Probability & Statistics

Chapter 1
Representation of data

Cambridge AS level

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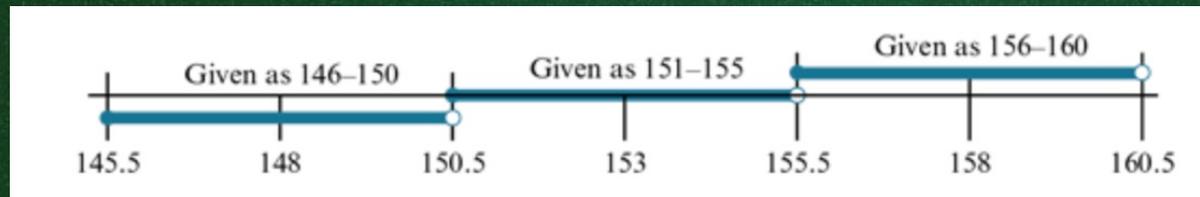
When values are rounded, gaps appear between classes of values and this can lead to a misunderstanding of continuous data because those gaps do not exist.

1.3 REPRESENTATION OF CONTINUOUS DATA: HISTOGRAMS

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

Lower class boundaries

Upper class boundaries

Class widths

Class mid-values

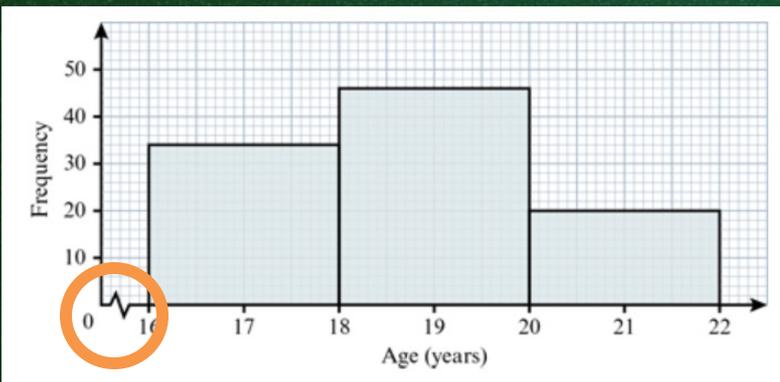# 1.3 REPRESENTATION OF CONTINUOUS DATA: HISTOGRAMS

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**.

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 resource.

TIP: We *concertina* part of an axis to show that a range of values has been omitted.
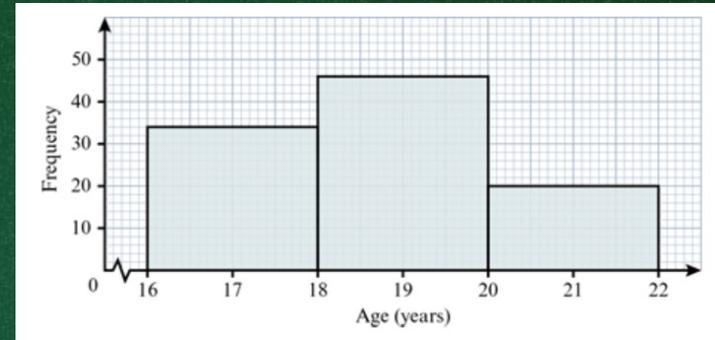
1.3 REPRESENTATION OF CONTINUOUS DATA: HISTOGRAMS

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

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

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

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

1.3 REPRESENTATION OF CONTINUOUS DATA: HISTOGRAMS

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.

KEY POINT: For a standard interval of unit of measurement,

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

which can be rearranged to give

$$Class\ frequency = class\ width \times frequency\ density$$

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

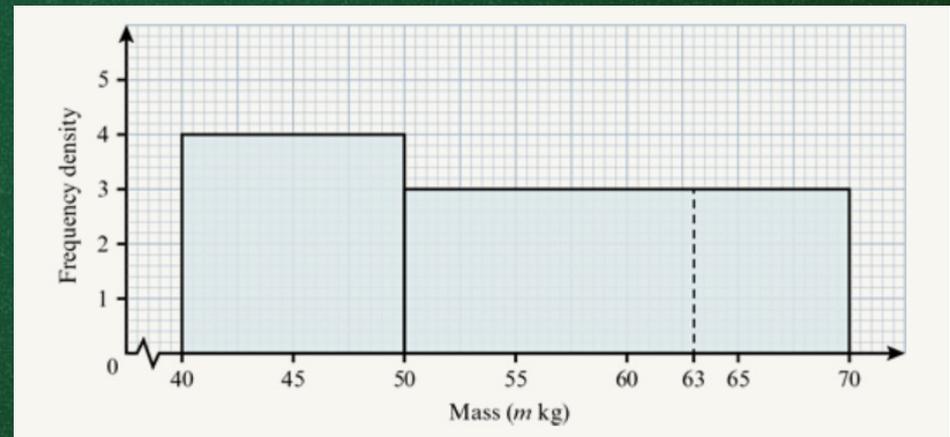| Mass (m kg)	$40 \leqslant m < 50$	$50 \leqslant m < 70$
No. children (f)	40	60

a.   Illustrate the data in a histogram.

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

a.

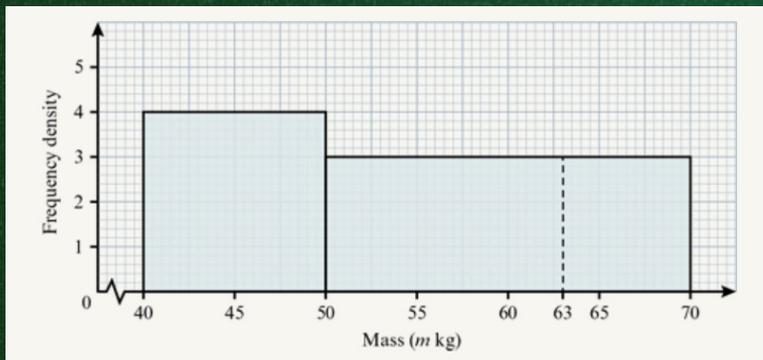| Mass (m kg)	$40 \leqslant m < 50$	$50 \leqslant m < 70$
No. children (f)	40	60
Class width (kg)	10	20
Frequency density	$40 \div 10 = 4$	$60 \div 20 = 3$

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

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

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

$Frequency = (63 - 50) \times 3 = 39$ children
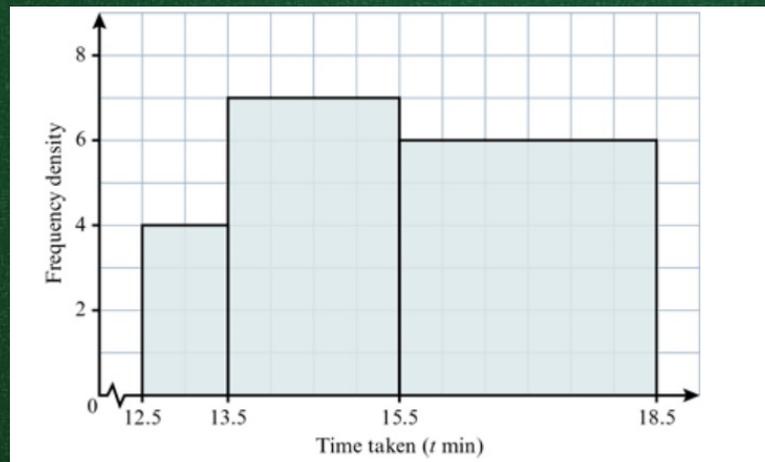
$20 + 39 = 59$ children

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

a.   Illustrate the data in a histogram.
b.   The number of athletes who took less than 13.0 minutes
c.   The number of athletes who took between 14.5 and 17.5 minutes
d.   The time taken to run the race by the slowest three athletes.

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	$4 \div 1 = 4$	$14 \div 2 = 7$	$18 \div 3 = 6$

TIP: Use class boundaries (rather than rounded values) to find class widths, otherwise incorrect frequency densities will be obtained.
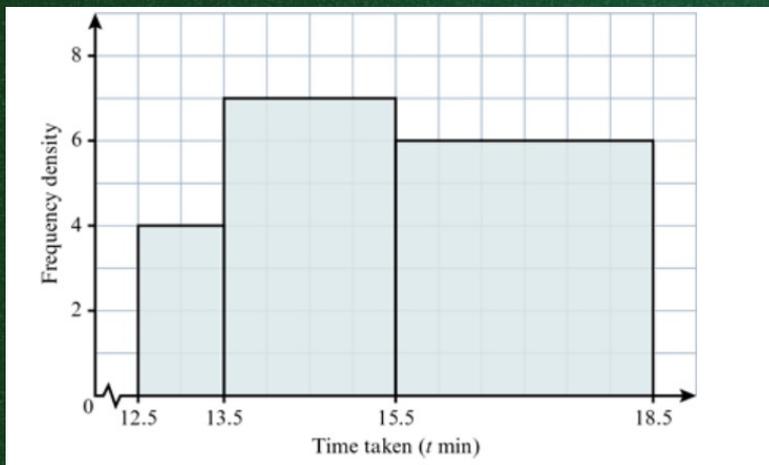
Consider the times taken, to the nearest minute, for 36 athletes to complete a race, as given in the table below.

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

b. The number of athletes who took less than 13.0 minutes
c. The number of athletes who took between 14.5 and 17.5 minutes
d. The time taken to run the race by the slowest three athletes.

TIP: The class with the highest frequency does not necessarily have the highest frequency density

$b.\ 4 \times \dfrac{1}{2} = 2$ athletes

$c.\ (15.5 - 14.5) \times 7 = 7$ athletes

$(17.5 - 15.5) \times 6 = 12$ athletes

$7 + 12 = 19$ athletes

$d.\ 18.0 - 18.5$ minutes

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.

5 Each of the 70 trainees at a secretarial college was asked to type a copy of a particular document. The times taken are shown, correct to the nearest 0.1 minutes, in the following table.

Time taken (min)	2.6–2.8	2.9–3.0	3.1–3.2	3.3–3.7
No. trainees (f)	15	25	20	10

   a Explain why the interval for the first class has a width of 0.3 minutes.

   b Represent the times taken in a histogram.

   c Estimate, to the nearest second, the upper boundary of the times taken by the fastest 10 typists.

   d It is given that 15 trainees took between 3.15 and b minutes. Calculate an estimate for the value of b when:

     i $b > 3.15$ **ii** $b < 3.15$.

6 A railway line monitored 15% of its August train journeys to find their departure delay times. The results are shown below. It is given that 24 of these journeys were delayed by less than 2 minutes.
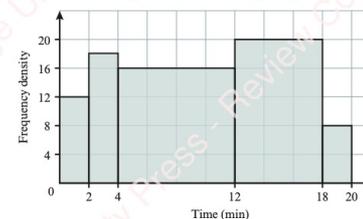
   a How many journeys were monitored?

   b Calculate an estimate of the number of these journeys that were delayed by:

     i 1 to 3 minutes   **ii** 10 to 15 minutes.

   c Show that a total of 2160 journeys were provided in August.

   d Calculate an estimate of the number of August journeys that were delayed by 3 to 7 minutes. State any assumptions that you make in your calculations.

7 A university investigated how much space on its computers' hard drives is used for data storage. The results are shown below. It is given that 40 hard drives use less than 20 GB for data storage.
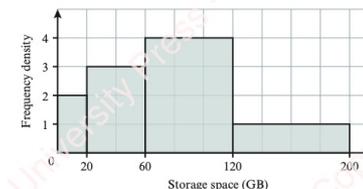
   a Find the total number of hard drives represented.

   b Calculate an estimate of the number of hard drives that use less than 50 GB.

   **c** Estimate the value of k, if 25% of the hard drives use k GB or more.

8 The lengths of the 575 items in a candle maker's workshop are represented in the histogram.

  a What proportion of the items are less than 25 cm long?

  b Estimate the number of items that are between 12.4 and 36.8 cm long.

  c The shortest 20% of the workshop's items are to be recycled. Calculate an estimate of the length of the shortest item that will not be recycled.
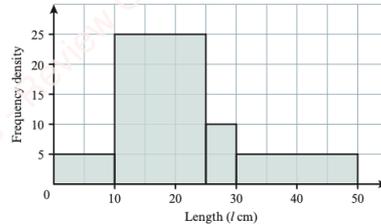
9 The thicknesses, k mm, of some steel sheets are represented in the histogram. It is given that $k < 0.4$ for 180 sheets.

  a Find the ratio between the frequencies of the three classes. Give your answer in simplified form.

  b Find the value of n, given that frequency density measures sheets per n mm.

  c Calculate an estimate of the number of sheets for which:

    i $k < 0.5$ ii $0.75 \le k < 0.94$

  d The sheets are classified as thin, medium or thick in the ratio $1 : 3 : 1$.

  Estimate the thickness of a medium sheet, giving your answer in the form $a \le k < b$. How accurate are your values for a and b?
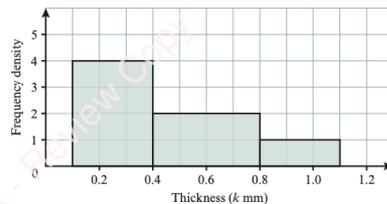
10 The masses, in kilograms, of the animals treated at a veterinary clinic in the past year are illustrated in a histogram. The histogram has four columns of equal height. The following table shows the class intervals and the number of animals in two of the classes.

Mass (kg)	3–5	6–12	13–32	33–44
No. animals (f)	a	371	1060	b

  a Find the value of a and of b, and show that a total of 2226 animals were treated at the clinic.

  b Calculate an estimate of the lower boundary of the masses of the heaviest 50% of these animals.

11 The minimum daily temperature at a mountain village was recorded to the nearest 0.5°C on 200 consecutive days. The results are grouped into a frequency table and a histogram is drawn.

The temperatures ranged from 0.5°C to 2°C on n days, and this class is represented by a column of height h cm.

The temperatures ranged from –2.5 to –0.5°C on d days. Find, in terms of n, h and d, the height of the column that represents these temperatures.

12 The frequency densities of the four classes in a histogram are in the ratio $4 : 3 : 2 : 1$. The frequencies of these classes are in the ratio $10 : 15 : 24 : 8$.

Find the total width of the histogram, given that the narrowest class interval is represented by a column of width 3 cm.

13 The percentage examination scores of 747 students are given in the following table.

Score (%)	p–50	51–70	71–80	81–q
No. students (f)	165	240	195	147

Given that the frequency densities of the four classes of percentage scores are in the ratio $5 : 8 : 13 : 7$, find the value of p and of q.