11 – 19 PROGRESSION

endorsed for edexcel

Edexcel GCSE (9–1)

Statistics

New edition



Pearson

Contents

How to use this book 5

1 Collection of data 6

1.1 Describing data 7
1.2 Grouping data 9
1.3 Primary and secondary data 14
1.4 Populations 17
1.5 Petersen capture–recapture formula 20
1.6 Random sampling 22
1.7 Non-random sampling 24
1.8 Stratified sampling 27
1.9 Collection of data 29
1.10 Questionnaires and interviews 33
1.11 Problems with collected data 38
1.12 Controlling extraneous variables 40
1.13 Hypotheses 43
1.14 Designing investigations 44
Check up 46
Strengthen 48
Extend 51
Summary 54
Test 55

2 Processing and representing data 56

2.1 Tables 57
2.2 Two-way tables 61
2.3 Pictograms 64
2.4 Bar charts 67
2.5 Stem and leaf diagrams 72
2.6 Pie charts 76
2.7 Comparative pie charts 79
2.8 Population pyramids 83
2.9 Choropleth maps 89
2.10 Histograms and frequency polygons 94
2.11 Cumulative frequency charts 98
2.12 The shape of a distribution 103
2.13 Histograms with unequal class widths 107
2.14 Misleading diagrams 113
2.15 Choosing the right format 116
Check up 121
Strengthen 126
Extend 131
Summary 135
Test 137

3 Summarising data 140

3.1 Averages 141
3.2 Averages from frequency tables 144
3.3 Averages from grouped data 148
3.4 Transforming data 155
3.5 Geometric mean and weighted mean 158
3.6 Measures of dispersion for discrete data 161
3.7 Measures of dispersion for grouped data 164
3.8 Standard deviation 170
3.9 Box plots and outliers 175
3.10 Skewness 180
3.11 Deciding which average to use 183
3.12 Comparing data sets 186
3.13 Making estimates 191
Check up 194
Strengthen 197
Extend 200
Summary 202
Test 204

4 Scatter diagrams and correlation 206

4.1 Scatter diagrams 207
4.2 Correlation 210
4.3 Causal relationships 212
4.4 Line of best fit 217
4.5 Interpolation and extrapolation 219
4.6 The equation of a line of best fit 224
4.7 Spearman's rank correlation coefficient 228
4.8 Calculating Spearman's rank correlation coefficient 231
4.9 Pearson's product moment correlation coefficient 233
Check up 237
Strengthen 239
Extend 241
Summary 242
Test 243

5 Time series 244

5.1 Line graphs and time series 245
5.2 Trend lines 248
5.3 Variations in a time series 250
5.4 Moving averages 253
5.5 Estimating seasonal variations and making predictions 257
Check up 264
Strengthen 266
Extend 267
Summary 269
Test 270

Contents

6 Probability 272

6.1 The meaning of probability **273**
6.2 Experimental probability **277**
6.3 Using probability to assess risk **280**
6.4 Sample space diagrams **282**
6.5 Venn diagrams **285**
6.6 Mutually exclusive and exhaustive events **290**
6.7 The general addition law **294**
6.8 Independent events **296**
6.9 Tree diagrams **298**
6.10 Conditional probability **302**
6.11 The formula for conditional probability **306**
Check up **308**
Strengthen **310**
Extend **313**
Summary **315**
Test **317**

7 Index numbers 318

7.1 Index numbers **319**
7.2 RPI, CPI and GDP **322**
7.3 Chain base index numbers **326**
7.4 Rates of change **328**
Check up **334**
Strengthen **336**
Extend **338**
Summary **339**
Test **340**

8 Probability distributions 342

8.1 Binomial distributions **343**
8.2 Normal distributions **347**
8.3 Standardised scores **355**
8.4 Quality assurance and control charts **356**
Check up **362**
Strengthen **364**
Extend **367**
Summary **368**
Test **369**

Thinking statistically 371

Preparing for your exams 375

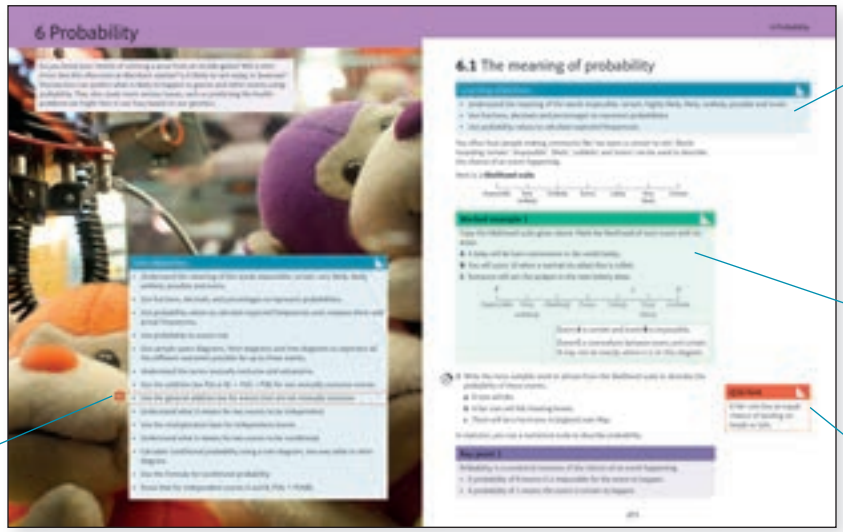
Answers 383

Index 432

How to use this book

This book is designed to give you the best preparation for your GCSE Statistics examination.

- Follows the same structure as the Edexcel scheme of work
- Supports both Foundation and Higher students
- Offers differentiated questions
- Features exam-style questions and exam preparation sections
- Gets you thinking statistically
- Includes support on calculators

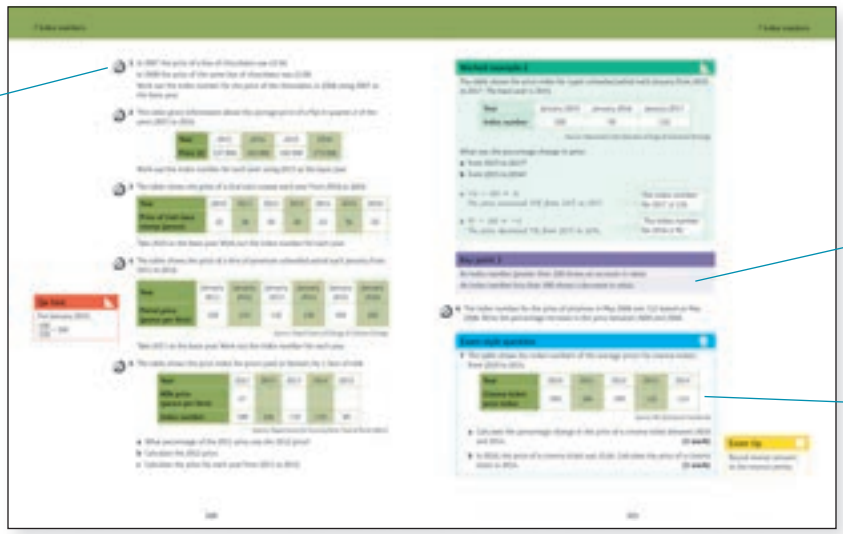


Each section opens with its learning objectives.

There are worked examples throughout each unit.

Hints are offered throughout the book to aid learning.

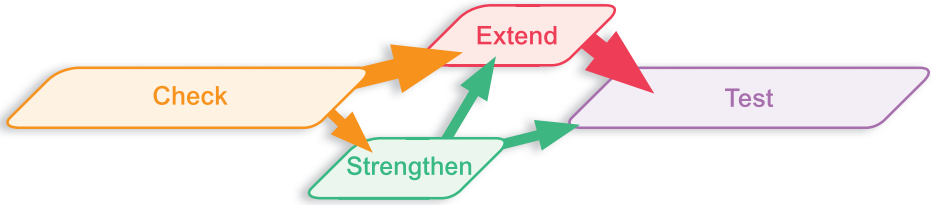
Higher tier content is clearly marked throughout the book.



Questions are tagged with Pearson Progression Steps to help offer differentiation.

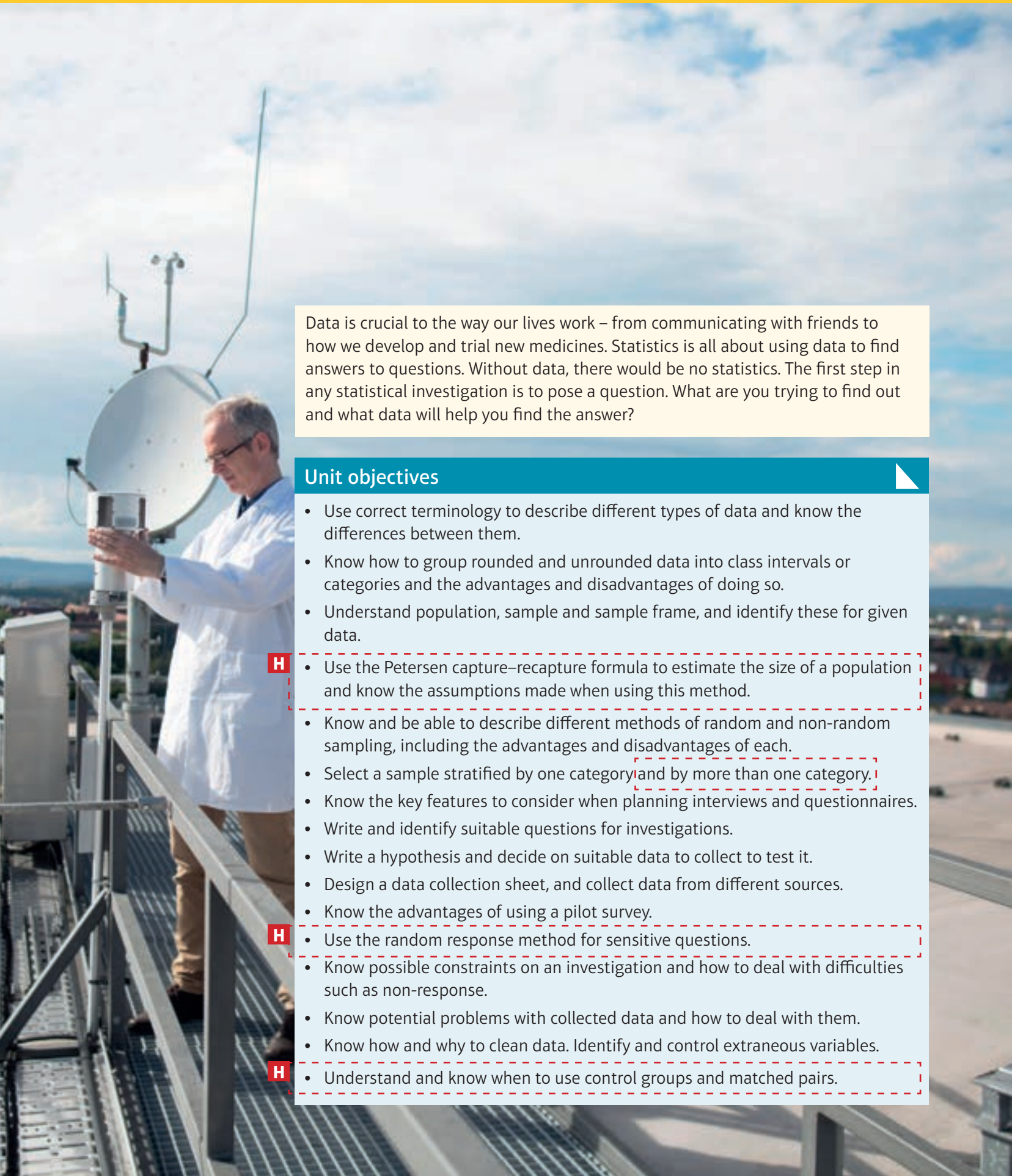
Key points are expanded.

Exam-style questions feature in every unit.



Each unit ends with a set of questions to check understanding and then routes students through to either Strengthen questions or Extend questions before the unit closes with a Test.

1 Collection of data



Data is crucial to the way our lives work – from communicating with friends to how we develop and trial new medicines. Statistics is all about using data to find answers to questions. Without data, there would be no statistics. The first step in any statistical investigation is to pose a question. What are you trying to find out and what data will help you find the answer?

Unit objectives

- Use correct terminology to describe different types of data and know the differences between them.
- Know how to group rounded and unrounded data into class intervals or categories and the advantages and disadvantages of doing so.
- Understand population, sample and sample frame, and identify these for given data.

H
- Use the Petersen capture–recapture formula to estimate the size of a population and know the assumptions made when using this method.

- Know and be able to describe different methods of random and non-random sampling, including the advantages and disadvantages of each.
- Select a sample stratified by one category and by more than one category.
- Know the key features to consider when planning interviews and questionnaires.
- Write and identify suitable questions for investigations.
- Write a hypothesis and decide on suitable data to collect to test it.
- Design a data collection sheet, and collect data from different sources.
- Know the advantages of using a pilot survey.

H
- Use the random response method for sensitive questions.

- Know possible constraints on an investigation and how to deal with difficulties such as non-response.
- Know potential problems with collected data and how to deal with them.
- Know how and why to clean data. Identify and control extraneous variables.

H
- Understand and know when to use control groups and matched pairs.

1.1 Describing data

Learning objectives

- Describe different types of data.
- Know the difference between quantitative and qualitative, discrete and continuous data.

Raw data is data just as it is collected – before it is ordered, grouped or rounded.

A statistical enquiry collects raw data on variables such as eye colour, height, price, number of followers, or level of education, to help investigate a hypothesis.

Key point 1

Raw data is either

quantitative – numerical observations or measurements, such as 10, 5.2, 39 cm

or **qualitative** – non-numerical observations, such as blue, A levels, cat.

6th **1** Which of these are qualitative data and which are quantitative data?

A Number of pets

B Height

C Make of car

Exam-style question

2 Maya is planning an investigation into this hypothesis:

'People with a university degree earn more than people without a university degree.'

State the **two** types of data she could collect to investigate this hypothesis and whether each type of data is qualitative or quantitative. **(2 marks)**

Key point 2

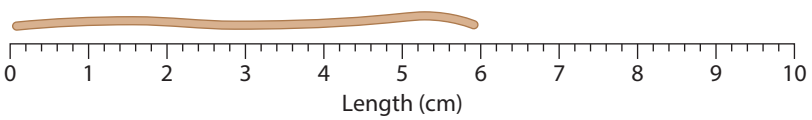
Quantitative data is either:

continuous – can take any value on a continuous numerical scale, such as length or mass

or **discrete** – can only take particular values on a continuous numerical scale, such as shoe size or number of pets.

The length of a piece of string could take any value on this scale.

It is continuous data.



3 Are these discrete data or continuous data?

A The weight of a dog

B The number of flowers in a bouquet

C The time it takes to bake a cake

4 Julita sold raffle tickets at a village fair.

The tickets were red, green, blue and yellow.

discrete **continuous** **qualitative** **quantitative**

Which of the words above can be used to describe:

a the number of tickets sold?

b the colour of tickets sold?

Key point 3

Categorical data can be sorted into non-overlapping categories.

Worked example 1

Jamal collects data on the colour and engine size of cars. Suggest categories for sorting the data.

Colour can be sorted into silver, red, blue, other.

Engine size e can be sorted into

$e \leq 1$ litre $1 \text{ litre} < e \leq 2$ litres $e > 2$ litres

Suggest some colour groups. Include 'other' to cover any you have not thought of, or mixed colours such as a red and black car.

Make sure numerical categories do not overlap. 1 litre can only be included in one category.

Ordering raw data can make it easier to use or display.

Questions such as 'Number your three favourite pizzas, with 1 for your first choice and so on' give data in a natural order, with 1 being the most popular.

Questions such as 'On a scale of 1–5, how likely are you to shop here again, where 1 is very unlikely and 5 is very likely?' use a numerical rating scale, so answers can be ordered by their rating score.

Key point 4

Ordinal data can be written in order or can be given a numerical rating scale.

6th **5** Is each data set categorical or ordinal?

a Students' year groups

b The league positions of football teams

6th **6** Write **two** types of categorical data that you could collect about mobile phones.

6th **7** Which of these could be ordinal data?

A The marks gained in a test by a group of students

B The position of dogs in a dog show

C The colours of sweets

Key point 5

Bivariate data involves pairs of related data.

In many statistical investigations you can investigate pairs of variables to find out how they are related or how changes in one variable affect the other variable. Examples include age and price of second-hand cars, or distance and time taken for train journeys.

Hint

'Bi' means 'two', as in bicycle (two wheels).

H

Key point 6

Multivariate data involves sets of three or more related data values.

For example, multivariate data for plants are colour, leaf size and height.

8th **8** Suggest words that make a pair of bivariate data in each case.

a Height and _____ of people

b Hours of work and _____

c Age of computer and _____

1.2 Grouping data

Learning objectives

- Group discrete data.
- Group continuous data.

Grouping data can help you to see the distribution of the data and spot patterns.

Key point 1

Discrete data can be grouped into classes that do not overlap, like this: 0–10, 11–20, 21–30, etc.

The intervals 0–10, 11–20, etc. are called **class intervals**.

When grouping data, think about the number of class intervals and the width of these intervals.

- If there are not enough classes, important detail may be lost.
- If there are too many classes, the classes will be very small which could hide any patterns.



1 A mathematical test is marked out of 100. Here are the marks for 60 students.

71	62	40	72	59	63	43	81	44	23
55	52	55	58	66	31	45	54	57	59
63	61	54	42	35	47	33	62	41	73
57	82	26	71	52	48	38	65	52	56
68	36	49	63	57	53	77	65	27	88
41	62	35	47	63	39	62	43	46	51

a Copy and complete the frequency table to show the students' marks.

Mark	Tally	Frequency
20–29		
30–39		
40–49		
50–59		
60–69		
70–79		
80–89		
Total		

b The pass mark for the test was 40 out of 100.
How many students passed the test?



2 A newsagent recorded the number of newspapers sold on each day in January:

40	62	67	40	49	52	57	42
46	44	48	55	53	51	56	58
58	59	60	44	52	63	48	49
42	53	57	56	53	61	51	

a Draw and complete a frequency table, using class intervals 40–44, 45–49, and so on.

b In order to cut costs, the newsagent decides that he will stock only 60 newspapers each day. In January, on how many days would he have sold out of newspapers?

Key point 2

Intervals do not need to be equal widths. Use narrower intervals where the data is close together and wider intervals where the data is spread out.

When you don't know the minimum or maximum possible value, you can use an open-ended class interval.

Worked example 1

Here are the ages of people on a bus who are streaming music on their phones:

10, 12, 13, 13, 14, 15, 16, 16, 16, 17, 17, 18,
18, 19, 20, 22, 24, 24, 27, 30, 34, 41, 56, 72

Suggest suitable class intervals for this data.

The minimum age is 0, but you don't know the maximum age, so use an open-ended class, >40.

Most of the data values are between 10 and 25, so put these in smaller class intervals.

Class intervals: 0–9, 10–14, 15–19, 20–24, 25–29, 30–40, >40

You need to select class intervals carefully. If you select too many or too few intervals, trends in the data can be obscured.

Calculations based on grouped data are less accurate than those based on raw data. In grouped data, individual data values are not known so you can only calculate estimates of the mean, mode and median.

4th **3** Sue and Lisa conducted a survey into the ages of 125 people at a classical music concert. They used the same data but drew different frequency tables.

These are the frequency tables.

Sue

Age	Frequency
0–9	1
10–19	0
20–29	2
30–39	55
40–49	56
50–59	8
60–69	2
70–79	0
80–89	1
Total	125

Lisa

Age	Frequency
0–29	3
30–34	18
35–39	37
40–44	43
45–49	13
50–59	8
>60	3
Total	125

a What is the main difference between the two frequency tables?

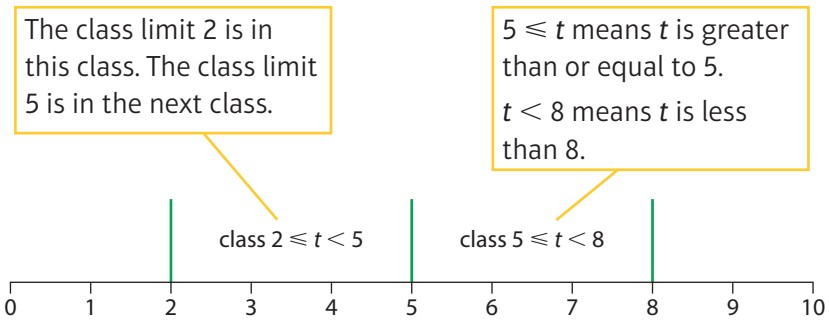
b Explain why Lisa has used such a wide class interval for people below the age of 30.

c Which frequency table shows more detail about the most common age ranges? Explain your answer.

d Why did Lisa leave the last interval open?

e How could Sue have improved her frequency table? Give two ways.

Continuous data can take any value on a continuous scale and can be sorted into classes.



Key point 3

For continuous data, the class intervals must not have gaps in between them or overlap each other.

6th

4 Twenty students take part in a 400 m race. These are the times taken (in seconds) for each person to complete the race.

54.0	58.0	69.3	82.2	70.4	63.2	69.0	78.0	54.4	66.2
53.0	56.2	71.4	76.3	80.0	84.0	72.2	68.4	56.4	62.3

Karthik, Richard and Serguei each tried to sort the data into a grouped frequency table. They each chose different class intervals.

Karthik

Time (s)
50 to 59
60 to 69
70 to 79
80 to 89

Richard

Time (s)
50–60
60–70
70–80
80–89

Serguei

Time, t (s)
$50 < t \leqslant 60$
$60 < t \leqslant 70$
$70 < t \leqslant 80$
$80 < t \leqslant 90$

a Why are Karthik and Richard's class intervals unsuitable?

b Comment on the suitability of Serguei's class intervals.

Another person runs 400 m in 105.8 seconds.

c How could Serguei change his final class interval to allow for times longer than 90 seconds?

You round continuous variables to a degree of accuracy, for example heights to the nearest centimetre, or times to the nearest tenth of a second.

You would probably measure the length of a field to the nearest metre. So, if the exact length is 235.3 m, it would be acceptable to say it is 235 m long.

Key point 4

A measurement given correct to the nearest whole unit can be inaccurate by up to $\pm\frac{1}{2}$ unit.

A field with a length of 231 m could measure between 230.5 m and 231.5 m.

You can write this as an inequality:

$230.5 \leqslant \text{length} < 231.5$

All the values from 230.5 up to but not including 231.5 round to 231.

Key point 5

When data values have been rounded, all possible values that round to the same number must fit into the same class interval.

Worked example 2

a Explain why Serguei's class intervals in question **4** are unsuitable if the times are rounded to the nearest second.

b Design a frequency table with suitable intervals.

a *A time shown as 70 seconds could have been anything in the range $69.5 \leqslant t < 70.5$. It may belong in the class interval $60 < t \leqslant 70$ or $70 < t \leqslant 80$.*

Serguei
Time, t (s)
$50 < t \leqslant 60$
$60 < t \leqslant 70$
$70 < t \leqslant 80$
$80 < t \leqslant 90$

The times 69.5 to 70 fit into this group.

Not all values that round to 70 fit into this class interval (e.g. 70.2).

b

Time, t (s)
$49.5 \leqslant t < 60.5$
$60.5 \leqslant t < 70.5$
$70.5 \leqslant t < 80.5$
$80.5 \leqslant t < 90.5$

All figures that round to 70 fit into this group.

7th **5** Gareth records the amount of rainfall each day in Runsby. Here is the raw data for January (in centimetres).

5.6	4.3	2.1	0	0.8	5.2	3.3	2.8	2.2	1.6	0.4
1.9	3.2	4.2	1.0	3.0	3.6	2.4	1.8	0.4	0	0
3.2	3.5	2.7	1.2	2.1	1.1	5.7	5.2	3.1		

Design and complete a frequency table with suitable class intervals for this data.

Q5 hint

The data has not been rounded.

6 Frank delivers parcels.
These are the masses in kilograms, to 2 decimal places, of the parcels that he delivers in one day.

2.44	1.57	2.35	1.13	2.52	1.59	2.53
0.65	2.56	1.60	2.67	1.22	2.89	1.72
2.99	0.27	3.00	1.77	3.13	1.34	3.22
1.81	0.74	1.88	1.37	1.91	0.48	2.11
1.48	2.36	0.85	2.22	1.53	2.29	

a What is the mass of the heaviest parcel?

b Frank begins to draw a frequency table.

Mass, m (kg)	Tally	Frequency
$0 \leqslant m < 0.5$		
$0.5 \leqslant m < 1$		

Copy and complete Frank's frequency table. Use classes of equal width.

7 Here are the weights of 30 boys, rounded to the nearest kilogram.

60	62	51	53	42	52	50	53	48	55
58	59	63	49	52	54	35	53	44	54
46	57	46	67	58	56	48	48	37	41

a What is the range of values that could be represented by the weight 48 kg?

b Kathleen wanted to use class intervals of $35 \leqslant w < 40$, $40 \leqslant w < 45$, $45 \leqslant w < 50$, etc. Explain why this is wrong.

c Choose class intervals of width 5 kg that would suit this rounded data.

d Use your class intervals from part **c** to create and complete a frequency table for this data.

1.3 Primary and secondary data

Learning objectives

- Know the difference between primary and secondary data.
- Understand the advantages and disadvantages of primary and secondary data.

Key point 1

Primary data is collected by, or for, the person who is going to use it.

Secondary data has been collected by someone else.

Examples of collecting primary data include:

- measuring the circumference of babies' heads in a hospital
- observing and tallying the colours of all the cars passing your house on a certain morning.

Sources of secondary data include websites, newspapers and magazines, research articles, databases and census returns.

1 Check up

Questionnaires

5th **1** Is this a closed or an open question?

'What do you think about the new hall?'

Give a reason for your answer.

Types of data

6th **2** Which of these words can be used to describe the data in parts **a** to **f**?

continuous discrete quantitative qualitative primary secondary

a Height

b Colour

c Number of aunts

d Time

e Census information on a website

f A tally you make of car types

7th **3** Which of these are primary data and which are secondary data?

A Data collected from a car magazine

B Data from the BBC website

C Data collected by asking questions of people at a supermarket

Grouping data

7th **4** A council kept a 30-day record of the number of absentees among its workers. The data is:

5	12	17	27	4	13	32	54	6	13
14	23	24	3	9	5	15	21	7	2
6	8	9	14	14	19	17	18	22	24

Sort this data into groups and draw and complete a grouped frequency table.

Q5 hint

Remember: the data is rounded.

7th **5** Thirty students were asked to time their journey to school to the nearest minute. These are the results.

6	18	29	55	7	34	28	56	33	4
2	41	33	23	7	43	26	53	44	41
32	46	16	17	3	26	17	47	22	17

Design and complete a frequency table to sort this data. Use class intervals of equal width.

1 Strengthen

Questionnaires

Q1 hint

In closed questions you choose an answer from a list. In open questions you write your own answer.

5th **1** State whether each question is open or closed.

a Where did you go on holiday last year?

b How many times a week do you buy a newspaper?

0 ☐ 1–3 ☐ 4–6 ☐ 7 ☐

Types of data

Q2 hint

Any data measured with a measuring instrument is continuous.

6th **2** Which of these are continuous data and which are discrete data?

A Time

B Number of dogs

C Volume of milk

Q3 hint

Quantitative data is 'quantities' or numbers.

6th **3** There are three horses in a field.

Use one of these words to copy and complete each sentence.

discrete quantitative qualitative continuous cumulative

a The colour of the horses is _____ data.

b The number of horses is _____ data.

Grouping data

Q4 hint

You could use an open-ended class for the higher scores.

8th **4** Here are the batting scores for 50 cricket players.

33	48	30	24	15	31	23	28	32	29
36	31	31	37	42	18	20	34	40	25
29	28	29	32	26	33	25	27	32	22
22	31	21	35	34	29	30	34	26	32
32	27	29	35	19	28	24	33	27	50

a Write the lowest score.

b Write the highest score.

c Design and complete a grouped frequency table for this data. Use classes of equal width.

d From your answer to part **c**, decide which class intervals contain the most data values.

Make a new frequency table, with:

- smaller class widths where there is most data
- wider class widths where there is not so much data.