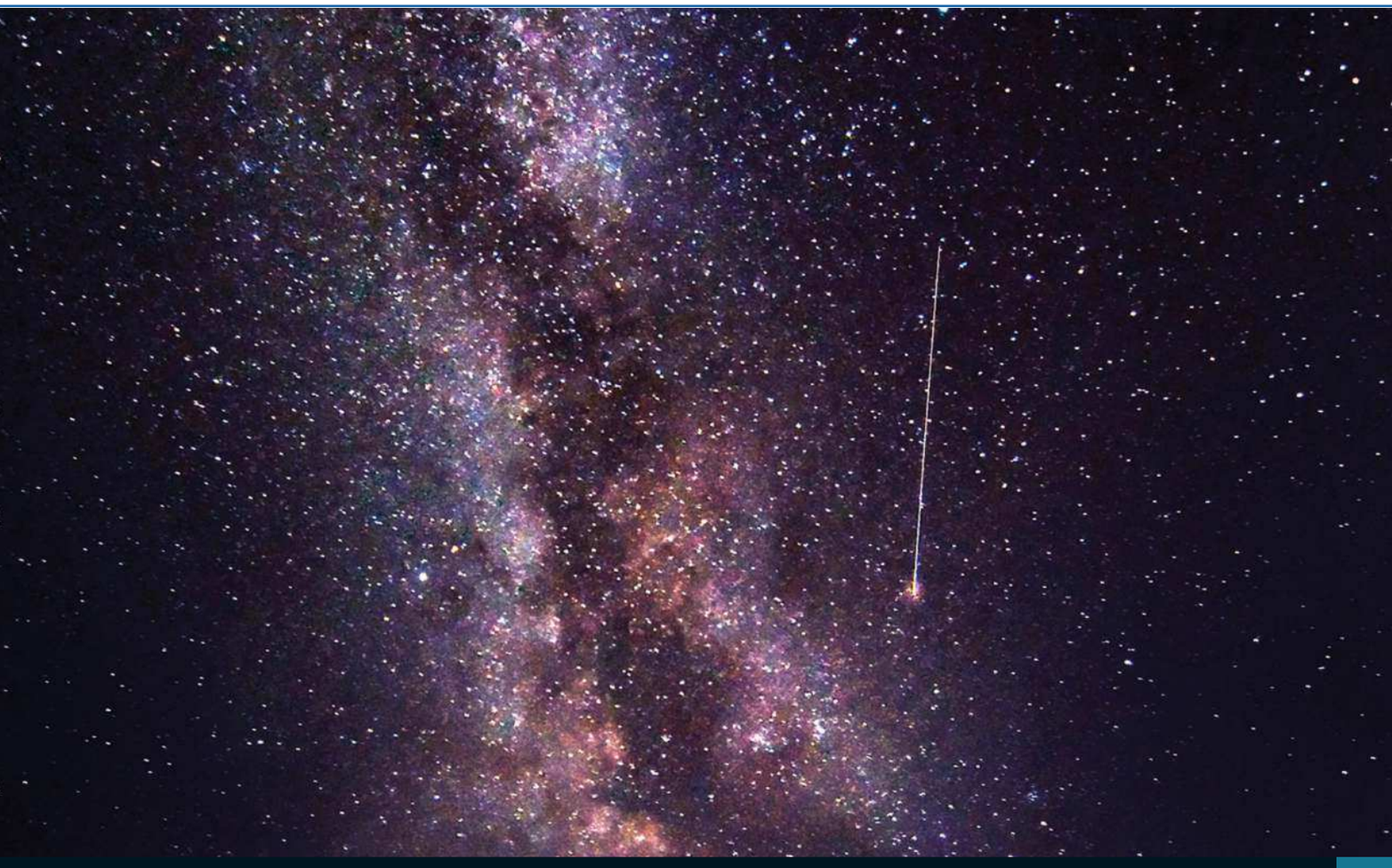
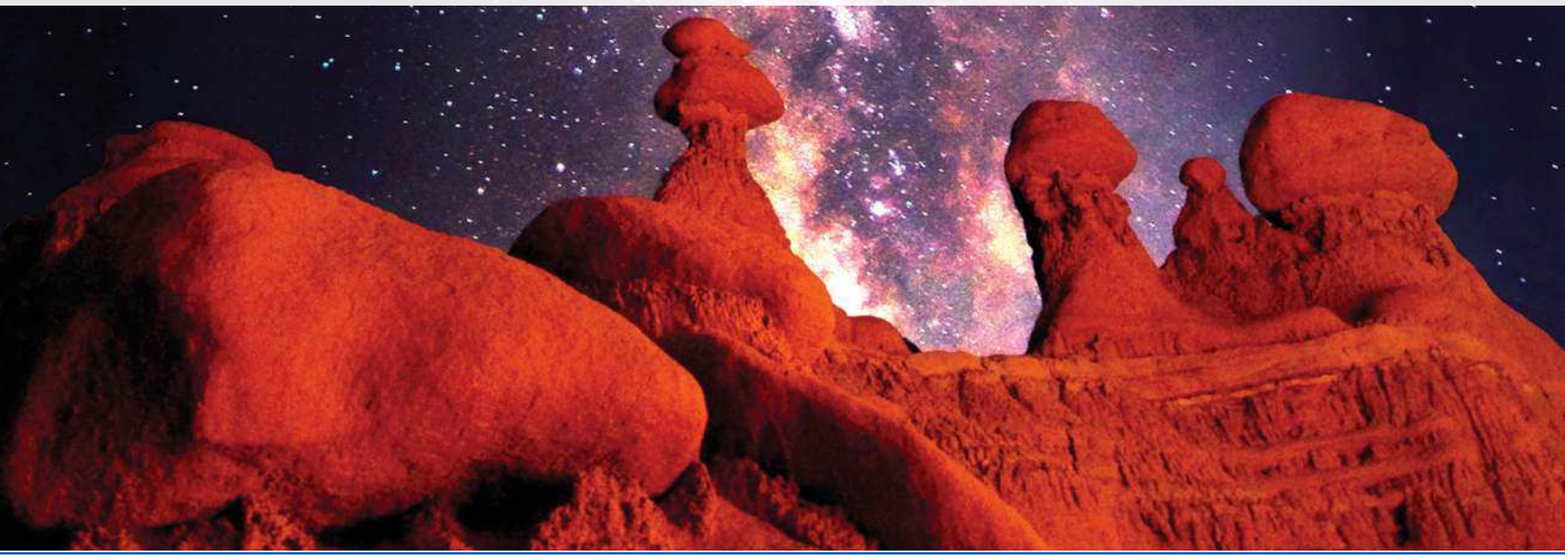


Chapter 1
Representation of data

In this chapter you will learn how to:

- display numerical data in stem-and-leaf diagrams, histograms and cumulative frequency graphs
- interpret statistical data presented in various forms
- select an appropriate method for displaying data.



PREREQUISITE KNOWLEDGE

Where it comes from	What you should be able to do	Check your skills
IGCSE® / O Level Mathematics	Obtain appropriate upper and lower bounds to solutions of simple problems when given data to a specified accuracy.	**1** A rectangular plot measures 20 m by 12 m, both to the nearest metre. Find: **a** its least possible perimeter **b** the upper boundary of its area.
IGCSE / O Level Mathematics	Construct and interpret histograms with equal and unequal intervals.	**2** A histogram is drawn to represent two classes of data. The column widths are 3 cm and 4 cm, and the column heights are 8 cm and 6 cm, respectively. What do we know about the frequencies of these two classes?
IGCSE / O Level Mathematics	Construct and use cumulative frequency diagrams.	**3** The heights of 50 trees are measured: 17 trees are less than 3 m; 44 trees are less than 4 m; and all of the trees are less than 5 m. Determine, by drawing a cumulative frequency diagram, how many trees have heights: **a** between 3 and 4 m **b** of 4 m or more.

Why do we collect, display and analyse data?

We can collect data by gathering and counting, taking surveys, giving out questionnaires or by taking measurements. We display and analyse data so that we can describe the things, both physical and social, that we see and experience around us. We can also find answers to questions that might not be immediately obvious, and we can also identify questions for further investigation.

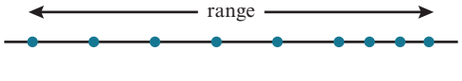
Improving our data-handling skills will allow us to better understand and evaluate the large amounts of statistical information that we meet daily. We find it in the media and from elsewhere: sports news, product advertisements, weather updates, health and environmental reports, service information, political campaigning, stock market reports and forecasts, and so on.

Through activities that involve data handling, we naturally begin to formulate questions. This is a valuable skill that helps us to make informed decisions. We also acquire skills that enable us to recognise some of the inaccurate ways in which data can be represented and analysed, and to develop the ability to evaluate the validity of someone else's research.

1.1 Types of data

There are two types of data: **qualitative** (or **categorical**) **data** are described by words and are non-numerical, such as blood types or colours. **Quantitative data** take numerical values and are either discrete or continuous. As a general rule, **discrete data** are counted and cannot be made more precise, whereas continuous data are measurements that are given to a chosen degree of accuracy.

Discrete data can take only certain values, as shown in the diagram.



The number of letters in the words of a book is an example of discrete quantitative data. Each word has 1 or 2 or 3 or 4 or… letters. There are no words with $3\frac{1}{3}$ or 4.75 letters.

Discrete quantitative data can take non-integer values. For example, United States coins have dollar values of 0.01, 0.05, 0.10, 0.25, 0.50 and 1.00. In Canada, the United Kingdom and other countries, shoe sizes such as $6\frac{1}{2}$, 7 and $7\frac{1}{2}$ are used.

Continuous data can take any value (possibly within a limited range), as shown in the diagram.

The times taken by the athletes to complete a 100-metre race is an example of continuous quantitative data. We can measure these to the nearest second, tenth of a second or even more accurately if we have the necessary equipment. The range of times is limited to positive real numbers.

KEY POINT 1.1

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

1.2 Representation of discrete data: stem-and-leaf diagrams

A **stem-and-leaf diagram** is a type of table best suited to representing small amounts of discrete data. The last digit of each data value appears as a *leaf* attached to all the other digits, which appear in a *stem*. The digits in the stem are **ordered** vertically, and the digits on the leaves are ordered horizontally, with the smallest digit placed nearest to the stem.

Each row in the table forms a **class** of values. The rows should have intervals of equal width to allow for easy visual comparison of sets of data. A **key** with the appropriate unit must be included to explain what the values in the diagram represent.

Stem-and-leaf diagrams are particularly useful because **raw data** can still be seen, and two sets of related data can be shown back-to-back for the purpose of making comparisons.

Consider the raw percentage scores of 15 students in a Physics exam, given in the following list: 58, 55, 58, 61, 72, 79, 97, 67, 61, 77, 92, 64, 69, 62 and 53.

To present the data in a stem-and-leaf diagram, we first group the scores into suitable equal-width classes.

Class widths of 10 are suitable here, as shown below.

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

Next, we arrange the scores in each row in ascending order from left to right and add a key to produce the stem-and-leaf diagram shown below.

5	3 5 8 8	Key: **5** | 3
6	1 1 2 4 7 9	represents
7	2 7 9	a score of 53%
8		
9	2 7	

In a back-to-back stem-and-leaf diagram, the leaves to the right of the stem ascend left to right, and the leaves on the left of the stem ascend right to left (as shown in Worked example 1.1).

TIP

The diagram should have a bar chart-like shape, which is achieved by aligning the leaves in columns. It is advisable to redraw the diagram if any errors are noticed, or to complete it in pencil, so that accuracy can be maintained.

WORKED EXAMPLE 1.1

The number of days on which rain fell in a certain town in each month of 2016 and 2017 are given.

Year 2016					
Jan: 17	Feb: 20	Mar: 13	Apr: 12	May: 10	Jun: 8
Jul: 0	Aug: 1	Sep: 5	Oct: 11	Nov: 16	Dec: 9

Year 2017					
Jan: 9	Feb: 13	Mar: 11	Apr: 8	May: 6	Jun: 3
Jul: 1	Aug: 2	Sep: 2	Oct: 4	Nov: 8	Dec: 7

Display the data in a back-to-back stem-and-leaf diagram and briefly compare the rainfall in 2016 with the rainfall in 2017.

Answer

2016		2017
9 8 5 1 0	**0**	1 2 3 4 6 7 8 8 9
7 6 3 2 1 0	**1**	1 3
0	**2**	

Key: 5 | **0** | 6 represents 5 days in a month of 2016 and 6 days in a month of 2017

We group the values for the months of each year into classes 0–9, 10–19 and 20–29, and then arrange the values in each class in order with a key, as shown.

It rained on more days in 2016 (122 days) than it did in 2017 (74 days).

No information is given about the amount of rain that fell, so it would be a mistake to say that more rain fell in 2016 than in 2017.

TIP

If rows of leaves are particularly long, repeated values may be used in the stem (but this is not necessary in Worked example 1.1). However, if there were, say, 30 leaves in one of the rows, we might consider grouping the data into narrower classes of 0–4, 5–9, 10–14, 15–19 and 20–24. This would require 0, 0, 1, 1 and 2 in the stem.

KEY POINT 1.2

Data in a stem-and-leaf diagram are ordered in rows of equal widths.

EXERCISE 1A

1 Twenty people leaving a cinema are each asked, "How many times have you attended the cinema in the past year?" Their responses are:

6, 2, 13, 1, 4, 8, 11, 3, 4, 16, 7, 20, 13, 5, 15, 3, 12, 9, 26 and 10.

Construct a stem-and-leaf diagram for these data and include a key.

2 A shopkeeper takes 12 bags of coins to the bank. The bags contain the following numbers of coins:

150, 163, 158, 165, 172, 152, 160, 170, 156, 162, 159 and 175.

a Represent this information in a stem-and-leaf diagram.

b Each bag contains coins of the same value, and the shopkeeper has at least one bag containing coins with dollar values of 0.10, 0.25, 0.50 and 1.00 only.

What is the greatest possible value of all the coins in the 12 bags?

3 This stem-and-leaf diagram shows the number of employees at 20 companies.

1	0 8 8 8 8 9 9	Key: **1** | 0
2	0 5 6 6 7 7 8 9	represents 10
3	0 1 1 2 9	employees

a What is the most common number of employees?

b How many of the companies have fewer than 25 employees?

c What percentage of the companies have more than 30 employees?

d Determine which of the three rows in the stem-and-leaf diagram contains the smallest number of:

i companies **ii** employees.

4 Over a 14-day period, data were collected on the number of passengers travelling on two ferries, A and Z. The results are presented to the right.

Ferry A (14)		Ferry Z (14)
8 7 6	**2**	
7 6 4 0	**3**	0 5 8
8 6 5 3	**4**	3 4 5 7 7 7
5 3 3	**5**	0 2 6 6 9

Key: 3 | **5** | 0 represents 53 passengers on A and 50 passengers on Z

a How many more passengers travelled on ferry Z than on ferry A?

b The cost of a trip on ferry A is \$12.50 and the cost of a trip on ferry Z is \$$x$. The takings on ferry Z were \$3.30 less than the takings on ferry A over this period. Find the value of x.

c Find the least and greatest possible number of days on which the two ferries could have carried exactly the same number of passengers.

5 The runs scored by two batsmen in 15 cricket matches last season were:

Batsman P: 53, 41, 57, 38, 41, 37, 59, 48, 52, 39, 47, 36, 37, 44, 59.

Batsman Q: 56, 48, 31, 64, 21, 52, 45, 36, 57, 68, 77, 20, 42, 51, 71.

a Show the data in a diagram that allows easy comparison of the two performances.

b Giving a reason for your answer, decide which of the batsmen performed:

i better **ii** more consistently.

6 The total numbers of eggs laid in the nests of two species of bird were recorded over several breeding seasons.

The numbers of eggs laid in the nests of 10 wrens and 10 dunnocks are:

Wrens: 22, 18, 21, 23, 17, 23, 20, 19, 24, 13.

Dunnocks: 28, 24, 23, 19, 30, 27, 22, 25, 22, 17.

a Represent the data in a back-to-back stem-and-leaf diagram with rows of width 5.

b Given that all of these eggs hatched and that the survival rate for dunnock chicks is 92%, estimate the number of dunnock chicks that survived.

c Find the survival rate for the wren chicks, given that 14 did not survive.

PS **7** This back-to-back stem-and-leaf diagram shows the percentage scores of the 25 students who were the top performers in an examination.

Girls (12)		Boys (13)
4 1	**8**	2
8 6 6	**8**	5 9
3 2 1 0	**9**	0 1 3 3 4 4
8 7 7	**9**	5 6 6 9

Key: 1 | **8** | 2 represents 81% for a girl and 82% for a boy

The 25 students are arranged in a line in the order of their scores. Describe the student in the middle of the line and find the greatest possible number of boys in the line who are not standing next to a girl.

1.3 Representation of continuous data: histograms

Continuous data are given to a certain degree of accuracy, such as 3 significant figures, 2 decimal places, to the nearest 10 and so on. We usually refer to this as *rounding*.

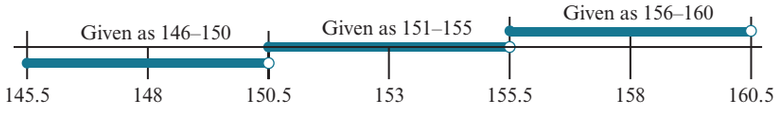
When values are rounded, gaps appear between classes of values and this can lead to a misunderstanding of continuous data because those gaps do not exist.

Consider heights to the nearest centimetre, given as 146–150, 151–155 and 156–160.

Gaps of 1cm appear between classes because the values are rounded.

Using h for height, the actual classes are $145.5 \leqslant h < 150.5$, $150.5 \leqslant h < 155.5$ and $155.5 \leqslant h < 160.5$ cm.

The classes are shown in the diagram below, with the **lower and upper boundary** values and the **class mid-values** (also called midpoints) indicated.



Lower **class boundaries** are 145.5, 150.5 and 155.5 cm.

Upper class boundaries are 150.5, 155.5 and 160.5 cm.

Class widths are $150.5 - 145.5 = 5$, $155.5 - 150.5 = 5$ and $160.5 - 155.5 = 5$.

Class mid-values are $\frac{145.5+150.5}{2} = 148$, $\frac{150.5+155.5}{2} = 153$ and $\frac{155.5+160.5}{2} = 158$.

A **histogram** is best suited to illustrating continuous data but it can also be used to illustrate discrete data. We might have to group the data ourselves or it may be given to us in a **grouped frequency table**, such as those presented in the tables below, which show the ages and the percentage scores of 100 students who took an examination.

Age (A years)	$16 \leqslant A < 18$	$18 \leqslant A < 20$	$20 \leqslant A < 22$
No. students (f)	34	46	20

Score (%)	10–29	30–59	60–79	80–99
No. students (f)	6	21	60	13

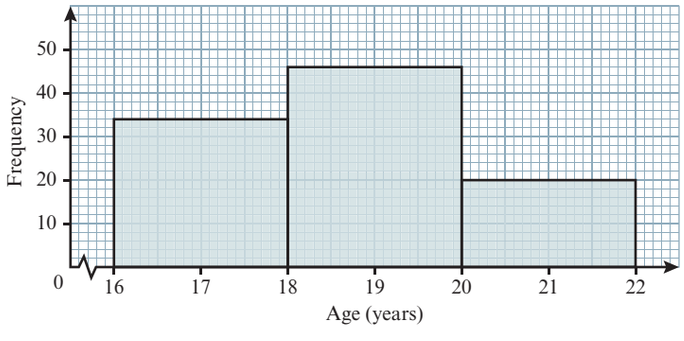
TIP

'No.' is the abbreviation used for 'Number of' throughout this book.

The first table shows three classes of continuous data; there are no gaps between the classes and the classes have equal-width intervals of 2 years. This means that we can represent the data in a frequency diagram by drawing three equal-width columns with column heights equal to the class frequencies, as shown below.

TIP

We *concertina* part of an axis to show that a range of values has been omitted.



The following table shows the areas of the columns and the **frequency** of each of the three classes presented in the diagram on the previous page.

	First	Second	Third
Area	$2 \times 34 = 68$	$2 \times 46 = 92$	$2 \times 20 = 40$
Frequency	34	46	20

From this table we can see that the ratio of the column areas, 68 : 92 : 40, is exactly the same as the ratio of the frequencies, 34 : 46 : 20.

In a histogram, the area of a column represents the frequency of the corresponding class, so that the area must be proportional to the frequency.

TIP

The symbol $\propto$ means 'is proportional to'.

We may see this written as 'area $\propto$ frequency'.

This also means that in every histogram, just as in the example above, the ratio of column areas is the same as the ratio of the frequencies, even if the classes do not have equal widths.

Also, there can be no gaps between the columns in a histogram because the upper boundary of one class is equal to the lower boundary of the neighbouring class. A gap can appear only when a class has zero frequency.

The axis showing the measurements is labelled as a continuous number line, and the width of each column is equal to the width of the class that it represents.

When we construct a histogram, since the classes may not have equal widths, the height of each column is no longer determined by the frequency alone, but must be calculated so that area $\propto$ frequency.

The vertical axis of the histogram is labelled **frequency density**, which measures frequency per standard interval. The simplest and most commonly used standard interval is 1 unit of measurement.

For example, a column representing 85 objects with masses from 50 to 60 kg has a frequency density of $\frac{85\,\text{objects}}{(60-50)\,\text{kg}} = 8.5$ objects per kilogram or 0.0085 objects per gram and so on.

KEY POINT 1.3

For a standard interval of 1 unit of measurement, $\text{Frequency density} = \frac{\text{class frequency}}{\text{class width}}$, which can be rearranged to give

$\text{Class frequency} = \text{class width} \times \text{frequency density}$

In a histogram, we can see the relative frequencies of classes by comparing column areas, and we can make estimates by assuming that the values in each class are spread evenly over the whole **class interval**.

WORKED EXAMPLE 1.2

The masses, m kg, of 100 children are grouped into two classes, as shown in the table.

Mass (m kg)	$40 \leqslant m < 50$	$50 \leqslant m < 70$
No. children (f)	40	60

a Illustrate the data in a histogram.

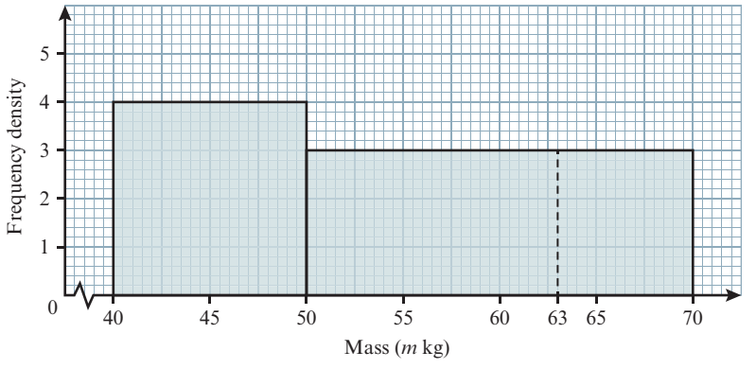
b Estimate the number of children with masses between 45 and 63 kg.

Answer

a

Mass (m kg)	$40 \leqslant m < 50$	$50 \leqslant m < 70$
No. children (f)	40	60
Class width (kg)	10	20
Frequency density	$40 \div 10 = 4$	$60 \div 20 = 3$

Frequency density is calculated for the unequal-width intervals in the table. The masses are represented in the histogram, where frequency density measures *number of children per 1 kg* or simply *children per kg*.



b There are children with masses from 45 to 63 kg in both classes, so we must split this interval into two parts: 45–50 and 50–63.

$\frac{1}{2} \times 40 = 20$ children

The class 40–50 kg has a mid-value of 45 kg and a frequency of 40.

$$\text{Frequency} = \text{width} \times \text{frequency density}$$
$$= (63 - 50)\,\text{kg} \times \frac{3 \text{ children}}{1\,\text{kg}}$$
$$= 13 \times 3 \text{ children}$$
$$= 39 \text{ children}$$

Our estimate for the interval 50–63 kg is equal to the area corresponding to this section of the second column.

Our estimate is $20 + 39 = 59$ children.

We add together the estimates for the two intervals.

TIP

Column areas are equal to class frequencies. For example, the area of the first column is

$$(50 - 40)\,\text{kg} \times \frac{4 \text{ children}}{1\,\text{kg}}$$
$$= 40 \text{ children}.$$

TIP

If we drew column heights of 8 and 6 instead of 4 and 3, then frequency density would measure *children per 2 kg*. The area of the first column would be

$$(50 - 40)\,\text{kg} \times \frac{8 \text{ children}}{2\,\text{kg}}$$
$$= 40 \text{ children}.$$

Consider the times taken, to the nearest minute, for 36 athletes to complete a race, as given in the table below.

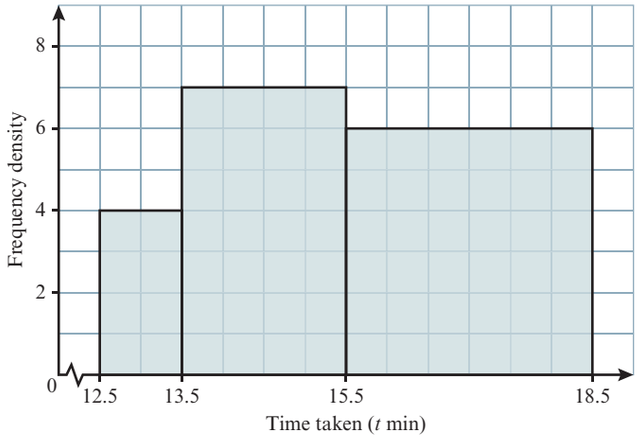
Time taken (min)	13	14–15	16–18
No. athletes (f)	4	14	18

Gaps of 1 minute appear between classes because the times are rounded.

Frequency densities are calculated in the following table.

Time taken (t min)	$12.5 \leqslant t < 13.5$	$13.5 \leqslant t < 15.5$	$15.5 \leqslant t < 18.5$
No. athletes (f)	4	14	18
Class width (min)	1	2	3
Frequency density	$4 \div 1 = 4$	$14 \div 2 = 7$	$18 \div 3 = 6$

This histogram represents the race times, where frequency density measures *athletes per minute*.



TIP

Use class boundaries (rather than rounded values) to find class widths, otherwise incorrect frequency densities will be obtained.

TIP

The class with the highest frequency does not necessarily have the highest frequency density.

TIP

Do think carefully about the scales you use when constructing a histogram or any other type of diagram. Sensible scales, such as 1 cm for 1, 5, 10, 20 or 50 units, allow you to read values with much greater accuracy than scales such as 1 cm for 3, 7 or 23 units. For similar reasons, try to use as much of the sheet of graph paper as possible, ensuring that the whole diagram will fit before you start to draw it.

WORKED EXAMPLE 1.3

Use the histogram of race times shown previously to estimate:

a the number of athletes who took less than 13.0 minutes

b the number of athletes who took between 14.5 and 17.5 minutes

c the time taken to run the race by the slowest three athletes.

Answer

We can see that two blocks represent one athlete in the histogram. So, instead of calculating with frequency densities, we can simply count the number of blocks and divide by 2 to estimate the number of athletes involved.

a $4 \div 2 = 2$ athletes

There are four blocks to the left of 13.0 minutes.

b $38 \div 2 = 19$ athletes

There are $14 + 24 = 38$ blocks between 14.5 and 17.5 minutes.

c Between 18.0 and 18.5 minutes.

The slowest three athletes are represented by the six blocks to the right of 18.0 minutes.

EXPLORE 1.1

Refer back to the table in Section 1.3 that shows the percentage scores of 100 students who took an examination.

Discuss what adjustments must be made so that the data can be represented in a histogram.

How could we make these adjustments and is there more than one way of doing this?

TIP

It is not acceptable to draw the axes or the columns of a histogram freehand. Always use a ruler!

EXERCISE 1B

1 In a particular city there are 51 buildings of historical interest. The following table presents the ages of these buildings, given to the nearest 50 years.

Age (years)	50–150	200–300	350–450	500–600
No. buildings (f)	15	18	12	6

a Write down the lower and upper boundary values of the class containing the greatest number of buildings.

b State the widths of the four class intervals.

c Illustrate the data in a histogram.

d Estimate the number of buildings that are between 250 and 400 years old.

2 The masses, m grams, of 690 medical samples are given in the following table.

Mass (m grams)	$4 \leqslant m < 12$	$12 \leqslant m < 24$	$24 \leqslant m < 28$
No. medical samples (f)	224	396	p

a Find the value of p that appears in the table.

b On graph paper, draw a histogram to represent the data.

c Calculate an estimate of the number of samples with masses between 8 and 18 grams.

3 The table below shows the heights, in metres, of 50 boys and of 50 girls.

Height (m)	1.2–	1.3–	1.6–	1.8–1.9
No. boys (f)	7	11	26	6
No. girls (f)	10	22	16	2

a How many children are between 1.3 and 1.6 metres tall?

b Draw a histogram to represent the heights of all the boys and girls together.

c Estimate the number of children whose heights are 1.7 metres or more.

4 The heights of 600 saplings are shown in the following table.

Height (cm)	0–	5–	15–	30–u
No. saplings (f)	64	232	240	64

a Suggest a suitable value for u, the upper boundary of the data.

b Illustrate the data in a histogram.

c Calculate an estimate of the number of saplings with heights that are:

i less than 25 cm **ii** between 7.5 and 19.5 cm.