Statistics & Probability · Unit S17 · Standard

Averages, Continued

Measure sixty journeys to the nearest tenth of a minute and no two learners give the same answer, so the frequency table you learned in S06 has sixty rows of 1. Continuous data has to be grouped — and the moment it is grouped the individual values are gone. This unit gets a mean, a median and a modal class out of what is left, and is honest, every single time, about the word estimate.

Standard (MYP 4–5) · about 50 minutes · 60 practice questions · Criteria A, B, C and D

BY THE END OF THIS UNIT YOU CAN

• **Explain** why continuous data has to be sorted into classes before any average can be found, and why it then has no mode

• **Calculate** the midpoint of a class and the product $f \times x$ for every row

• **Estimate** the mean of grouped continuous data as $\Sigma fx \div \Sigma f$, to 3 significant figures

• **Explain** the midpoint assumption — and why it makes every grouped mean an estimate

• **Identify** the class that holds the median by counting to $n/2$, then **estimate** the median itself from the cumulative frequency curve

• **Explain** what changes, and what does not, when the same data is regrouped at a different class width

• **Discuss** the honest choices for an open-ended final class such as 60+, and state the one you took

• **State** the modal class, and **explain** why a continuous measurement has a modal class and not a mode

THE ONE IDEA BEHIND ALL OF IT

Grouping is a trade. You give up every individual measurement and you get back a table small enough to read. The estimated mean is the price list for that trade: it assumes that every value in a class sits exactly at the **midpoint** of that class, works out $\Sigma fx \div \Sigma f$ as if that were true, and hands you a number that is close to the real mean but is not it. The

whole of this unit is that one assumption — where it comes from, how big an error it costs, what happens to the error when you change the class width, and what to do when the last class has no top at all. Every answer you write from a grouped table is an estimate, because the raw values are lost when the data is grouped.

estimated mean = Σfx ÷ Σf

where x is the midpoint of each class

WHERE THIS UNIT SITS

What you already have. S06 *The Three Averages* owns the discrete versions: mean, median and mode from a list, Σ*fx* ÷ Σ*f* from a frequency table, and a first look at midpoints for grouped *discrete* data. None of that is retaught here. This unit extends it to **continuous** data, where grouping is not a convenience but the only way the data can be written down at all.

Who owns which rule. S15 *Climbing Curve* owns the cumulative frequency curve — building the column, plotting at the upper boundaries, reading in at *n*/2. This unit uses that curve once, to turn a median *class* into a median *value*, and points back to S15 for the drawing. S22 *Bars With Meaning* owns the histogram: touching bars, the shape language, frequency polygons and the unequal-width problem. The equal-width bars on this page are here only so you can see where the classes and their midpoints sit.

Accuracy. N04 *Round It Right* set the rule this unit leans on hardest: an estimate is quoted to **3 significant figures** and labelled an estimate. A grouped mean written to six figures is not more accurate, it is less honest.

Where the numbers come from. Every dataset on this page was produced once by the seeded generator at the top of this unit’s build script and then grouped by *counting* the raw values into classes. The 60 journey times are printed below in full, which is what lets this page put each estimate beside the value the raw data really has.

CONCEPT 1 · CLASSES ARE NOT OPTIONAL

why continuous data has to be grouped

Continuous data is measured, not counted, so it can take any value in a range and two measurements almost never land on the same number. Write the 60 journey times in a frequency table one value at a time and you get sixty rows, nearly all of them with a frequency of 1 — a longer version of the list you started with, and no use to anybody. So the values are sorted into **classes**: equal intervals that *meet*, written $0 \le x < 10$, $10 \le x < 20$, and so on. The rule at the join is the one thing to be careful about: a journey of exactly 20.0 minutes belongs to $20 \le x < 30$, the class whose *lower* boundary it is, because that is what the inequality signs say. Once the table is written the individual times are gone. The table can still tell you how many learners are in each class; it can no longer tell you what any one of them scored, and everything in the rest of this unit follows from that.

EXAMPLE · the 60 journey times, in minutes, exactly as they were recorded. 42.4, 47.2, 18.6, 14.5, 16.7, 26.6, 36.5, 19.1, 26.7, 27.6, 15.2, 41.4, 50.5, 21.1, 24.3, 11.0, 21.1, 42.0, 16.9, 18.9, 40.7, 7.1, 21.8, 25.2, 54.4, 9.5, 20.0, 24.5, 34.2, 3.3, 32.4, 46.8, 34.5, 31.5, 51.5, 20.1, 25.1, 11.1, 33.2, 16.4, 40.0, 16.7, 34.2, 31.6, 27.7, 27.3, 16.3, 31.3, 7.0, 7.3, 41.1, 24.6, 24.9, 10.4, 45.5, 35.7, 21.9, 25.2, 24.6, 22.6. Counting them into six classes of width 10 gives the frequencies 5, 13, 20, 10, 9, 3.

60 journey times · classes and their midpoints

Time, x (minutes)	Midpoint	Frequency, f
0 ≤ x < 10	5	5
10 ≤ x < 20	15	13
20 ≤ x < 30	25	20
30 ≤ x < 40	35	10
40 ≤ x < 50	45	9
50 ≤ x < 60	55	3
Total		60

Every class is 10 minutes wide, and the classes meet.

The table the rest of the unit works from. The midpoint column is written first, before any multiplying, because every number after it depends on it.

EXAMPLE. The fastest of the 60 journeys was 3.3 minutes and the slowest 54.4, so six classes of width 10 cover them all. The class $20 \leq x < 30$ holds 20 learners — the biggest group — but the table can no longer name a single one of their times. A learner who took exactly 20 minutes is in $20 \leq x < 30$, not in $10 \leq x < 20$: the boundary belongs upwards.

CONCEPT 2 · MIDPOINTS ON PARADE

Σfx ÷ Σf, and the assumption behind it

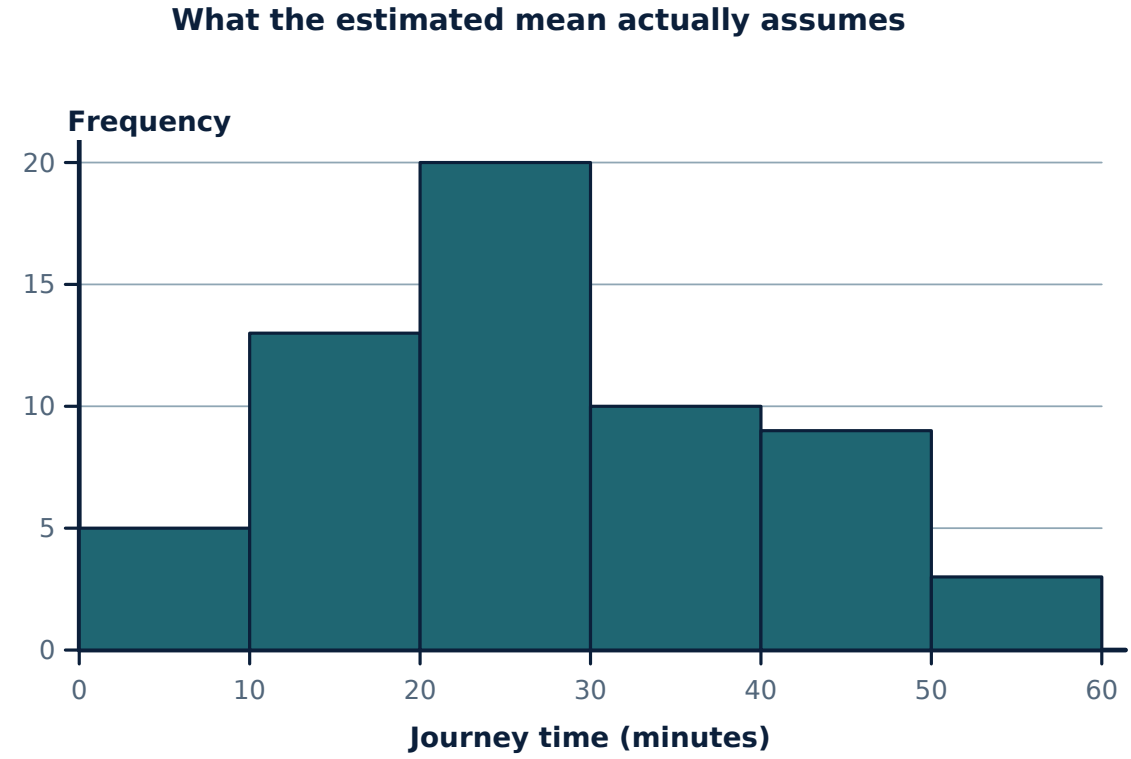
The mean needs a total, and the total needs a value for every learner — which the table no longer has. So the table makes an assumption, and it is the only one available: **every value in a class is treated as sitting at the midpoint of that class**. The midpoint is the average of the two boundaries, (lower + upper) ÷ 2, so the class $20 \leq x < 30$ has midpoint 25. Multiply each midpoint x by its frequency f to get the total that class contributes, add those products down the page for Σfx, add the frequencies for Σf, and divide. That is the **estimated mean**, $\Sigma fx \div \Sigma f$, and it must be written with the word estimate attached and rounded to 3 significant figures. It is not the mean. It is what the mean would be if every learner in a class had taken exactly the midpoint time, and none of them did.

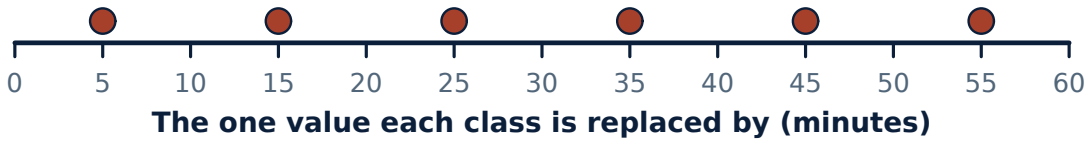
The estimated mean · Σfx ÷ Σf

Midpoint, x (minutes)	Frequency, f	f × x
5	5	25
15	13	195
25	20	500
35	10	350
45	9	405
55	3	165
Total	60	1640

1640 ÷ 60 = 27.3 minutes, an estimate.

Multiply across, add down, divide once. The total 1640 is the journey time the midpoint model assumes — not the total anybody actually travelled.





The midpoint assumption, drawn: six values stand for sixty.

What the estimate really claims. Each bar is a class; the marker beneath it is the single time every learner in that class is assumed to have taken. The bars are here to locate the classes, nothing more — the histogram itself belongs to S22.

Worked example 1 · Midpoints on parade

1	$0 \le x < 10$: midpoint $(0 + 10) \div 2 = 5$, frequency 5.	**fx = 25**
2	$10 \le x < 20$: midpoint $(10 + 20) \div 2 = 15$, frequency 13.	**fx = 195**

3	$20 \le x < 30$: midpoint $(20 + 30) \div 2 = 25$, frequency 20.	**fx = 500**
4	$30 \le x < 40$: midpoint $(30 + 40) \div 2 = 35$, frequency 10.	**fx = 350**
5	$40 \le x < 50$: midpoint $(40 + 50) \div 2 = 45$, frequency 9.	**fx = 405**
6	$50 \le x < 60$: midpoint $(50 + 60) \div 2 = 55$, frequency 3.	**fx = 165**
7	Add the fx column for Σfx.	**1640**
8	Add the frequency column for Σf, and check it is n.	**60 = n** ✓
9	Divide, and round to 3 significant figures.	**1640 ÷ 60 = 27.3 min**
10	Write the answer with the word it needs.	**estimated mean 27.3 min**

Two habits are being drilled here. The first is dividing by Σf and not by the number of rows: 60, not 6. The second is the label. The answer is 27.3 minutes, and it is an estimate, because the raw values are lost when the data is grouped. The true mean of the 60 recorded times is 26.6 minutes, so the midpoint model was 0.708 of a minute out — close enough to be useful and far enough to matter.

EXAMPLE. Midpoints first: 5, 15, 25, 35, 45, 55. Then the products *fx*: 25, 195, 500, 350, 405, 165. Adding them gives Σ*fx* = **1640** and Σ*f* = 60, so the estimated mean is 1640 ÷ 60 = **27.3 minutes** to 3 significant figures. The true mean of the 60 recorded times is 26.6 minutes, so the midpoint model is out by about 0.708 of a minute — small, real, and exactly what the word estimate is for.

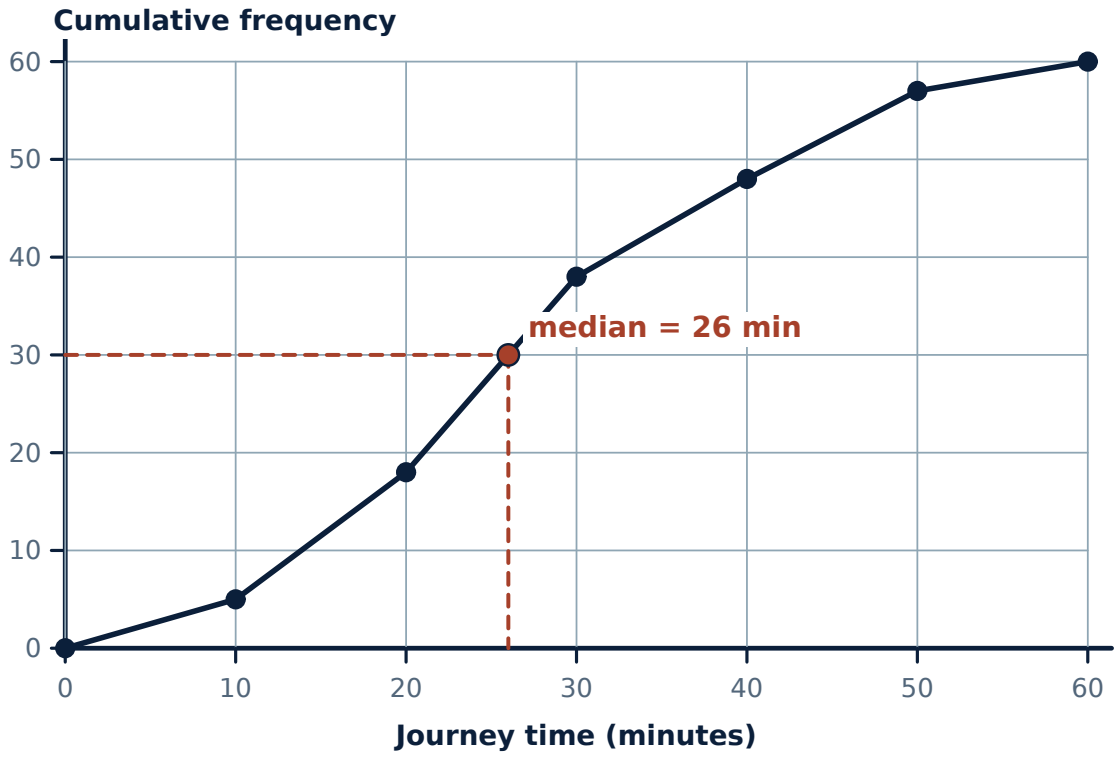
CONCEPT 3 · WHICH CLASS HOLDS THE MEDIAN?

count to n/2, then read the curve

For grouped data the median is found in two moves. First find the **median class**, which is pure counting: the middle of *n* values sits at position *n*/2, so run the frequencies as a running total until the total first reaches *n*/2, and the class you are standing in is the median class. That answer is *exact* — the table really does say the middle value is in that class. Second, if a single value is wanted, **estimate** it from the cumulative frequency curve: in at *n*/2 on the count axis, across to the curve, down to the value axis. That second answer is not exact, because

the curve has to assume the values are spread evenly across the class. Notice the position: $n/2$, not $(n+1)/2$. The $(n+1)/2$ rule counts to a middle *item* in a list, and a grouped table has no items left to count to.

The median estimated from the climbing curve



In at 30 on the count axis, across, then down.

Counting gives the median class exactly. Only the single value read down from the curve is an estimate, and the dashed guides are part of the answer.

Worked example 2 · Which class holds the median?

1	n = 60, so the middle sits at n ÷ 2.	**position 30**
2	Run the frequencies as a running total.	**5, 18, 38, 48, 57, 60**
3	Find the first total that reaches 30.	**38, in the class 20 ≤ x < 30**
4	State the median class. This part is exact.	**20 ≤ x < 30**
5	For a single value, go in at 30 on the cumulative frequency curve.	**across, then down**
6	Read the median off the value axis.	**median ≈ 26 min**

*Two answers, two different standards of truth, and a good solution separates them. The median **class** $20 \le x < 30$ is a fact the table states: 18 learners had already been counted by the bottom of it and 38 by the top, so the 30th value is inside it. The median **value** of about 26 minutes is an estimate, because the raw values are lost when the data is grouped. The true median of the 60 recorded times is 25 minutes.*

EXAMPLE. n = 60, so count to 30. The running total goes 5, 18, 38, 48, 57, 60 — it first reaches 30 in the class **$20 \le x < 30$**, which is the median class, and that much is exact. Reading the curve in at 30 puts the median at about **26 minutes**. The true median of the 60 recorded times is 25 minutes, so the curve was 1 of a minute out, which is what spreading a class evenly costs.

CONCEPT 4 · CHANGE THE WIDTH, CHANGE THE ANSWER

regrouping, and the drift it causes

Class width is a **choice**, and every choice gives a different estimated mean from exactly the same measurements. Group the 60 journey times in tens and the estimate is one number; group the same 60 times in twenties and it is another. Nothing about the learners changed — only how much of each one you threw away. The pattern is worth stating carefully: the wider the classes, the further a value can sit from its midpoint, so the wider the classes the bigger the error you should expect. Narrow classes estimate better and are more work to read; wide classes read at a glance and blur the data. What never changes is the true mean of the raw measurements: the drift you see is entirely the grouping’s, not the data’s, and saying so is half of a good answer.

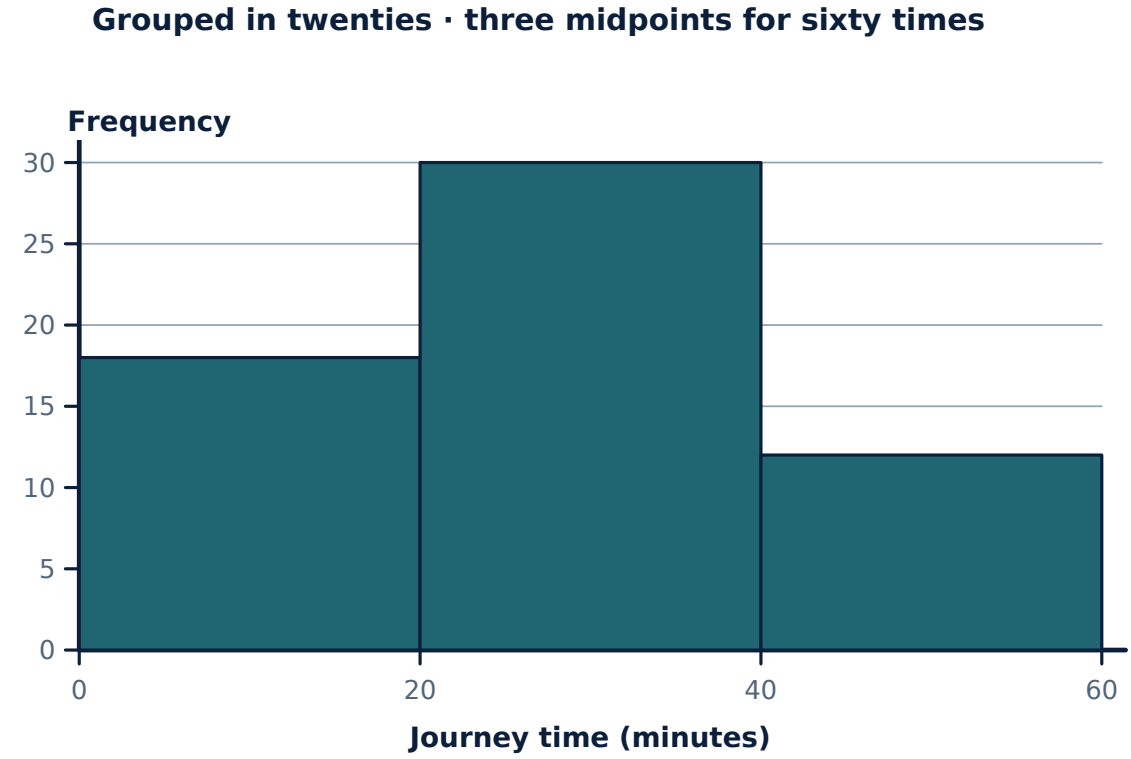
The same data, grouped two ways

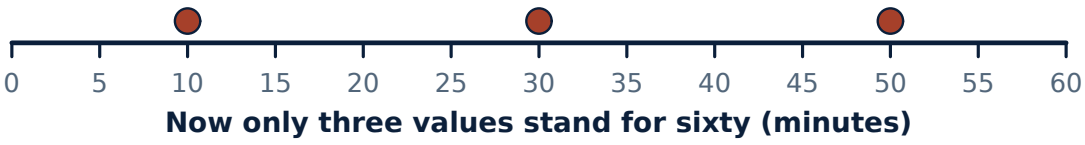
Time, x (minutes)	Midpoint	Frequency, f
0 ≤ x < 10	5	5
10 ≤ x < 20	15	13
20 ≤ x < 30	25	20
30 ≤ x < 40	35	10
40 ≤ x < 50	45	9
50 ≤ x < 60	55	3
Total		60

Time, x (minutes)	Midpoint	Frequency, f
0 ≤ x < 20	10	18
20 ≤ x < 40	30	30
40 ≤ x < 60	50	12
Total		60

Same 60 times, two tables, two different estimates.

The same 60 measurements, twice. Nothing in the data changed between these two tables; the midpoints did, and that is enough to move the estimated mean.





Wider classes, cruder midpoints, a bigger error.

Grouped in twenties, three markers now stand for all 60 journeys. Wider classes leave every value further from its stand-in, which is exactly where the extra error comes from.

Worked example 3 · The same data, regrouped

1	Group the 60 times in tens: the midpoints are these.	**5, 15, 25, 35, 45, 55**
2	Σfx ÷ Σf for that table.	**1640 ÷ 60 = 27.3 min**

3	Now group the same 60 times in twenties: three classes.	**18, 30, 12**
4	The midpoints have moved.	**10, 30, 50**
5	Σfx ÷ Σf for the wider table.	**1680 ÷ 60 = 28.0 min**
6	Measure the drift between the two estimates.	**0.667 min**

The data did not move; the classes did. Every learner in 20 ≤ x < 40 is now assumed to have taken 30 minutes instead of 25 or 35, and 30 learners sit in that one class, so the totals shift and the estimate shifts with them. Against the true mean of 26.6 minutes the width-10 table is out by 0.708 and the width-20 table by 1.375: wider classes, bigger error. Both answers are an estimate, because the raw values are lost when the data is grouped.

EXAMPLE. Grouped in tens the estimate is 27.3 minutes; grouped in twenties, with midpoints 10, 30, 50, it is 28.0 minutes. The same 60 measurements have moved their own average by 0.667 of a minute, and the true mean, 26.6 minutes, sat still the whole time. Both estimates are too high here, and the wider grouping is the further out — 1.375 against 0.708.

CONCEPT 5 · THE OPEN END

60+, and the three honest readings

A final class printed **60+**, or “60 and over”, has no upper boundary, so it has no midpoint, so Σ*fx* cannot be worked out until somebody decides what the top is. There are three honest ways to decide, and a fourth honest way to refuse. You may give the class **the same width as the others**, which keeps the table consistent. You may **stretch it to the largest value you actually know about**, if the sheet records one. You may **double the width** as a cautious reading when nothing is known about the top. Or you may decline to estimate a mean at all and report the **median and the modal class** instead, neither of which needs the top of the last class. What you may never do is pick a midpoint quietly. The choice moves the answer, so the choice goes in the answer: *“taking 60 ≤ x < 70, the estimated mean is...”*

Reading the open class · choice 1 of 3

Age, x (years)	Midpoint	Frequency, f
20 ≤ x < 30	25	7
30 ≤ x < 40	35	12
40 ≤ x < 50	45	16
50 ≤ x < 60	55	10
60 ≤ x < 70	65	5
Total		50

The sheet says 60+. This table has chosen to read it as 60 to 70.

One of three defensible readings of 60+, and the figure says so out loud. A midpoint chosen silently is the trap this concept exists to prevent.

Worked example 4 · The open end

1	Work out what the four closed classes contribute, once.	**Σfx so far = 1865**
2	Choice 1 — the same width as every other class: the class is 60 ≤ x < 70, midpoint 65.	**estimate 43.8 years**
3	Choice 2 — stretched to the oldest member recorded, 71: the class is 60 ≤ x < 72, midpoint 66.	**estimate 43.9 years**
4	Choice 3 — double width, when nothing is known about the top: the class is 60 ≤ x < 80, midpoint 70.	**estimate 44.3 years**
5	Compare the three answers.	**a spread of 0.5 years**
6	Write the answer so the reader can see the choice in it.	**“taking 60 ≤ x < 72…”**

*All three are defensible and they do not agree, so the choice has to be stated. Here the sheet records an oldest member of 71, which quietly rules choice 1 out: reading 60+ as 60 ≤ x < 70 puts at least one member in a class they are not in. Choice 2 uses what is known and is the best supported of the three, so its estimate is the one to quote. And notice the fourth option, always available: the **median class** and the **modal class** can be*

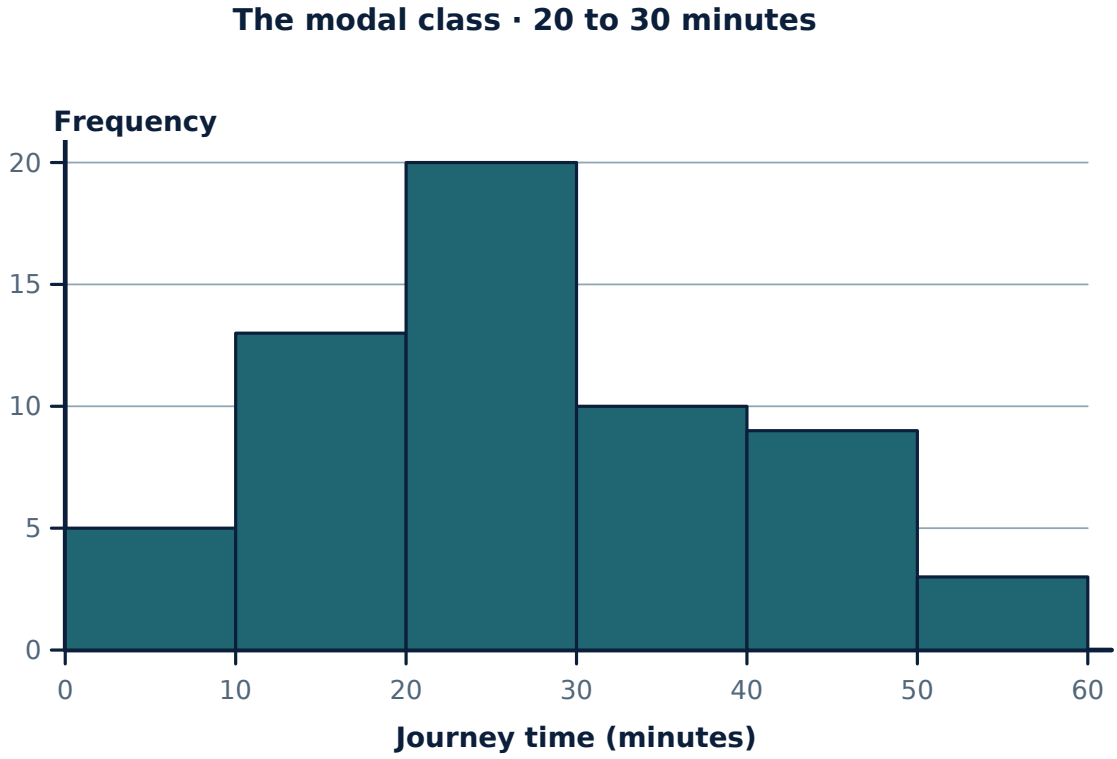
stated exactly without deciding anything about the top at all.

EXAMPLE. The closed classes contribute 1865, whatever happens next. Read 60+ as $60 \le x < 70$ and the estimated mean is **43.8 years**; stretch it to the oldest member, 71, and it is **43.9**; double the width and it is **44.3**. Three defensible choices, three different answers, spread over half a year — which is why the sentence that reports the answer has to name the choice that produced it.

CONCEPT 6 · MODE OR MODAL CLASS?

what continuous data can be asked

A **mode** is the value that occurs most often. Continuous data measured finely enough has no value occurring more than once, so it has **no mode** — and a question that asks for one has asked the wrong question. What it has instead is a **modal class**: the class with the highest frequency. Two disciplines come with it. First, the answer is a *class*, written $20 \le x < 30$, and never the frequency 20 and never the midpoint 25. Second, the modal class is only read off the bar heights when every class is **the same width**, which is the only kind of grouping this unit uses; with unequal widths a tall bar can simply be a wide one, and the repair for that belongs to S22. Like every other grouped answer, a modal class is a statement about a class and not about any learner in it.



The tallest bar names a class, never a single time.

The tallest bar names a class, 20 ≤ x < 30, and never a single journey time. Read it this way only because every class here is the same width.

Worked example 5 · Mode or modal class?

1	Ask whether any journey time occurs more than once.	**measured to 0.1 min, almost never**
2	So decide what can be asked of continuous data.	**a modal class, not a mode**
3	Find the highest frequency in the table.	**20**
4	Name the class it belongs to — the class, not the number.	**20 ≤ x < 30**
5	Check the classes are all the same width before trusting a bar.	**all 10 minutes wide** ✓
6	Regroup in twenties and answer the same question again.	**20 ≤ x < 40**

*The three wrong answers to “state the mode” are all common and all different: **20** is the frequency, **25** is the midpoint, and “there is no mode” on its own is true but unhelpful. The answer is the class 20 ≤ x < 30. And because the classes are a choice, so is the answer: regrouped in twenties*

the modal class becomes 20 ≤ x < 40, which is not a contradiction, just a coarser question. Nothing here is an estimate: a modal class is read straight off the frequencies, and needs no assumption at all.

> **EXAMPLE.** The modal class of the journey times is **20 ≤ x < 30**, the class holding 20 of the 60 learners. It is not 20, which is a frequency, and it is not 25, which is a midpoint. Regroup the same times in twenties and the modal class becomes **20 ≤ x < 40** — a different answer to the same question, from the same data, because the classes moved.

THE GROUPED-AVERAGE ROUTINE

1. Write the midpoint column first. Halfway between the two boundaries, one row at a time. Get these wrong and every number after them is wrong, and nothing in the working will look odd.

2. Multiply, add, divide - in that order. f times x for each row, add the column for Sigma fx, add the frequencies for Sigma f, then divide. Never divide by the number of rows.

3. Count to n/2 for the median class. A running total of frequencies, stopped the moment it reaches n/2. The class is exact; only the value read off the curve is estimated.

4. Say the word estimate, and say what you assumed. Three significant figures, the word estimate, and - if a class was open - the top you chose for it. An answer without those is incomplete even when the arithmetic is perfect.

THE MISTAKES THAT COST THE MOST MARKS

Trap 1 · Using the class boundaries instead of the midpoints

Wrong. Writing Σfx with the lower boundaries: 5 × 0 + 13 × 10 + 20 × 20 + … = 1340, then 1340 ÷ 60 = 22.3 minutes.

Right. Use the **midpoint** of each class: 5, 15, 25, 35, 45, 55. Then Σfx = 1640 and the estimated mean is 1640 ÷ 60 = **27.3 minutes**.

Lower boundaries assume everybody in a class took the shortest time in it, which drags the estimate down to 22.3 minutes; upper boundaries assume everybody took the longest, which pushes it up to 32.3. The true mean is 26.6, and it sits between them because the midpoint is the only value in a class with nothing to favour. The arithmetic looks perfect either way, which is what makes this the most expensive slip in the topic.

Trap 2 · Quoting the estimated mean as though it were exact

Wrong. “The mean journey time is 27.3333 minutes.”

Right. “The **estimated** mean journey time is **27.3 minutes**, to 3 significant figures.” It is an estimate, because the raw values are lost when the data is grouped.

Six figures claim an accuracy the table cannot support. The true mean of the 60 recorded times is 26.6 minutes, so every digit after the first three was invented by the division and not by the data. Two marks usually hang on this: one for the number, one for the word estimate and the sensible rounding that goes with it.

Trap 3 · Giving the open class a midpoint without saying so

Wrong. Reading 60+ as $60 \leq x < 70$, writing the estimated mean as 43.8 years, and saying nothing about where the 65 came from.

Right. State the assumption in the same sentence as the answer: “taking the last class as $60 \leq x < 72$, the estimated mean age is 43.9 years”. If the sheet records an oldest member — here 71 — use it, and the estimate becomes 43.9 years.

The three defensible readings of 60+ give 43.8, 43.9 and 44.3 years. A reader who is not told which one was taken cannot check the answer, cannot compare it with anybody else’s, and has been handed a number with a hidden decision inside it. Naming the choice costs one clause and turns a guess into a result.

PRACTICE · 60 QUESTIONS

Answers are at the back. Work each one out before you look.

Concept 1 · Classes are not optional

P1 (Fluency)	**Explain** in one sentence why continuous data is sorted into classes before an average is found.
P2 (Fluency)	The 60 journey times, in minutes, are grouped in tens: $0 \le x < 10$: 5; $10 \le x < 20$: 13; $20 \le x < 30$: 20; $30 \le x < 40$: 10; $40 \le x < 50$: 9; $50 \le x < 60$: 3. **State** the class that a journey of exactly 20 minutes belongs to.
P3 (Fluency)	The 60 journey times, in minutes, are grouped in tens: $0 \le x < 10$: 5; $10 \le x < 20$: 13; $20 \le x < 30$: 20; $30 \le x < 40$: 10; $40 \le x < 50$: 9; $50 \le x < 60$: 3. **State** the frequency of the class $20 \le x < 30$.
P4 (Fluency)	The 60 journey times, in minutes, are grouped in tens: $0 \le x < 10$: 5; $10 \le x < 20$: 13; $20 \le x < 30$: 20; $30 \le x < 40$: 10; $40 \le x < 50$: 9; $50 \le x < 60$: 3. **Calculate** how many of the 60 learners took less than 20 minutes.
P5 (Applied)	**Explain** why a journey of exactly 20 minutes is counted in $20 \le x < 30$ and not in $10 \le x < 20$.
P6 (Applied)	The 60 journey times, in minutes, are grouped in tens: $0 \le x < 10$: 5; $10 \le x < 20$: 13; $20 \le x < 30$: 20; $30 \le x < 40$: 10; $40 \le x < 50$: 9; $50 \le x < 60$: 3. **Calculate** the class width and **state** the number of classes.
P7 (Applied)	**Explain** why the 60 journey times have no mode, even though the table has a tallest class.
P8 (Applied)	The 60 journey times, in minutes, are grouped in tens: $0 \le x < 10$: 5; $10 \le x < 20$: 13; $20 \le x < 30$: 20; $30 \le x < 40$: 10; $40 \le x < 50$: 9; $50 \le x < 60$: 3. **Calculate** Σf and **explain** what it counts.

P9 (Challenge)	**Discuss** what is lost, and what is kept, when the 60 raw journey times are replaced by the six-class table.
P10 (Challenge)	The 60 journey times, in minutes, are grouped in tens: $0 \le x < 10$: 5; $10 \le x < 20$: 13; $20 \le x < 30$: 20; $30 \le x < 40$: 10; $40 \le x < 50$: 9; $50 \le x < 60$: 3. **Determine** the smallest and the largest possible total of all 60 journey times, using the table alone.

Concept 2 · Midpoints on parade

P11 (Fluency)	**Calculate** the midpoint of the class $20 \le x < 30$.
P12 (Fluency)	The class $20 \le x < 30$ holds 20 learners. **Calculate** $f \times x$ for that class.
P13 (Fluency)	**State** the formula for the estimated mean of grouped data.
P14 (Fluency)	**State** the one assumption the estimated mean makes about the values inside a class.
P15 (Applied)	The 60 journey times, in minutes, are grouped in tens: $0 \le x < 10$: 5; $10 \le x < 20$: 13; $20 \le x < 30$: 20; $30 \le x < 40$: 10; $40 \le x < 50$: 9; $50 \le x < 60$: 3. **Calculate** Σfx for this table.
P16 (Applied)	The 60 journey times, in minutes, are grouped in tens: $0 \le x < 10$: 5; $10 \le x < 20$: 13; $20 \le x < 30$: 20; $30 \le x < 40$: 10; $40 \le x < 50$: 9; $50 \le x < 60$: 3. **Estimate** the mean journey time, to 3 significant figures.
P17 (Applied)	**Explain** why the answer to the previous question is called an estimate and not the mean.
P18 (Applied)	A learner writes the estimated mean journey time as 27.3333 minutes. **Explain** what is wrong with the answer.

P19 (Challenge)	The three learners in the class $50 \le x < 60$ are removed from the table. **Determine** the new estimated mean, to 3 significant figures.
P20 (Challenge)	**Explain** why an estimated mean can never come out smaller than the smallest class midpoint or larger than the largest.

Concept 3 · Which class holds the median?

P21 (Fluency)	60 journey times are grouped. **State** the position you count to in order to find the median class.
P22 (Fluency)	The 60 journey times, in minutes, are grouped in tens: $0 \le x < 10$: 5; $10 \le x < 20$: 13; $20 \le x < 30$: 20; $30 \le x < 40$: 10; $40 \le x < 50$: 9; $50 \le x < 60$: 3. **State** the median class.
P23 (Fluency)	The 60 journey times, in minutes, are grouped in tens: $0 \le x < 10$: 5; $10 \le x < 20$: 13; $20 \le x < 30$: 20; $30 \le x < 40$: 10; $40 \le x < 50$: 9; $50 \le x < 60$: 3. **State** the running total of frequencies at the top of the class $10 \le x < 20$.
P24 (Fluency)	The 60 journey times, in minutes, are grouped in tens: $0 \le x < 10$: 5; $10 \le x < 20$: 13; $20 \le x < 30$: 20; $30 \le x < 40$: 10; $40 \le x < 50$: 9; $50 \le x < 60$: 3. **Calculate** how many learners took 30 minutes or more.
P25 (Applied)	The 60 journey times, in minutes, are grouped in tens: $0 \le x < 10$: 5; $10 \le x < 20$: 13; $20 \le x < 30$: 20; $30 \le x < 40$: 10; $40 \le x < 50$: 9; $50 \le x < 60$: 3. **Explain** how the median class is found by counting, and why that answer is exact.
P26 (Applied)	A cumulative frequency curve is drawn for the 60 journey times, using the running totals 5, 18, 38, 48, 57, 60. **Estimate** the median time from it.

P27 (Applied)	The curve estimates the median journey time at 26 minutes. The true median of the 60 recorded times is 25 minutes. **Explain** the gap.
P28 (Applied)	The 60 journey times, in minutes, are grouped in tens: $0 \le x < 10$: 5; $10 \le x < 20$: 13; $20 \le x < 30$: 20; $30 \le x < 40$: 10; $40 \le x < 50$: 9; $50 \le x < 60$: 3. **Calculate** the fraction of the median class you must go into to reach the 30th value, and turn it into minutes.
P29 (Challenge)	The same 60 journey times grouped in twenties: $0 \le x < 20$: 18; $20 \le x < 40$: 30; $40 \le x < 60$: 12. **Determine** the median class of this wider table, and **estimate** the median from it.
P30 (Challenge)	**Discuss** whether the median class or the estimated mean is the safer thing to quote from a grouped table.

Concept 4 · Change the width, change the answer

P31 (Fluency)	The 60 journey times are regrouped into classes of width 20. **State** how many classes there now are.
P32 (Fluency)	The same 60 journey times grouped in twenties: $0 \le x < 20$: 18; $20 \le x < 40$: 30; $40 \le x < 60$: 12. **State** the frequency of the class $20 \le x < 40$.
P33 (Fluency)	**Calculate** the midpoint of the class $0 \le x < 20$.
P34 (Fluency)	The same 60 journey times grouped in twenties: $0 \le x < 20$: 18; $20 \le x < 40$: 30; $40 \le x < 60$: 12. **Estimate** the mean journey time from this table, to 3 significant figures.
P35 (Applied)	The width-10 table estimates the mean at 27.3 minutes and the width-20 table at 28.0 minutes. **Calculate** the drift between them.

P36 (Applied)	**Explain** why regrouping the same measurements at a different class width changes the estimated mean.
P37 (Applied)	The true mean of the 60 recorded times is 26.6 minutes. **Compare** the two estimates with it.
P38 (Applied)	**Explain** which of the two groupings you would quote in a report, and why.
P39 (Challenge)	The 60 times are put into a single class, $0 \le x < 60$. **Determine** the estimated mean, and **explain** what has happened.
P40 (Challenge)	**Discuss** the trade-off in choosing a class width for continuous data.

Concept 5 · The open end

P41 (Fluency)	The 50 ages are grouped: $20 \le x < 30$: 7; $30 \le x < 40$: 12; $40 \le x < 50$: 16; $50 \le x < 60$: 10; and the last class is printed **60+**, holding 5 members. **State** how many members are in the class printed 60+.
P42 (Fluency)	The class 60+ is read as $60 \le x < 70$. **State** the midpoint that this gives.
P43 (Fluency)	The 50 ages are grouped: $20 \le x < 30$: 7; $30 \le x < 40$: 12; $40 \le x < 50$: 16; $50 \le x < 60$: 10; and the last class is printed **60+**, holding 5 members. **Calculate** Σf for the age table.
P44 (Fluency)	**Explain** why the class 60+ has no midpoint as it stands.
P45 (Applied)	The 50 ages are grouped: $20 \le x < 30$: 7; $30 \le x < 40$: 12; $40 \le x < 50$: 16; $50 \le x < 60$: 10; and the last class is printed **60+**, holding 5 members. **Estimate** the mean age, to 3 significant figures, taking the last class as $60 \le x < 70$.

P46 (Applied)	The record sheet notes that the oldest member is 71. **Estimate** the mean age, taking the last class as $60 \le x < 72$.
P47 (Applied)	**Estimate** the mean age once more, taking the last class as $60 \le x < 80$ instead.
P48 (Applied)	The three readings give 43.8, 43.9 and 44.3 years. **Explain** which one this record sheet supports best.
P49 (Challenge)	**Discuss** what should be reported if nothing at all is known about the top of an open final class.
P50 (Challenge)	The 5 members in the class 60+ are all aged at most 71. **Determine** the smallest and largest values the true mean age could possibly take.

Concept 6 · Mode or modal class?

P51 (Fluency)	The 60 journey times, in minutes, are grouped in tens: $0 \le x < 10$: 5; $10 \le x < 20$: 13; $20 \le x < 30$: 20; $30 \le x < 40$: 10; $40 \le x < 50$: 9; $50 \le x < 60$: 3. **State** the modal class.
P52 (Fluency)	**Explain**, in one sentence, why these journey times have no mode.
P53 (Fluency)	The same 60 journey times grouped in twenties: $0 \le x < 20$: 18; $20 \le x < 40$: 30; $40 \le x < 60$: 12. **State** the modal class of this wider table.
P54 (Fluency)	The 80 waiting times, in minutes, are grouped in tens: $0 \le x < 10$: 7; $10 \le x < 20$: 14; $20 \le x < 30$: 23; $30 \le x < 40$: 15; $40 \le x < 50$: 12; $50 \le x < 60$: 4; $60 \le x < 70$: 3; $70 \le x < 80$: 2. **State** the modal class of the waiting times.
P55 (Applied)	**Explain** the difference between the mode and the modal class.
P56 (Applied)	**Explain** why the modal class can change when the same data is regrouped.

P57 (Applied)	A learner answers “the modal class is 20”. **Explain** the mistake and give the correct answer.
P58 (Applied)	**Explain** why a modal class must not be read off bar heights when the classes have different widths.
P59 (Challenge)	**Discuss** whether the modal class is a useful average for the 60 journey times.
P60 (Challenge)	Two more learners are added to the 60 journey times. **Determine** whether the modal class could change, and **explain** your answer.

CRITERION B · INVESTIGATING PATTERNS

The midpoint bet

Forty learners were timed, to the nearest tenth of a minute, on a mental-arithmetic test. Every one of the forty times is printed below, so for once you can do the thing a grouped table normally makes impossible: work out the **true** mean, and then see how close the estimate from a grouped table comes to it. The bet is this. The estimated mean assumes every value sits at its class midpoint. Does that assumption cost you more when the classes are wide, or when they are narrow — and by how much?

The 40 test times, in minutes, in full

13.5, 15.7, 11.5, 17.2, 16.3, 11.8, 36.9, 16.0, 7.3, 28.0, 13.9, 32.8, 20.0, 22.3, 37.5, 15.7, 18.4, 22.6, 14.5, 5.8, 13.9, 27.0, 13.7, 18.6, 25.7, 22.3, 14.7, 18.5, 31.6, 24.9, 25.3, 28.7, 15.0, 9.3, 8.8, 18.2, 19.5, 3.8, 12.7, 26.7.

Part 1 · the structured investigation

1. **Predict**, before you calculate anything, which of the three class widths 5, 10 and 20 will give the estimate closest to the true mean, and which will be furthest away. Write the prediction down and do not change it afterwards.

2. **Calculate** the true mean of the forty printed times, exactly. This is the target every estimate is aiming at, and it never changes.

3. **Tabulate**, for each of the three widths, the grouped frequency table, the midpoints, Σfx, the estimated mean to 3 significant figures, and the size of the error — that is, the difference between the estimate and the true mean.

4. **Identify** the pattern linking the class width to the size of the error, and say whether the estimates fall on the same side of the true mean each time.

5. **Determine** a general statement connecting class width to the accuracy of the estimated mean, written as a sentence a classmate could apply to a dataset they have never seen.

6. **Predict** the estimated mean the same forty times would give at a class width of 2, using your general statement and before you work it

out. Then **test and verify** the prediction by building the width-2 table.

7. **Justify** the pattern from the midpoint assumption itself: say how far a value can possibly sit from its own midpoint in a class of width *w*, and use that to explain why a narrower class cannot do worse in the long run.

The table to fill in

Class width	Classes	Sigma fx	Estimated mean (3 s.f.)	Error
20				
10				
5				
2 - predict first				
True mean				

Every entry in this table is yours to compute from the printed times. Nothing in it has been given to you.

Part 2 · your own method choice

Part 2 · your own investigation. The general statement you have just justified is about the long run, and a single dataset can break it. **Construct** a dataset of your own — twenty values or so, anywhere in the range 0 to 40 — for which the **width-10** estimate is *closer* to the true mean than the **width-5** estimate is. Print your dataset in full, tabulate both groupings, and show the two errors. Then **explain** the fluke: what has to happen inside the classes for the coarser grouping to win, and why that is luck rather than a better method. You choose the values, the way you search for them and the way you present the result — but the reader must be able to check every number you claim.

Criterion B · Investigating patterns. Levels 1–2: groups the data at one or two widths and computes at least one estimated mean correctly. **Levels 3–4:** all three tables complete and correct, the true mean found, the three errors computed, and the pattern described in words. **Levels 5–6:** states a correct general statement linking class width to accuracy, and tests it on the width-2 table with the prediction made in advance. **Levels 7–8:** in addition, justifies the statement from the midpoint assumption — a value can sit at most half a class width from its midpoint, so halving the width halves the worst error any single value can contribute — and carries Part 2 through to a constructed dataset where the coarser grouping wins, with the cancellation that causes it explained rather than merely reported.

Remember, you should communicate in an organized and coherent manner.

CRITERION D · APPLYING MATHEMATICS IN REAL-LIFE CONTEXTS

What can the clinic honestly promise?

A clinic in Gaborone timed how long each of 80 patients waited between arriving and being called, one morning session, and grouped the times into classes of ten minutes. The sister in charge wants to put a notice at the door telling patients what to expect. She has the grouped table and a sheet of other information about the morning; the individual arrival slips were thrown away, so the actual waiting times no longer exist anywhere.

Morning session record sheet

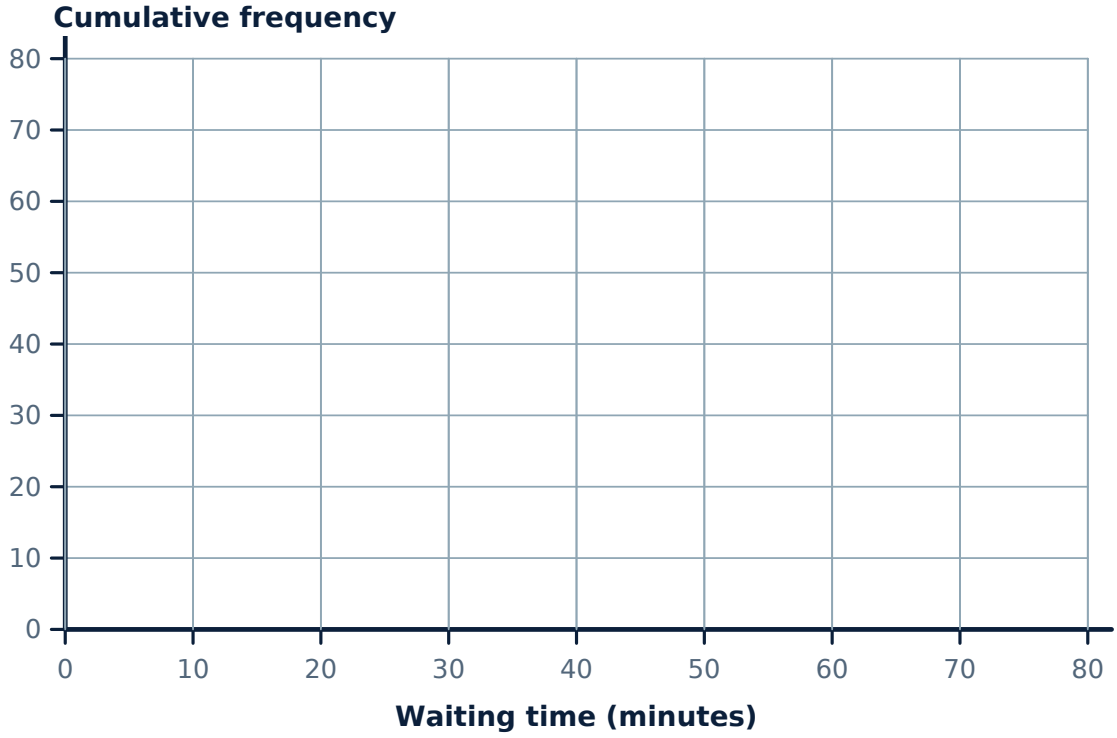
Entry	Recorded
Session	Tuesday morning, general consulting
Patients seen	80
Nurses on duty	3
Doctor on duty	1, from 09:00
Weather	hot and clear
Times recorded by	the reception clerk, to the nearest minute
Individual arrival slips	not kept — only the tally survives
Longest recorded wait	over 70 minutes

Waiting time, x (minutes)	Number of patients
$0 \le x < 10$	7
$10 \le x < 20$	14
$20 \le x < 30$	23
$30 \le x < 40$	15
$40 \le x < 50$	12
$50 \le x < 60$	4
$60 \le x < 70$	3
$70 \le x < 80$	2

Waiting time, x (minutes)	Number of patients
Total	80

This record sheet is an invented illustrative example written for the task. It is not taken from any business, price list or publication, and no organisation, person or source is named or implied.

Criterion D · plot the curve here



Blank axes for the Criterion D curve. Plot at the upper class boundaries, as S15 sets out, and show the guides.

1. **Interpreting the situation. Explain** which parts of the record sheet the waiting-time question actually needs, and which are surplus to it. For two of the surplus entries, say what question they would be the right data for instead.

2. **Selecting the mathematics. Explain** why the mean has to be estimated here rather than calculated, and why the modal class and the median class can be stated exactly. State which average you would put on a notice and why.

3. **Applying the mathematics. Calculate** the midpoint of every class and use them to **estimate** the mean waiting time, to 3 significant figures. **State** the modal class and the median class. Then **draw** the cumulative frequency curve on the printed axes and **estimate** the

median waiting time from it, showing the dashed read-off guides.

4. **Accuracy. Explain** what the word **estimate** costs this decision: state how far out the estimated mean could be if the values in each class are not evenly spread, and **discuss** which of your answers would survive the clinic regrouping its table into classes of twenty minutes instead of ten.

5. **Sense in context. Discuss** what the notice at the door should actually say. Consider whether a patient is better served by an average or by a statement such as “three quarters of patients are seen within … minutes”, what the 5 patients who waited an hour or more would make of the notice, and whether a promise should be made from a grouped table at all when the clinic could record the individual times from tomorrow.

Criterion D · Applying mathematics in real-life contexts. Levels 1–2: identifies some relevant information and begins the midpoint column. **Levels 3–4:** separates the needed data from the surplus, computes the estimated mean correctly and states the modal class as a class. **Levels 5–6:** estimated mean, modal class, median class and the curve estimate of the median all correct, the word estimate used and explained, and a defensible choice of which figure belongs on the notice. **Levels 7–8:** in addition, discusses the size and direction of the estimation error with reasons — the midpoint model assumes an even spread inside every class, and waiting times bunch — considers what regrouping would do to each answer, and reaches a defended recommendation about what the clinic may honestly promise and what it should start recording instead.

ANSWERS

Every answer here is recomputed from the datasets printed in the questions before this file is written. Every mean taken from a grouped table is an estimate quoted to 3 significant figures: a small window either side of the printed value is correct too.

P1	because measured values almost never repeat, so an ungrouped frequency table would have one row per value and tell you nothing	**P31**	3
P2	$20 \le x < 30$	**P32**	30
P3	20	**P33**	10
P4	18	**P34**	an estimate of 28.0 minutes
P11	25	**P41**	5
P12	500	**P42**	65
P13	$\Sigma fx \div \Sigma f$, where x is the class midpoint	**P43**	50
P14	that every value in the class sits at the midpoint of that class	**P44**	because it has no upper boundary, and a midpoint is halfway between two boundaries
P21	30	**P51**	$20 \le x < 30$
P22	$20 \le x < 30$	**P52**	because they are continuous measurements, so no value occurs more often than any other
P23	18	**P53**	$20 \le x < 40$
P24	22	**P54**	$20 \le x < 30$

Applied and Challenge · fully worked

P5 because the class 10 ≤ x < 20 stops below 20, while 20 ≤ x < 30 starts at it

Read the inequality signs. The class $10 \le x < 20$ accepts every value that is greater than or equal to 10 and strictly less than 20, so 20 itself is excluded from it and included in the class above. Classes that meet like this cover every possible measurement exactly once, with no gap and no overlap, which is the whole point of writing them that way.

P6 width 10 minutes, 6 classes

The width of a class is upper boundary minus lower boundary: $10 - 0 = 10$ minutes, and every class here is the same width, which is what lets the bars and the modal class be compared by height. Counting the rows gives 6 classes, and their frequencies add to 60, the number of learners.

P7 because the times are measured, so almost no value occurs twice; the tallest class gives a modal class, not a mode

A mode is a value that occurs most often. These times are continuous and were recorded to the nearest tenth of a minute, so repeats are accidental and mean nothing. What the table can give is the class with the highest frequency — the **modal class**, here $20 \le x < 30$ — and that is a statement about a class of learners, not about any single journey.

P8 60, the number of learners in the whole dataset

Add the frequency column downwards: 5 + 13 + 20 + 10 + 9 + 3. The total is 60, which is n — every learner is in exactly one class, because the classes meet and do not overlap. That total is the divisor in $\Sigma fx \div \Sigma f$, and it is also the free check that no learner has been counted twice or dropped.

P9 the individual times are lost; the shape of the distribution and the class totals are kept, so only an estimate of the mean is possible

Lost: every learner’s actual time, and with it the true mean, the true median, the range and any outlier. The table cannot name the fastest journey or the slowest — only the classes they fall in. Kept: how many learners sit in each interval, which is enough for the shape of the distribution, the modal class, the median class and an estimate of the mean. The trade is deliberate: six numbers a reader can hold in their head, bought with sixty that they could not.

P10 1340 minutes and 1940 minutes

Every learner in a class took at least the lower boundary and less than the upper one, so the total is at least 5 × 0 + 13 × 10 + 20 × 20 + 10 × 30 + 9 × 40 + 3 × 50 and less than 5 × 10 + 13 × 20 + 20 × 30 + 10 × 40 + 9 × 50 + 3 × 60. Dividing by 60, the true mean must lie between 22.3 and 32.3 minutes. The midpoint estimate of 27.3 sits between those two limits, which is the strongest guarantee the grouped table can give: not the right answer, but an answer that cannot be far wrong.

P15 1640

Find the six midpoints first: 5, 15, 25, 35, 45, 55. Multiply each by its frequency to get the fx column: 25, 195, 500, 350, 405, 165. Adding that column gives Σfx = 1640 minutes — the total journey time the midpoint model assumes, not the total anybody actually travelled.

P16 27.3 minutes

Σfx = 1640 and Σf = 60, so the estimated mean is 1640 ÷ 60 = 27.3 minutes to 3 significant figures. Say the word: it is an estimate, because the raw values are lost when the data is grouped.

P17 because the individual times were thrown away by grouping, every learner in a class has to be treated as taking the midpoint time, so the answer can only be an estimate

The table no longer holds a single journey time, so a true total cannot be formed. The midpoint is used as a stand-in for every value in its class, which is right on average only if the values in that class are spread evenly about it — and real measurements bunch. The answer is therefore an estimate, because the raw values are lost when the data is grouped. Here the true mean of the 60 recorded times is 26.6 minutes against the estimate of 27.3.

P18 it is quoted to six significant figures, which claims an accuracy an estimate cannot have; write 27.3 minutes

The division does produce those digits, but they come from the arithmetic, not from the data: the table cannot locate any learner closer than the nearest 10 minutes. Three significant figures is what an estimate off a grouped table can support, so the answer is 27.3 minutes, and the word estimate belongs in the sentence too.

P19 25.9 minutes

Removing that class takes 3 × 55 = 165 off Σfx and 3 off Σf. The new totals are 1475 and 57, so the estimate becomes 1475 ÷ 57 = 25.9 minutes. It has fallen by about 1.46 of a minute, because the three longest journeys were pulling the mean up — and it is still an estimate, because the raw values are lost when the data is grouped.

P20 because it is a weighted average of the midpoints, so it must lie between the smallest and the largest of them

Σfx ÷ Σf is the mean of the midpoints, each one counted as many times as its frequency. Replacing every midpoint by the smallest, 5, would give exactly 5; replacing every one by the largest, 55, would give exactly 55. The real mixture must sit between those two, so the estimate is trapped in the interval 5 to 55. It is a cheap check worth making: an estimated mean outside that range is an arithmetic error, every time.

P25 run the frequencies as a running total until it first reaches 30; the class you stop in is the median class, and the table really does say the middle value lies in it

The running total is 5, 18, 38, 48, 57, 60. It is still 18 at the top of 10 ≤ x < 20 and reaches 38 at the top of 20 ≤ x < 30, so values 19 to 38 all lie in that class — and the 30th value is one of them. No assumption about how the values are spread has been used, so the median **class** 20 ≤ x < 30 is exact. Only a single median **value** would have to be estimated.

P26 26 minutes

Go in at n ÷ 2 = 30 on the cumulative frequency axis, across to the curve and down. The climb across the median class runs from 18 to 38, a rise of 20 learners, and 30 is 0.6 of the way up it, so the median is 20 + 0.6 × 10 = 26 minutes. It is an estimate, because the raw values are lost when the data is grouped.

P27 the curve has to assume the 20 times in the median class are spread evenly across it, and they are not

Between two plotted points the curve is a straight line, which is exactly the assumption that the values inside the class are evenly spread. In this dataset they bunch towards the lower end of 20 ≤ x < 30, so the even-spread model reads a little high: 26 against a true 25, a gap of 1 of a minute. The gap is not an error in the method — it is the price of grouping, and it is why the answer is an estimate, because the raw values are lost when the data is grouped.

P28 0.6 of the class, which is 6 minutes

18 learners had been counted before the median class and 20 are inside it, so the 30th value is the 12th of those 20: 30 − 18 = 12 of the way in, which is 0.6 of the class. The class is 10 minutes wide, so that is 0.6 × 10 = 6 minutes past 20, giving a median of about 26.

P29 20 ≤ x < 40, median about 28 minutes

The running totals are 18, 48, 60, so the total first reaches 30 in the class 20 ≤ x < 40 — a wider, less useful answer than the width-10 table gave. Interpolating across it, 30 is 0.4 of the way through a class 20 minutes wide, so the median is estimated at 28 minutes. The width-10 table put it at 26 and the true median is 25: the wider the class, the further an interpolated median can drift, and both answers are an estimate, because the raw values are lost when the data is grouped.

P30 the median class is exact and needs no assumption, so it is safer; the estimated mean is more useful but rests on the midpoint assumption

The median class is read straight off the running totals and is a fact the table states, so nothing can go wrong with it beyond arithmetic. The estimated mean needs every value in a class to be replaced by its midpoint, so it can be pulled about by a skewed class or a wide one, and an open final class can stop it being computable at all. Against that, the mean uses every observation and can be combined and compared, which a class label cannot. A good report gives both, and states which of them is an estimate, because the raw values are lost when the data is grouped.

P35 0.667 of a minute

28 − 27.3 = 0.667 minutes to 3 significant figures. Neither table is wrong and the learners did not change their journeys: the drift is entirely the grouping’s doing, because a wider class moves every midpoint and therefore every product fx. Both numbers are an estimate, because the raw values are lost when the data is grouped.

P36 because every value is replaced by its class midpoint, and changing the classes changes which midpoint each value is replaced by

Take the 30 learners in 20 ≤ x < 40. In the width-10 table they are split between midpoints 25 and 35; in the width-20 table all 30 of them are assumed to have taken 30 minutes. Different stand-in values give a different Σfx, so the estimate moves. The measurements never moved, and neither did their true mean of 26.6 minutes.

P37 the width-10 estimate is 0.708 minutes out and the width-20 estimate is 1.375 minutes out, so the narrower grouping is the better estimate

Width 10: 27.3 − 26.6 = 0.708 minutes too high. Width 20: 28.0 − 26.6 = 1.375 minutes too high. Both overestimate, because in this dataset the times bunch below the midpoints of the crowded classes, and the wider table overestimates by nearly twice as much. Narrower classes leave each value closer to its stand-in, so they estimate better — at the cost of a longer table.

P38 the width-10 grouping, because narrower classes keep each value closer to its midpoint and so estimate the mean more closely

The width-10 table gives 27.3 minutes against a true 26.6; the width-20 table gives 28.0. Six classes are still small enough to read at a glance, so nothing is gained by the coarser table except a shorter page. The argument would change if the classes became so narrow that most held one or two learners: then the table stops summarising anything. Either way the figure quoted is an estimate, because the raw values are lost when the data is grouped.

P39 30.0 minutes — the midpoint of the only class, so the data has stopped contributing anything

With one class the midpoint is 30 and every one of the 60 learners is assumed to have taken it, so $\Sigma fx = 60 \times 30 = 1800$ and the estimate is exactly 30.0 minutes. The answer no longer depends on the data at all — the same figure would come out for any 60 times between 0 and 60. It is the limiting case of the width rule: as the classes widen the estimate learns less, until at one class it has learned nothing.

P40 narrow classes estimate better but make a long table with small, noisy frequencies; wide classes read easily but blur the data and worsen the estimate

Every value sits at most half a class width from its midpoint, so halving the width halves the worst possible error each value can contribute — that is the case for narrow classes. Against it: with 60 measurements, classes of width 2 would give 30 rows holding one or two learners each, which is a list wearing a table’s clothes, and a modal class chosen from such frequencies would be pure chance. The usual compromise is between five and about a dozen classes, chosen so that the shape shows and the frequencies are worth counting. Whatever is chosen, the resulting mean is an estimate, because the raw values are lost when the data is grouped.

P45 43.8 years

The four closed classes contribute 1865, and the last class contributes 5 × 65 = 325. So Σfx = 2190, Σf = 50, and the estimated mean is 2190 ÷ 50 = 43.8 years. The sentence must carry the choice: taking 60+ as 60 ≤ x < 70. It is an estimate, because the raw values are lost when the data is grouped.

P46 43.9 years

The midpoint of 60 ≤ x < 72 is (60 + 72) ÷ 2 = 66, so the last class contributes 5 × 66 = 330 and Σfx = 2195. Dividing by 50 gives 43.9 years. This is the best supported of the three readings, because it uses a fact the sheet actually records instead of a convention. It is still an estimate, because the raw values are lost when the data is grouped.

P47 44.3 years

Now the midpoint of the last class is 70, so it contributes 5 × 70 = 350, Σfx = 2215 and the estimate is 44.3 years. Doubling the width of the open class has moved the answer by 0.5 of a year against choice 1 — five members out of fifty, each shifted 5 years, is enough to matter. It is an estimate, because the raw values are lost when the data is grouped.

P48 the reading 60 ≤ x < 72, because the sheet records an oldest member of 71 and that is a fact rather than a convention

Choice 1 keeps the class widths equal, which is the usual default, but here it is known to be wrong: reading 60+ as 60 ≤ x < 70 would put the oldest member, aged 71, in a class they are not in. Choice 3 is the cautious reading for a sheet that says nothing about the top, and this sheet does say something. So choice 2, 60 ≤ x < 72, is the best supported, giving 43.9 years — and the answer names the choice.

P49 report the median class and the modal class, which do not need the top at all, and give an estimated mean only with the assumption stated

Neither the median class nor the modal class uses the upper boundary of the last class: both are read off frequencies alone, so both can be stated exactly. An estimated mean cannot be computed without inventing a top, so either it is given with the invention named — “taking the last class as twice the width of the others” — or a range is given by computing it under two different readings. Quoting a single mean with no assumption stated is the one unacceptable option, because the reader cannot tell how much of the answer is data and how much is choice.

P50 between 38.8 years and 48.9 years

Put every member at the bottom of their class: 1940, a mean of 38.8 years. Put every one at the top, with the five oldest at 71: 2445, a mean of 48.9 years. The true mean of the 50 ages must lie between them, and the three estimates of 43.8, 43.9 and 44.3 years all sit inside that window — which is the check that none of the three readings of 60+ is absurd. The window is 10.1 years wide, so it settles very little on its own; the estimate is worth having precisely because it is much tighter than the guarantee.

P55 the mode is a single value that occurs most often; the modal class is the class with the highest frequency, and is an interval

A mode is a value: for a list of shoe sizes it is a size somebody actually wore. A modal class is an interval: for the journey times it is $20 \leq x < 30$, the class holding 20 of the 60 learners. Discrete data can have either; continuous data measured finely can only have a modal class, because no single value repeats. The answer to “state the modal class” is therefore always written as a class, never as a frequency and never as a midpoint.

P56 because the classes themselves change, so a different interval can end up holding the most values

Grouped in tens the modal class is $20 \leq x < 30$, holding 20 learners. Grouped in twenties, 30 of them fall in $20 \leq x < 40$, which becomes the modal class. Nothing about the journeys changed — the boundaries moved, and the boundaries are a choice. It follows that a modal class should always be quoted together with the grouping that produced it, and that two modal classes from two different tables cannot be compared directly.

P57 20 is the frequency, not the class; the modal class is $20 \leq x < 30$

The number 20 answers a different question: how many learners are in the modal class. The modal class itself is the interval those learners sit in, $20 \leq x < 30$. The other common wrong answer is the midpoint, 25, which is a single value the class was replaced by for the mean and is not the class either. Write the inequality out in full and the mistake cannot happen.

P58 because a wide class collects more values simply by being wide, so a tall bar need not mean a crowded interval

Every class in this unit is the same width, which is what makes the tallest bar the modal class. Widen one class to twice the others and it will tend to hold about twice as many values without being any more crowded, so its bar wins a comparison it should not be in. The repair — plotting frequency per unit width instead of frequency — belongs to S22 *Bars With Meaning*; here the rule is simply to check the widths are equal before trusting a height.

P59 it is useful as a quick statement of where the data crowds, but it is coarse, it depends on the grouping, and it ignores everything outside the tallest class

In its favour: $20 \le x < 30$ holds 20 of the 60 learners, so it does describe the typical journey, it needs no arithmetic and no assumption, and it survives an open final class. Against it: it is an interval 10 minutes wide rather than a value, it changes to $20 \le x < 40$ if the data is regrouped in twenties, and it says nothing about the 40 learners outside it — including the three whose journeys are over 50 minutes. For this dataset the estimated mean of 27.3 minutes and the median class carry more information, and the honest report gives all three.

P60 no — the modal class must still be $20 \le x < 30$

The modal class holds 20 learners and the next largest class holds 13. Two extra learners can raise a rival to at most $13 + 2 = 15$, which is still below 20, and if either of them joins the modal class its lead only grows. So no arrangement of two additions can change the answer, and it stays $20 \le x < 30$. Notice how the question was settled: by the worst case for the current winner, not by trying examples.