenson and Molly scored in 11 cricket games.

Jenson: 3, 21, 45, 66, 12, 4, 12, 65, 46, 55, 31

Molly: 34, 57, 12, 98, 17, 22, 17, 43, 23, 76, 44

a Draw box and whisker plots showing their scores.

b Compare their scores, using your diagrams and calculations to help you.

5 Consider this grouped frequency distribution:

Length (cm)	Frequency
17.5–17.9	15
18.0–18.4	27
18.5–18.9	18
19.0–19.9	12
20.0–24.9	15
25.0–29.9	4

a Identify the upper class values, if the measurements were recorded to the nearest mm.

b Construct a cumulative frequency diagram of the data.

c Using your graph, estimate the median.

d Use linear interpolation to estimate the median correct to 2 decimal places.

e Compare your arithmetic findings to your graphical estimate.

f Compare the median and the mean.

CM **6** The marks of 45 students randomly selected from those who sat a statistics test are displayed below:

36, 39, 39, 40, 41, 42, 42, 43, 44, 45, 46, 46, 46, 48, 50, 52, 53, 53, 54, 54, 55, 55, 56, 57, 57, 59, 60, 60, 60, 60, 61, 63, 64, 64, 64, 65, 65, 66, 67, 68, 69, 71, 72, 73, 73

a What is the modal mark?

b What are the lower quartile, the median and the upper quartile?

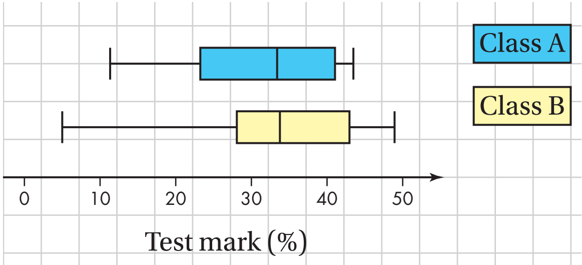
c Represent the data as a box and whisker plot.

d Are there any outliers in the data set?

e Represent the data as a cumulative frequency diagram.

f What is the range in marks between the 90th percentile and 10th percentile? Why might this be a useful measure?

7 This diagram shows the raw test marks of two student groups for the same test.



a Use it to determine which of the following statements are true.

A The median is the same for both classes.

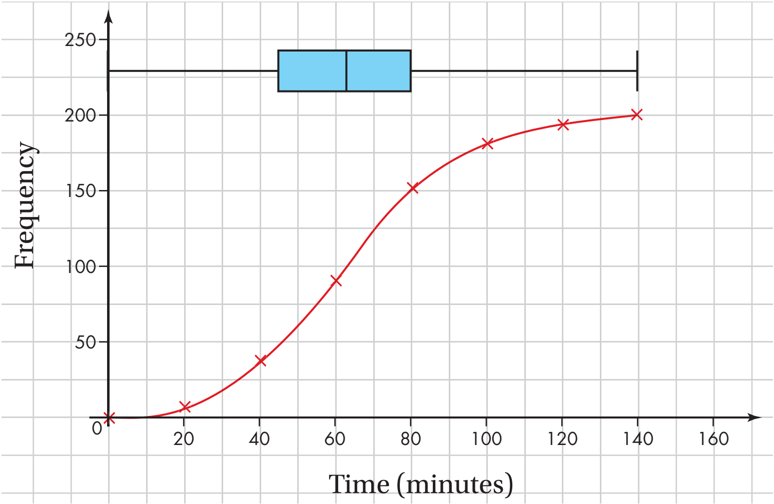
B Q_1 for Class A is 5 marks less than Q_1 for Class B.

C The interquartile range for Class A is $\frac{2}{3}$ the interquartile range of Class B.

D $P_{75} - P_{50}$ is the same for both classes.

b Write a comparison of the student groups.

8 This diagram shows the times that 200 17-year-olds spent using mobile phone apps.



a Use the diagram to estimate the following statistics:

i the median time spent using apps

ii the range of times spent using apps

iii $Q_3 - Q_1$

iv The amount of time 20% of 17-year-olds spend using mobile phone apps.

b How many students spent more than 1.6 hours using mobile phone apps?

Histograms

Histograms are best used for large sets of data when the data has been grouped into classes. They are most commonly used for continuous data and often have bars of varying width, representing unequal class intervals. The frequency of the data is shown by the area of the bars and not the height.

The vertical axis of a histogram is labelled Frequency density, which is calculated by the following formula:

$$\text{Frequency density} = \frac{\text{frequency}}{\text{class width}}$$

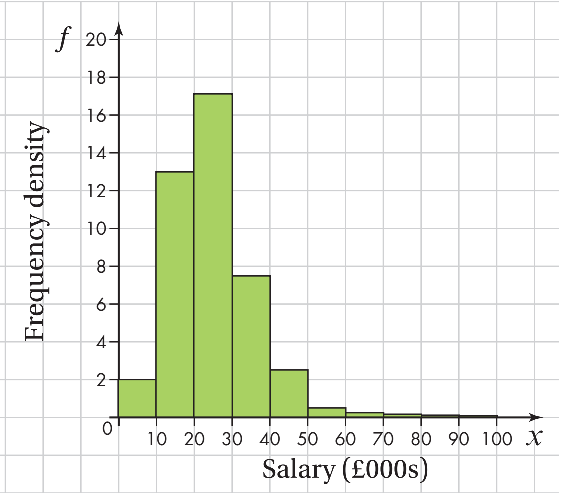
The table below shows data on gross salaries for 427 professions, drawn from the Office for National Statistics' 2010 salary tables (*Annual Survey of Hours and Earnings*).

Salary (£)	Frequency	Class width	Frequency density
0–9999	21	10	2.1
10 000–19 999	129	10	12.9
20 000–29 999	172	10	17.2
30 000–39 999	74	10	7.4
40 000–49 999	24	10	2.4
50 000–59 999	4	10	0.4
60 000–69 999	1	10	0.1
70 000–79 999	1	10	0.1
80 000–89 999	0	10	0
90 000–99 999	1	10	0.1

TECHNOLOGY

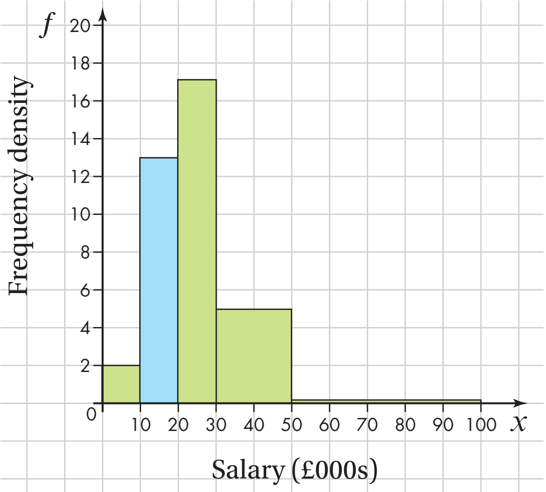
You could use graphing software to create a histogram of this data.

Plotting the frequency against the classes gives the following diagram:



An alternative is to use unequal classes to display the same data.

Salary (£)	Frequency	Class width	Frequency density
0–9999	21	10	2.1
10 000–19 999	129	10	12.9
20 000–29 999	172	10	17.2
30 000–49 999	98	20	4.9
50 000–99 999	7	50	0.14



This diagram gives an impression of the overall distribution of the data which tallies with that given by the first diagram. The data is now fairly represented, even though it is grouped into intervals with different widths.

Look at the bar shown in blue. The width is 10 and the frequency density is 12.9. So the area of the bar is 10 × 12.9 = 129, which equals the frequency.

On all histograms the vertical axis should either be labelled as Frequency density, or with the units of the frequency density.

Using the large data set 12.5

a Choose a category from your large data set, group the data using unequal intervals and display using a histogram.

b Using your table from **part a**, use linear interpolation to find the median of your category.

c Could the data gathered from your category be used to make predictions about other categories? Explain your reasons.

Exercise 12.3B

Answers page 546

1 Mr Hardy selects a random sample of 40 students and records, to the nearest hour, the time they spent gaming in a particular week.

Time (hours)	1–5	6–10	11–15	16–20	21–30	31–49
Frequency	3	7	11	14	4	1
Midpoint		8	13	18		40

a Find the midpoints of the 1–5 hour and 21–30 hour groups.

On a histogram representing this data set, the 6–10 hour group is represented by a bar of width 2 cm and height 5 cm.

b Find the width and height of the 31–49 hour group.

c Estimate the mean.

d Using linear interpolation, estimate the median time spent gaming by these students.

CM **2** Danielle and Hannah decide to weigh cookies served in the canteen.

Here is a grouped frequency table showing the masses, in grams, of the cookies.

Mass, *m* (grams)	$30 \leqslant m < 40$	$40 \leqslant m < 50$	$50 \leqslant m < 60$	$60 \leqslant m < 80$	$80 \leqslant m < 120$
Frequency	13	37	56	8	6

a Draw a cumulative frequency graph and use it to estimate:

i the median mass of a cookie

ii the upper and lower quartile masses of the cookies.

b Construct a frequency density table and use it to draw a histogram displaying the masses.

c Danielle wants to know the maximum weight a cookie can be before it is deemed an outlier. Show how she would do this.

CM **3** The lengths of runner beans were measured to the nearest whole cm. Eighty observations are given in the table below.

Length (cm)	3–8	9–13	14–25
Frequency	47	22	11

On a histogram representing this data set, the bar representing the 3–8 class has a width of 2 cm and a height of 4 cm. For the 9–13 class find:

a the width

b the height of the bar representing this class.

CM **4** The following table summarises the distances, to the nearest km, that 7359 people travelled to attend a small festival.

Distance (km)	Number of people
40–49	67
50–59	124
60–64	4023
64–69	2981
70–84	89
85–149	75

a Give a reason to justify the use of a histogram to represent these data.

b Calculate the frequency densities needed to draw a histogram for these data.

CM **5** An agriculturalist is studying the mass, m kg, of courgette plants. The data from a random sample of 70 courgette plants are summarised in the table.

(You may use $\Sigma fx = 750$ and $\Sigma fx^2 = 11\,312.5$)

A histogram has been drawn to represent these data.

The bar representing the mass $5 \leqslant m < 10$ has a width of 1.5 cm and a height of 6 cm.

Yield, m (kg)	Frequency, f
$0 \leqslant m < 5$	11
$5 \leqslant m < 10$	29
$10 \leqslant m < 15$	18
$15 \leqslant m < 25$	8
$25 \leqslant m < 35$	4

a Calculate the width and the height of the bar representing the mass $25 \leqslant m < 35$.

b Use linear interpolation to estimate the median mass of the courgette plants.

c Estimate the mean and the standard deviation of the mass of the courgette plants.

6 Here is a frequency table for the variable t, which represents the time taken, in minutes, by a group of people to run 3 km.

a Copy and complete the frequency table for t.

t	5–10	10–14	14–18	18–25	25–40
Frequency		15	22		18
Frequency density	2			3	

b Estimate the number of people who took longer than 20 minutes to run 3 km.

c Find an estimate for the mean time taken.

d Find an estimate for the standard deviation of t.

e Estimate the median and quartiles for t.

CM **7** A survey of 100 households gave the following results for their monthly shopping bills, £y.

Shopping, y (£)	Midpoint	Frequency, f
$0 \leqslant y < 200$	100	6
$200 \leqslant y < 240$	220	20
$240 \leqslant y < 280$	260	30
$280 \leqslant y < 350$	315	24
$350 \leqslant y < 500$	425	12
$500 \leqslant y < 800$	650	8

A histogram was drawn and the class $200 \leqslant £y < 240$ was represented by a rectangle of width 2 cm and height 7 cm.

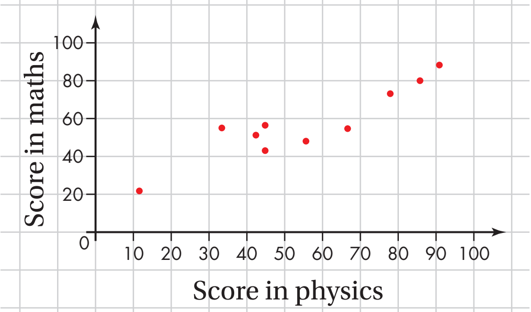
a Calculate the width and the height of the rectangle representing the class $280 \leqslant £y < 350$.

b Use linear interpolation to estimate the median shopping bill to the nearest pound.

c Estimate the mean and the standard deviation of the shopping bill for these data.

Scatter diagrams

You may remember meeting scatter diagrams at GCSE level in both maths and other subjects. You use scatter diagrams when you are comparing two variables (**bivariate data**), such as scores in maths and scores in physics.



You used them to suggest if there was a **correlation** in data. You will also have drawn lines of best fit by eye to predict values. You now need to be able to state the relationship between the variables and calculate the equation of the line of best fit for a set of bivariate data. Although the calculations you will learn give you more reliable information than the ‘by eye’ methods, scatter diagrams are still an important part of the process.

KEY INFORMATION

Scatter diagrams allow you to spot any obvious outliers.

Correlation in scatter diagrams

Correlation describes the relationship or link between two variables. A scatter diagram visually shows the relationship between the two variables. If the values of both variables are increasing, then they are *positively* correlated.

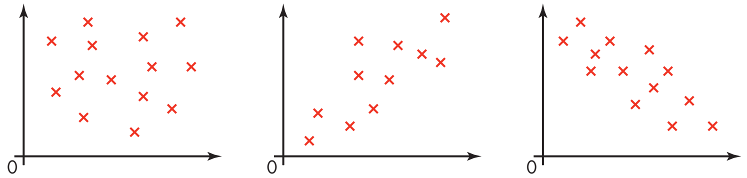
If one variable is increasing but the other is decreasing, there is a *negative* correlation.

If there is no pattern, and the points are scattered about the axes, there is no correlation.

KEY INFORMATION

Positive correlation: as one variable increases, so does the second variable.

Negative correlation: as one variable increases, the second variable decreases.



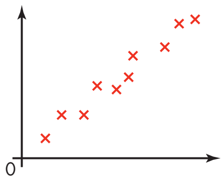
No correlation

Positive correlation

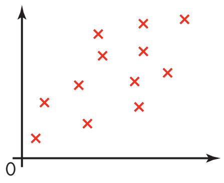
Negative correlation

The positions of the data items within a scatter diagram may show linear correlation. A line of best fit can be drawn by eye, but only when the points lie almost on a straight line. The closer to a straight line the points lie, the stronger the correlation. The line of best fit is a representation of all the data. There is, however, one point on the scatter diagram that the line of best fit *must* go through, which is the point $(\bar{x}, \bar{y})$, where $\bar{x}$ is the mean of x and

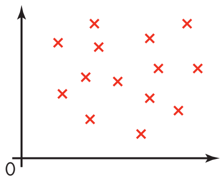
$\bar{y}$ is the mean of y. This helps the line of best fit to be central to all the data points.



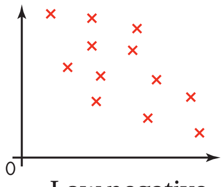
High positive correlation



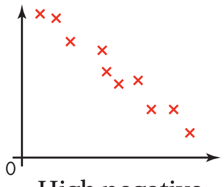
Low positive correlation



No correlation

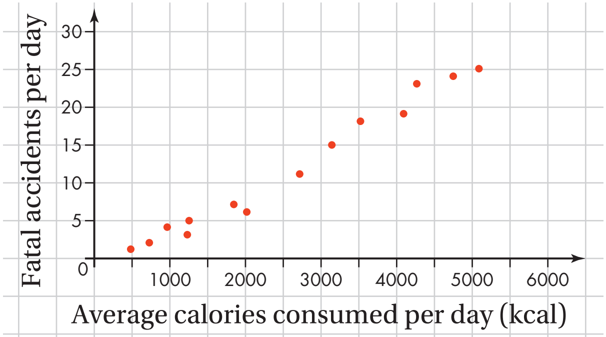


Low negative correlation



High negative correlation

The following diagram illustrates why it is important to interpret scatter diagrams with caution.



The scatter diagram shows the number of fatal accidents plotted against the average calories consumed on a single day, for different regions. Does the diagram indicate that the more calories you consume the more likely you are to have a fatal accident?

In fact, the average calories consumed and the number of fatal accidents in a day for a particular region both also correlate with a third variable – the number of units of alcohol consumed. The more alcohol is consumed, the more calories a person absorbs. The more alcohol consumed, the more fatal accidents happen on roads. Combining just the average calories consumed against the number of fatal accidents gives a false impression about the relationship between the two variables – this is spurious correlation.

In general, correlation does *not* imply causation.

Using the large data set 12.6

a Choose two variables to compare from your large data set. Calculate the standard deviation of each of your categories.

b For each of your categories, what would be deemed an outlier?

c Display the data as a scatter diagram, clearly indicating the mean point.

d Using the diagram, comment on any trends.

Exercise 12.3C

Answers page 547

CM **1** The table below shows the height and weight of each of 10 students.

Name	Height (cm)	Weight (kg)
Chloe	172	68
Arron	157	66
Melbin	190	75
Dufia	141	50
Emily	155	74
Cody	169	74
Tahlia	151	60
Thomas	177	70
Zineb	183	68
Joseph	139	52

a Plot the data on a scatter diagram.

b Describe what you notice from the diagram. Is there any correlation?

c Draw a line of best fit. Make sure it goes through the mean point.

d How heavy would you expect Uche to be if he is 196 cm tall?

2 Write down a real-life example scenario and sketch a scatter diagram for each of the following cases:

a a weak positive correlation

b a strong negative correlation

c no correlation

d a nonsense correlation.

3 The following table shows information about six students' test results in two subjects. Draw a scatter diagram and draw on a line of best fit for the data.

Subject	A	B	C	D	E	F
biology	43	64	35	57	69	92
geography	46	81	31	55	58	87

4 **a** Plot the following additional data on the scatter diagram you drew in **question 3** and draw on a new line of best fit.

Subject	G
biology	72
geography	78

b Is there an outlier within the data?

PS CM 5 The table shows information about the number of cups of coffee sold in a café on 14 consecutive days in January.

Temperature (°F)	Cups of coffee sold
62	37
67	62
66	27
69	12
58	80
62	62
68	19
64	39
65	32
56	96
68	11
64	38
69	5
60	72

a Plot the data on a scatter diagram.

b Describe what you notice from the diagram. Is there any correlation?

c Draw a line of best fit. Make sure it goes through the mean point.

d On another day the temperature was 58 °F. How many cups of coffee would you expect to be sold on this day?

Regression lines

A correlation coefficient provides you with a measure of the level of association between two variables in a bivariate distribution. If a relationship is indicated, you will want to know what that means. Drawing a line of best fit by eye is one way to spot a linear correlation but it is not a very accurate method. A more accurate way to find the line of best fit is to find and use the linear equation of the line.

The general form of an equation of a straight line is:

$$y = mx + c$$

- m is the gradient (the steepness of the line)
- c is the y-intercept (where the line crosses the y-axis).

See **Chapter 4 Coordinate geometry 1: Equations of straight lines** for more about gradients and intercepts.

An accurate way of plotting a line of best fit is to draw the **regression line**. This ensures that the distance between each point and the line of best fit is reduced to a minimum – these amounts that are left over are known as residuals. The regression line goes through the middle of the points plotted.

You use the same equation but with different letters:

$$y = a + bx$$

where

- b is the gradient
- a is the y-intercept.

If b is positive then there is a positive correlation.

If b is negative then there is a negative correlation.

The line of best fit with the equation $y = a + bx$ is called the regression line.

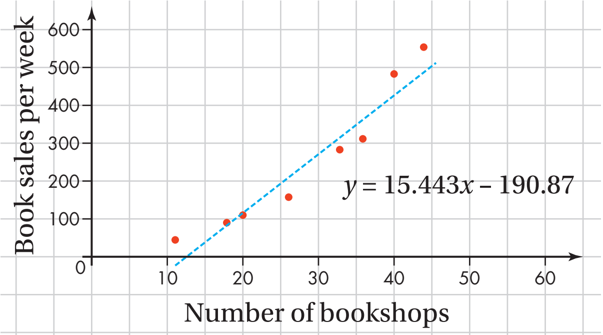
KEY INFORMATION

Regression line: a line of best fit for a given set of values, using the equation of a straight line, $y = a + bx$.

Variables and the line of best fit

In bivariate data, the **independent (explanatory) variable** is plotted on the x-axis and is independent of the other variable. The **dependent (response) variable** is plotted on the y-axis and is determined by the independent variable. The regression line is the line of best fit.

For example, your two variables could be sales of a particular book and the number of bookshops selling it. The book's sales will be dependent (y) on the number of bookshops selling it (x) – but the number of bookshops is independent.



The equation of a regression line

The equation of the regression line of y on x is:

$$y = a + bx$$

The values of a and b will be given to you, you need to know how to apply these in the equation.

Subject	A	B	C	D	E	F
biology	76	84	75	54	62	86
geography	65	43	73	54	58	77

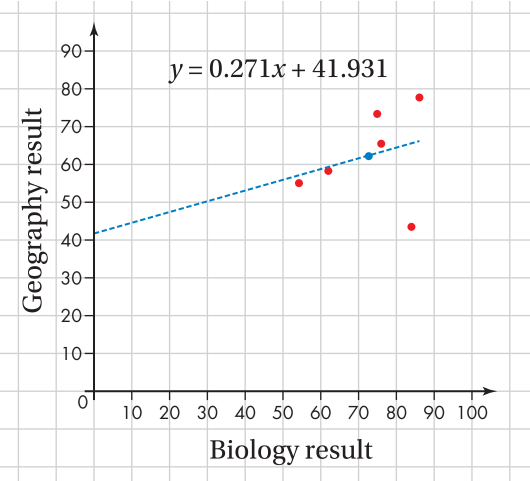
In this data set the values of a and b are $a = 41.931$ and $b = 0.271$, so the equation of the regression line is:

$$y = 41.931 + 0.271x$$

where x is the biology result and y is the geography result.

You can now use the equation of the regression line and the mean point to draw the regression line on the scatter diagram.

- From the equation, you know the y-intercept is at (0, 41.931).
- You can calculate the coordinates of the mean point to be (72.83, 61.67).
- Drawing a line between these two points give you the regression line.



TECHNOLOGY

You can use a spreadsheet software package to do the calculation. Enter the data into a spreadsheet with one column for the x values and another column for the corresponding y values. Then use the function INTERCEPT to calculate a and SLOPE to calculate b and follow the onscreen instructions.

Interpolation and extrapolation

Once you have drawn a line of best fit, either by eye or more accurately by plotting the equation of the regression line, you can start using the line of best fit to make predictions.

A regression equation can be used to predict the value of the dependent variable, based on a chosen value of the independent variable.

- **Interpolation** is estimating a value that is within the range of the data you have.
- **Extrapolation** is estimating a value outside the range of the data that you have. As the value is outside the range of the data you have, extrapolated values can be unreliable.

Generally, avoid extrapolating values unless asked and, even then, treat answers with caution.

Example 11

An experiment in which different masses were hung on a metal spring and the resulting length of the spring was measured produced the following results. The equation of the regression line is $y = 32.7 + 0.402x$.

Mass, x (kg)	10	20	30	50	100
Length, y (cm)	24	44	53	59	68

a Estimate the value for y when $x = 35$ kg. Is this interpolation or extrapolation?

b Estimate the value for y when $x = 120$ kg. Is this interpolation or extrapolation?

Solution

a Using the equation of the regression line:

$$y = 32.7 + 0.402x$$

Substitute in the x value of 35 kg.

$$y = 32.7 + (0.402 \times 35)$$

$$= 46.77 \text{ cm}$$

This is interpolation as $x = 35$ is within the range of the data you have.

b Setting x to be 120 kg, you get:

$$y = 32.7 + 0.402x$$

$$= 32.7 + (0.402 \times 120)$$

$$= 80.94 \text{ cm}$$

This is extrapolation as $x = 120$ is outside the range of the data you have.

Using the large data set 12.7

In **Using the large data set 12.6** you drew a scatter diagram.

a Describe the correlation of your scatter diagram.

b Using a spreadsheet, find the equation of the regression line. Explain what it means.

Exercise 12.3D

Answers page 548

CM **1** The weight, w grams, and the length, l cm, of each of 10 randomly selected new-born babies are given in the table below.

l	49.0	52.0	53.0	54.5	44.1	53.4	50.0	41.6	49.5	51.2
w	3424	3234	3479	3910	2855	3596	3001	2108	2906	1954

a The equation of the regression line of w on l is written in the form $w = a + bl$. Explain what a and b represent in this instance.

b The equation of the regression line is $w = 97.25l - 1799.4$. Use the equation of the regression line to estimate the weight of a new-born baby of length 60 cm.

c Comment on the reliability of your estimate, giving a reason for your answer.

2 The table below gives the number of hours spent studying for a science exam (x) by seven students, and their final exam percentage (y). The equation of the regression line is $y = 48.9 + 7.93x$. Predict the score of a student who studies for 6.5 hours.

x	3	5	1	0	4	2	6
y	72	91	60	43	78	71	94

3 The table below shows the lengths in cm (nose to tail) and corresponding weights (in kg) of cats. If a cat was 78 cm in length, predict its weight. State if this prediction is reliable or not.

Length (cm)	60	62	64	66	68	70	72
Weight (kg)	1.81	1.76	1.24	2.32	1.98	2.68	3.24

4 A football coach is doing a study on inside leg length and the distance a football can be kicked. This model allows him to determine the length of a kick when only the leg length is given. The inside leg length and kick distance for 10 males are given in the table.

Inside leg length (cm)	Kick distance (m)
72.5	41.9
73.0	44.7
74.8	43.8
77.8	47.1
79.4	52.4
79.9	50.2
80.4	57.3
81.6	58.9
83.4	54.1
87.5	50.8

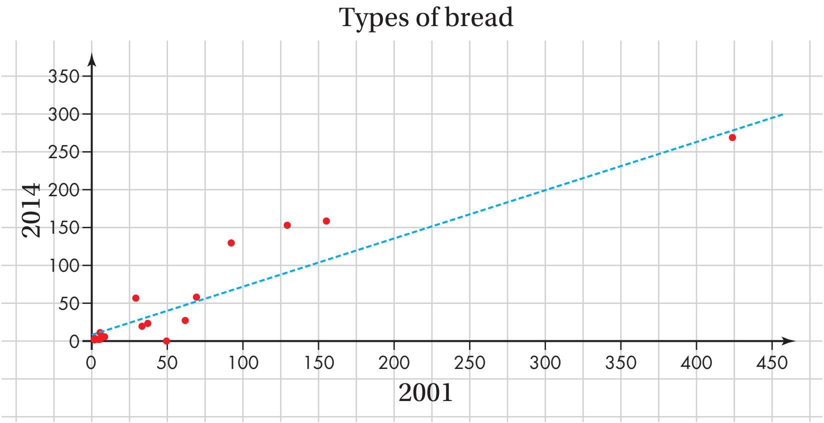
a Draw a scatter diagram to represent the data.

b The equation of the regression line that models the data is written in the form $y = a + bx$. If the intercept of the line is –18.8 and the gradient of the line is 0.87, write out the equation of the regression line.

c The coach measures a kick at 56 m. How long was the person's inside leg?

d If a person has an inside leg length of 70 cm, what does the model predict for their kick distance? Is this an accurate model?

CM **5** The amount, in grams, of different types of bread products consumed weekly per household was monitored in 2001 and 2014. The results are presented in the scatter diagram below.



a What correlation does this show?

b Using the scatter graph, what amount in grams would you expect to be consumed in 2014, if the amount consumed in 2001 was 90 grams?

c Was this interpolation or extrapolation?

The equation of the regression line is $y = 0.6427x + 5.7408$.

d What does this mean?

In 2014 the amount per week of one type of product rose to 300 grams.

e What amount would you expect to have been consumed in 2001?

f Was this interpolation or extrapolation? What are the dangers of this?

6 Jen studied the relationship between rainfall, r mm, and humidity h%, for a random sample of 11 days from a city in the UK. She obtained the following results.

h	99	96	98	93	99	99	96	97	83	83	96
r	1.7	8.9	6.3	8.2	0	0.6	5.8	7.3	8.8	7.5	12.1

Jen examined the figures for rainfall and found the following statistics:

$Q_1 = 1.8$

$Q_2 = 5.8$

$Q_3 = 8.25$

a Determine if there are any outliers in her sample.

b Present her data as a graph.

c Describe the correlation between rainfall and humidity.

d Jen found the the gradient of her line was –0.7255 and the intercept of her line was 98.88. Write the equation of the regression line in the form $r = a + bh$.

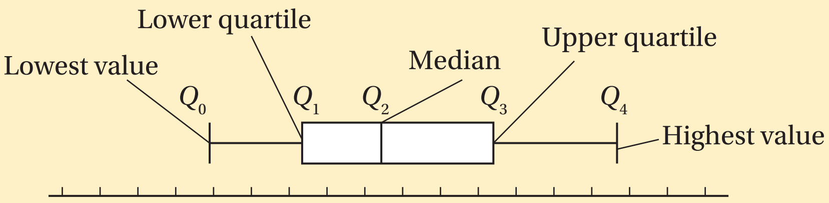
e Use your equation to estimate the humidity with a rainfall of 9.3 mm.

SUMMARY OF KEY POINTS

- Qualitative data are not numerical, but categorical.
- Quantitative data, or numerical data, can be subdivided into discrete and continuous.
- Discrete data may only take separate values, for example whole numbers.
- Continuous data may take any value, usually measured; these are usually within a range.
- The median is the middle value when the data items are placed in numerical order.
- The mode is the most common or most frequent item of data.
- The mean ($\bar{x}$) is found by adding the data values together and dividing by the number of values $\bar{x} = \frac{\sum x}{n}$
- Variance $= \frac{\sum(x-\bar{x})^2}{n}$
- Standard deviation $= \sqrt{\frac{\sum(x-\bar{x})^2}{n}}$
- An outlier can be identified as follows (IQR stands for interquartile range):
  - any data which are 1.5 × IQR below the lower quartile
  - any data which are 1.5 × IQR above the upper quartile
  - any data which are more than 2 standard deviations away from the mean.

Sometimes statisticians use mean ± 3 × standard deviation to define outliers.

- For scatter diagrams, positive correlation means that as one variable increases, so does the second variable. If one variable is increasing, but the other is *decreasing*, there is a negative correlation.
- A regression line is a line of best fit for a given set of values, using the equation of a straight line, $y = a + bx$.
- Box and whisker plots display the quartiles and minimum and maximum points visually.



EXAM-STYLE QUESTIONS 12

Answers page 548

PS CM

1 The amount of money invested, *I*, measured in £1000, and the amount of profit returned, *P*, measured in pounds, is modelled below for six different saving accounts using a risky stocks and shares account.

Savings account	A	B	C	D	E	F
***I* (£1000)**	4.3	2.8	6.0	0.3	6.8	9.5
***P* (£)**	1564	2098	2592	988	3200	4805

a Draw a scatter diagram to display this data and comment upon any correlation. **[2 marks]**

b The equation of the regression line is $y = 597 + 393x$. Explain what the equation shows. **[4 marks]**

c If you had £3500 to invest, what profit would you expect to make? **[1 mark]**

CM **2** A city in the UK had a mean annual salary for graduates of £28 500 with standard deviation of £9000.

What salaries would be outliers for the graduates? **[3 marks]**

3 Powerlifters have a certain amount of combined weight to lift in their category. The total of the combined weights, w, lifted by 17 powerlifters is 4930 kg and the standard deviation is 10.5.

a What is the maximum weight a powerlifter could lift before a judge would become concerned that it was too much? **[1 mark]**

Another powerlifter joins the competition and lifts a combined weight of 292 kg.

b Find the new mean. **[2 marks]**

4 It is reported in the news that teenagers use social media for a long time each day. A random sample of 11 students were interviewed and asked how long they spent using social media in an average week.

The total duration, y minutes, for the 11 students were:

7, 98, 121, 132, 151, 187, 204, 255, 260, 277, 357

a Find the median and quartiles for these data. **[2 marks]**

b Show that there are no outliers. **[2 marks]**

CM **5** Sixth-form students carry out an investigation on the lengths of children's feet before they have their teenage growth spurt. The sixth-form students measured the foot lengths of a random sample of 100 Year 7 students. The lengths are shown in the table.

Foot length, l (cm)	Number of children
$12 \leqslant l < 17$	2
$17 \leqslant l < 19$	14
$19 \leqslant l < 21$	38
$21 \leqslant l < 23$	34
$23 \leqslant l < 25$	12

a Using a written method, find an estimate of the median foot length. **[4 marks]**

b Using a cumulative frequency curve, find an estimate of the median foot length. **[2 marks]**

c What is the percentage error between your two estimates of the medians? **[2 marks]**

d Using a calculator, estimate the mean and standard deviation. **[3 marks]**

e What type of average would you use to best describe this data? **[1 mark]**

PS CM **6** A group of students are asked to do a blindfolded reaction test. When they hear a beep, they have to press a button which records the speed of their answer. A random sample of 104 17-year-olds are asked to take part and the results are recorded to the nearest millisecond. The results are summarised in this table.

Time (milliseconds)	Midpoint	Frequency
0–9	4.5	6
10–19	14.5	14
20–29	24.5	34
30–39	34.5	27
40–49	44.5	19
50–99	74.5	4

In a histogram, the group '10–19 milliseconds' is represented by a rectangle 2.5 cm wide and 4 cm high.

a Calculate the width of the rectangle representing the group '50–99 milliseconds'. **[2 marks]**

b Calculate the height of the rectangle representing the group '20–29 milliseconds'. **[2 marks]**

c Using a written method, estimate the median and interquartile range. **[3 marks]**

d The estimate of the mean for the data is 30.2. Would you use the mean or the median when reporting your findings? **[2 marks]**

CM **7** A car manufacturer monitored the number of times, *x*, 20 cars went back in to the garage for warranty work on engine parts. The distribution was as follows.

Number of times back, *x*	0	1	2	3	4	5	6	7	8 or more
Number of cars, *f*	5	8	2	0	2	1	1	1	0

a Use a suitable diagram to display the data. **[2 marks]**

b Find the median of the data set. **[2 marks]**

c Calculate the mean and the standard deviation of the data set. **[3 marks]**

d An outlier is a value that can be defined as outside 'the mean ± 2 × the standard deviation'. What is the maximum number of times a car could return for warranty work before the manufacturer should be concerned? **[4 marks]**

CM **8** Peter invests in a free holding and wants to make homemade ice cream. He decides to buy a Jersey cow and he monitors the amount of milk, to the nearest litre, that she produces each day during June. His results are as follows:

24	11	22	21	15	8	19	16	15	26
20	17	31	17	19	25	21	11	8	22
18	27	16	12	14	19	19	21	22	23

a Draw a box and whisker plot to display the data. **[4 marks]**

b An outlier is a value that is more than 1.5 times the interquartile range below the lower quartile or more than 1.5 times the interquartile range above the upper quartile. Show that there are no outliers in the data set. **[3 marks]**

c Compare the mean and median amounts of milk produced and comment on your answer. **[3 marks]**