Statistics & Probability · Unit S22 · Extended

Bars With Meaning

A bar chart is a picture of categories, and its gaps mean something: nothing lives between Maths and Setswana. A histogram is a picture of a measured scale, and its bars have to touch, because something lives everywhere along that scale. Once the bars touch, the outline they make has a name, a shape and a story — and this unit is about reading that story honestly, and about the one decision that can quietly rewrite it: how wide you make the classes.

Extended and Enrichment · about 55 minutes · 60 practice questions · Criteria A, B, C and D

BY THE END OF THIS UNIT YOU CAN

• **Draw** a histogram for continuous data grouped in equal fixed intervals, with the bars touching and both axes labelled

• **Justify** why the bars of a histogram touch and the bars of a bar chart do not, from the type of data each one shows

• **Identify** the shape of a distribution — symmetrical, skewed to the right, skewed to the left or bimodal — and **justify** the choice from the histogram itself

• **Explain** what each shape says about the measurements that made it

• **Construct** a frequency polygon by joining the midpoints of the class tops, and **overlay** two of them to compare two distributions

• **Justify** a comparison of two distributions, including why the two charts must use the same classes and the same axes

• **Determine** the effect of changing the class width on the shape a histogram shows, and **justify** a sensible choice of width

• **Explain** why a histogram states a modal *class* and why the tallest bar is not the average

THE ONE IDEA BEHIND ALL OF IT

A histogram is not a bar chart with the gaps removed. It is a different claim about a different kind of data. Bar chart: separate categories, separate bars, and the gaps say so. Histogram: one continuous

measured scale cut into **equal fixed intervals** that meet, so the bars touch because the classes touch — a gap would claim that no measurement can land there, and that is false. Once the bars touch, the outline becomes information: its shape names how the measurements are spread, and joining the midpoints turns that outline into a line you can lay over another one. But the shape is not handed to you by the data alone. You chose the class width, and a different width can turn one hump into two, or two into none. Choosing it, and saying why you chose it, is part of the answer.

bars touch ⇔ the scale is continuous

and the shape you see is partly the width you chose

WHERE THIS UNIT SITS

What you already have. S17 *Averages, Continued* owns the grouped machinery: class boundaries, the rule at the join, midpoints, Σ*fx*, the estimated mean and the median class. None of it is retaught here. This unit borrows exactly one piece — the **midpoint** of a class — because a frequency polygon is built on midpoints, and it sends you back to S17 for everything else.

Shape has been building for a while. S14 *Box and Whisker* started the shape language: a box pushed to one end of its whiskers is already telling you about skew. A histogram says the same thing with far more detail, and this unit gives the shapes their names.

Honesty. S12 *Lies and Statistics* set the rule this unit leans on hardest: a chart that changes its scale between two pictures is not comparing anything. Here that becomes concrete — two histograms may only be compared when the classes and the frequency axis are identical, and this unit refuses the comparison otherwise.

Where the numbers come from. Every dataset on this page was produced once by a seeded generator in this unit’s build script and then grouped by *counting* the raw values into classes. The 60 pumpkin masses and the 80 seedling heights are printed below in full, which is what lets you regroup them yourself and check that the shape really does change.

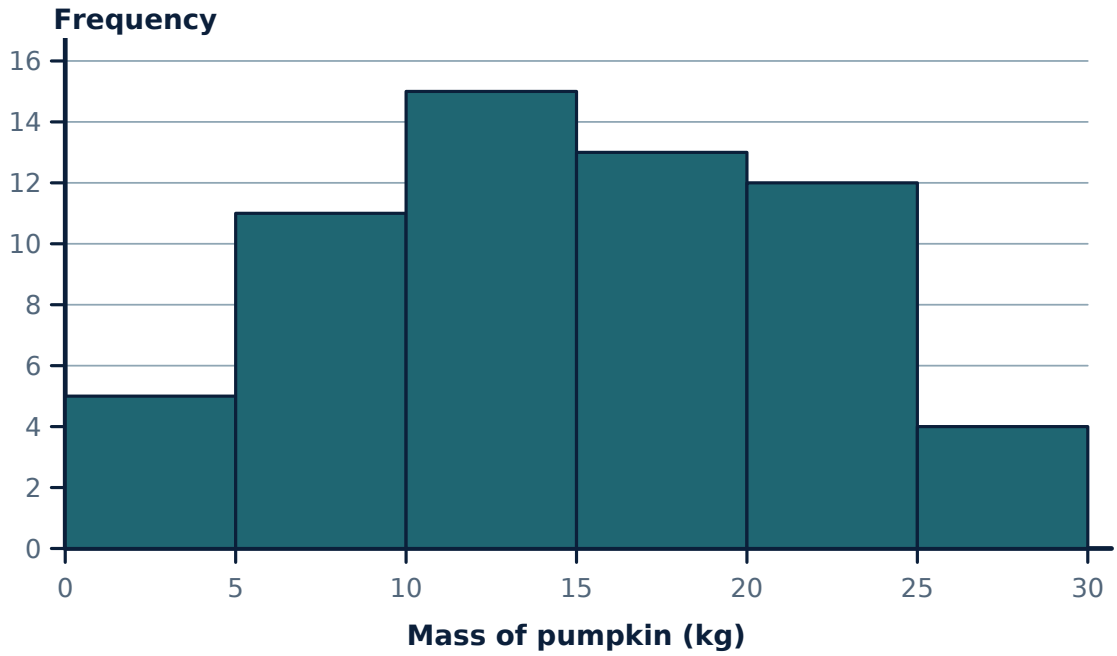
CONCEPT 1 · BARS THAT TOUCH

the histogram for continuous data with equal fixed intervals

Continuous data is **measured**, not counted, so it can take any value along a scale. Grouping it into **equal fixed intervals** that meet — $0 \le x < 5$, $5 \le x < 10$, and so on — is how it becomes a table, and S17 set that up. The histogram is what that table looks like. Along the bottom runs the measured scale itself, marked at the class *boundaries*, not at category names. Up the side runs frequency, starting at 0. Each bar stands over its own class, from boundary to boundary, at the height of its frequency — and because the top of one class is the bottom of the next, the bars touch because the classes touch. A gap would be a claim, and the claim would be false: it would say that no pumpkin in the field weighs 5 kg, when in fact 11 of them weigh between 5 and 10 kg. Two more rules follow from the same idea. The intervals must all be the **same width**, or the heights stop being comparable; and the frequency axis must start at **0**, or the heights stop being frequencies.

EXAMPLE · the 60 pumpkin masses, in kilograms, exactly as they were recorded. 18.7, 7.8, 4.6, 17.8, 9.5, 22.7, 12.9, 17.5, 17.2, 10.2, 21.5, 20.3, 16.1, 18.3, 5.2, 16.7, 12.4, 10.6, 13.2, 23.2, 20.5, 1.8, 9.6, 10.7, 12.0, 13.0, 14.4, 8.1, 24.8, 20.3, 8.6, 15.7, 8.4, 21.3, 28.5, 20.4, 9.5, 13.7, 16.8, 11.5, 8.0, 11.6, 2.3, 18.6, 14.8, 5.4, 25.4, 11.9, 20.7, 23.0, 8.8, 16.0, 3.9, 14.9, 16.4, 28.7, 25.6, 0.8, 24.5, 19.5. Counting them into 6 classes of width 5 gives the frequencies 5, 11, 15, 13, 12, 4.

60 pumpkin masses · classes 5 kg wide



Six classes, each 5 kg wide, and every one meets the next.

The chart the rest of the unit works from. The mass scale runs along the bottom marked at the class boundaries, the frequency axis starts at 0, and every bar meets its neighbours.

Worked example 1 · Touching bars

1	Check the data type before drawing anything.	**measured in kg → histogram**
2	Check the classes: equal fixed intervals that meet.	**6 classes, all 5 kg wide ✓**
3	$0 \le x < 5$ holds 5 pumpkins: a bar from 0 to 5.	**height 5**
4	$5 \le x < 10$ holds 11 pumpkins: a bar from 5 to 10.	**height 11**
5	$10 \le x < 15$ holds 15 pumpkins: a bar from 10 to 15.	**height 15**
6	$15 \le x < 20$ holds 13 pumpkins: a bar from 15 to 20.	**height 13**
7	$20 \le x < 25$ holds 12 pumpkins: a bar from 20 to 25.	**height 12**
8	$25 \le x < 30$ holds 4 pumpkins: a bar from 25 to 30.	**height 4**
9	Add the frequencies as a check, then label the mass axis at the boundaries 0, 5, 10, 15, 20, 25, 30 and the frequency axis from 0.	**60 = n ✓**

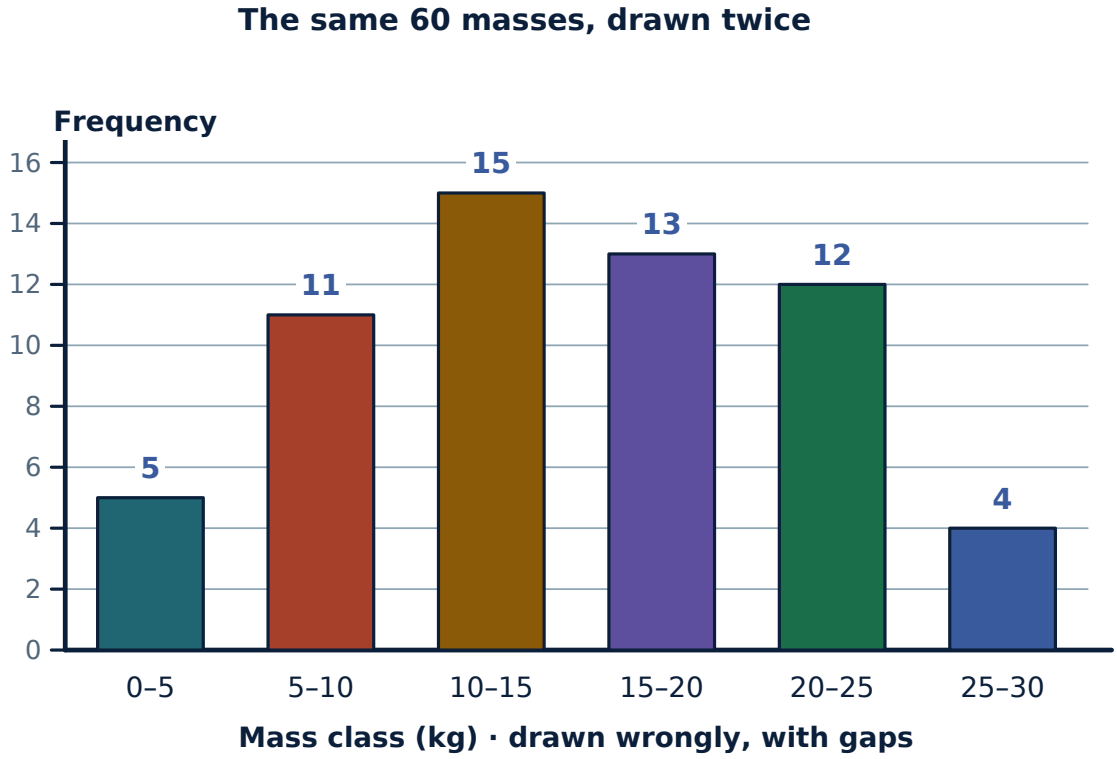
Justify the touching. *The bar for $5 \le x < 10$ runs to 10 and the bar for $10 \le x < 15$ starts at 10, so they share an edge and there is nowhere for a gap to go. That is not a drawing convention: a gap at 10 kg would be a claim that no pumpkin in this field can weigh 10 kg, and 15 of them weigh between 10 and 15 kg. In one sentence, the bars touch because the classes touch — and because every class here is 5 kg wide, the heights may be compared directly, which is what makes the tallest bar mean anything at all.*

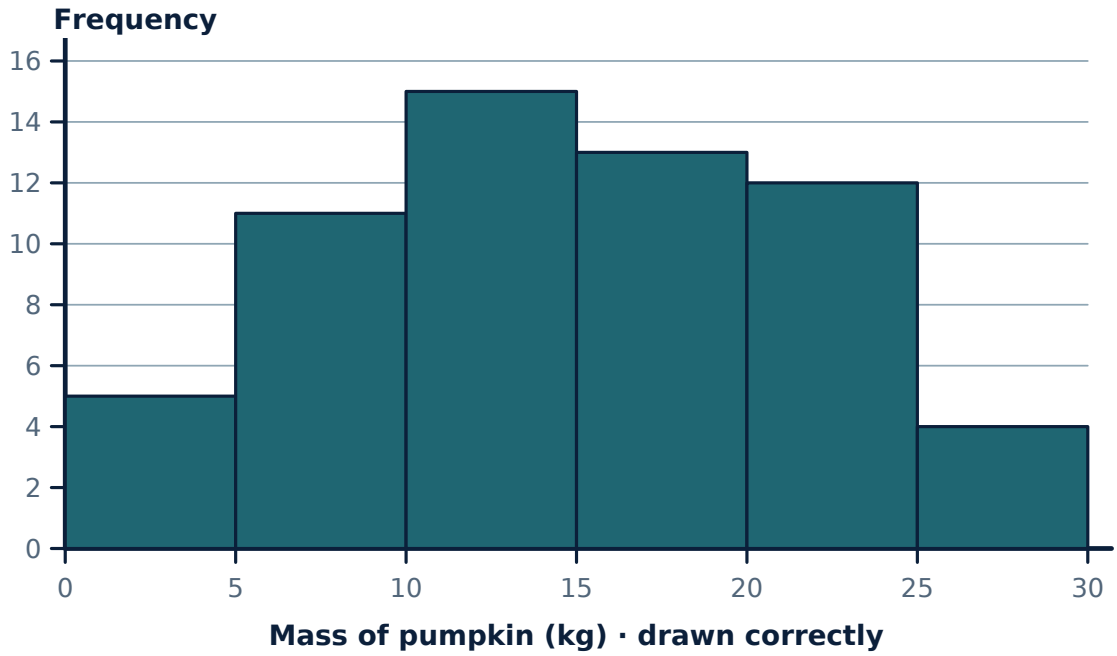
EXAMPLE. The class $5 \leq x < 10$ holds 11 of the 60 pumpkins, so its bar is drawn from 5 to 10 along the mass axis at a height of 11. The next bar begins exactly where that one stops. A pumpkin of exactly 10 kg is counted in the class above — the inequality signs say so — and on the chart it sits on the line where the two bars meet, which is why there is no room for a gap.

CONCEPT 2 · BAR CHART OR HISTOGRAM?

the data type chooses the chart

Both charts draw rectangles whose heights are frequencies, and there the resemblance stops. A **bar chart** shows data that comes in separate lumps: categories with no order (favourite subject), or whole numbers that are counted (children in a household). Between one lump and the next there is genuinely nothing, so the bars are drawn apart and the gaps are part of the message. A **histogram** shows data that was measured on a scale where every value in between is possible, so the bars touch because the classes touch. The test is one question asked of the variable, before you draw anything: *could a value fall between two of the labels I am about to write?* Between Maths and Setswana, no — bar chart. Between 5 kg and 10 kg, yes, endlessly — histogram. Notice that counted numbers such as “2 children” look numeric but behave like categories: nobody has 2.4 children, so the gaps stay.

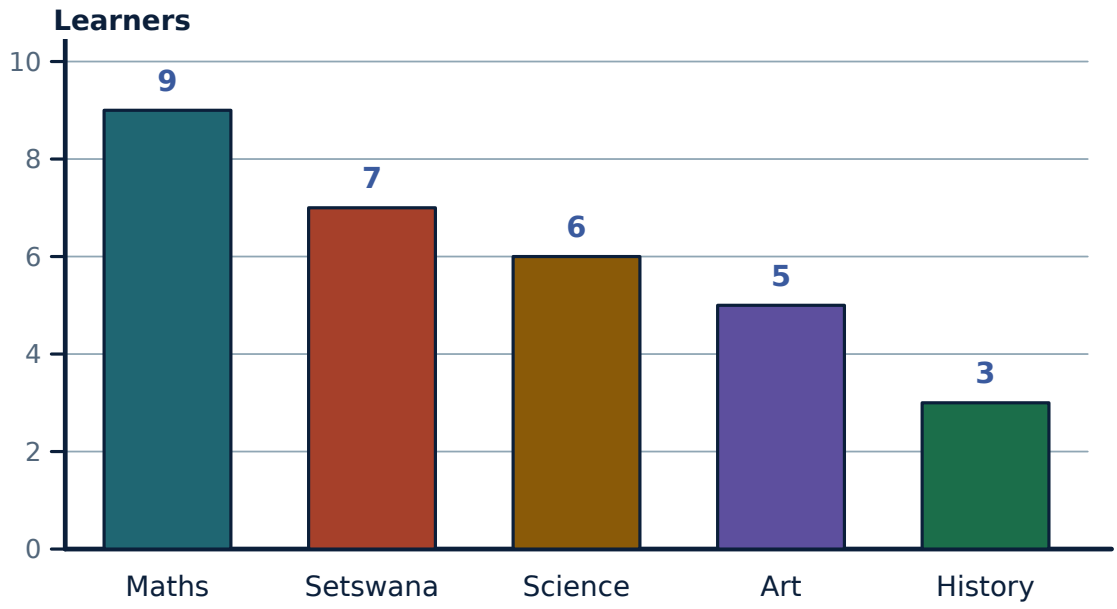




Above: gaps. Below: the bars touch because the classes touch.

The same six frequencies, drawn as a bar chart and as a histogram. Only one of them is telling the truth about a measured quantity, and the difference is the gaps.

Bar chart or histogram?



Favourite subject · a real bar chart, gaps and all

Which chart does the data choose?

Bar chart Favourite subject	**Bar chart** Children per home	**Histogram** Mass of a pumpkin	**Histogram** Mass of a tomato

The data type decides, not the chart you feel like drawing.

A real bar chart above — the gaps are information, because nothing exists between Maths and Setswana. Below, the same decision made for four datasets at once.

Worked example 2 · Bar chart or histogram?

1	**Dataset 1** — favourite subject of 30 learners: Maths: 9; Setswana: 7; Science: 6; Art: 5; History: 3.	**categorical → bar chart**
2	Why: could a value fall between two labels?	**no — gaps**
3	**Dataset 2** — number of children in 30 households: 0: 4; 1: 7; 2: 9; 3: 6; 4: 3; 5: 1.	**discrete → bar chart**

4	Why: is 1.5 children possible?	**no — gaps**
5	**Dataset 3** — The 60 pumpkin masses, in kilograms, grouped in fives: 0 ≤ x < 5: 5; 5 ≤ x < 10: 11; 10 ≤ x < 15: 15; 15 ≤ x < 20: 13; 20 ≤ x < 25: 12; 25 ≤ x < 30: 4.	**continuous → histogram**
6	Why: is 12.7 kg possible?	**yes — bars touch**
7	**Dataset 4** — The 60 early-planting tomato masses, in grams, grouped in tens: 0 ≤ x < 10: 2; 10 ≤ x < 20: 4; 20 ≤ x < 30: 12; 30 ≤ x < 40: 16; 40 ≤ x < 50: 14; 50 ≤ x < 60: 10; 60 ≤ x < 70: 2; 70 ≤ x < 80: 0; 80 ≤ x < 90: 0; 90 ≤ x < 100: 0.	**continuous → histogram**
8	Why: is 36.4 g possible?	**yes — bars touch**

Justify each verdict from the variable, not from the table. *Datasets 1 and 2 are drawn as bar charts with gaps because their values come in separate lumps — there is no subject between Maths and Setswana, and no household with 1.5 children — so the gap is true information. Datasets 3 and 4 were measured on continuous scales where every intermediate value exists, so they are grouped into equal fixed intervals and drawn as histograms in which the bars touch because the classes touch. The give-away is never how the numbers look on the page: 2 and 41 are both whole numbers, and one of them was counted while the other was rounded from a measurement.*

EXAMPLE. Thirty learners named a favourite subject: Maths: 9; Setswana: 7; Science: 6; Art: 5; History: 3. That is a bar chart — there is no scale running from Maths to History, and no value in between. The same thirty households reported 0, 1, 2, 3, 4 or 5 children: 0: 4; 1: 7; 2: 9; 3: 6; 4: 3; 5: 1. Still a bar chart, because 1.5 is impossible. But the mass of a pumpkin can be anything at all between 0.8 and 28.7 kg, so it is grouped and drawn as a histogram.

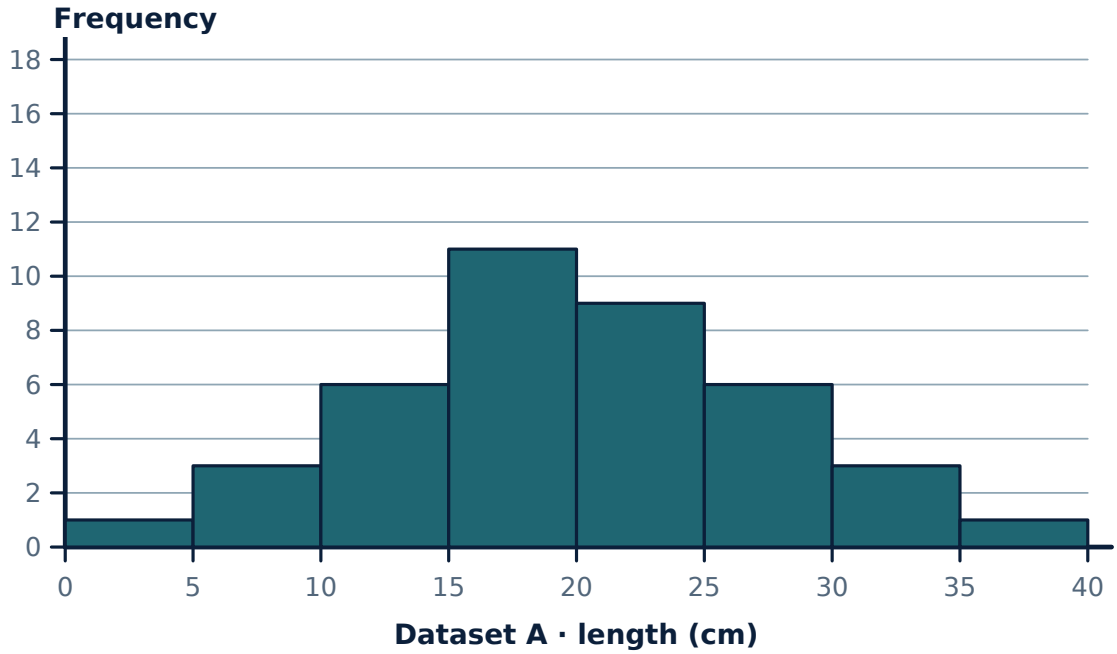
CONCEPT 3 · READING THE SHAPE

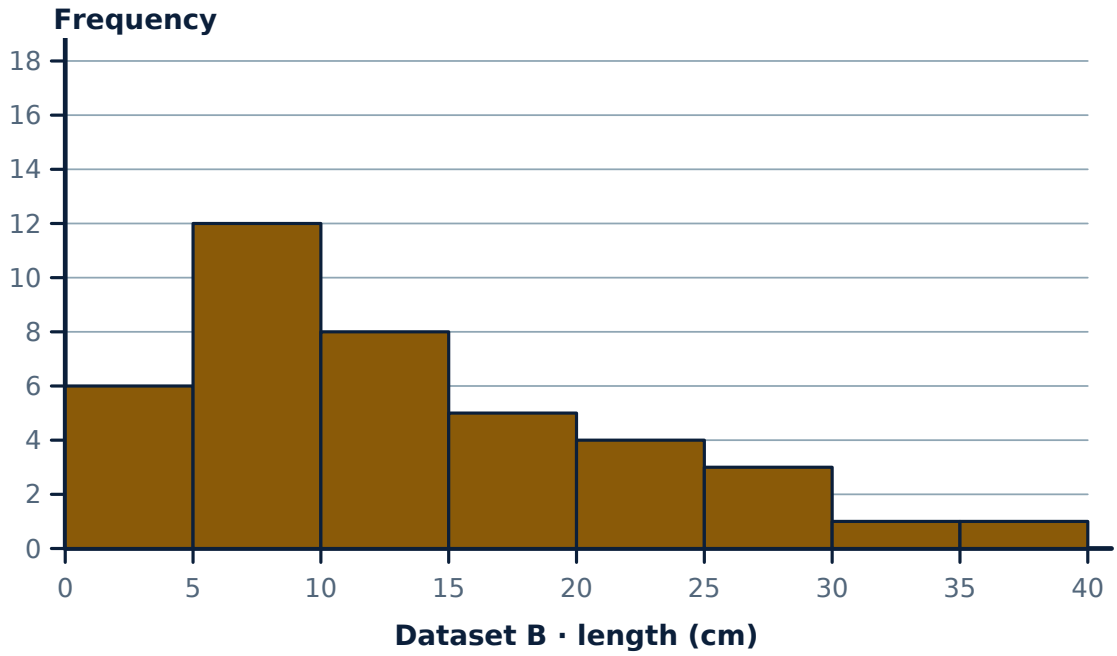
symmetrical, skewed, bimodal - and the story each one tells

Once the bars touch, run your eye along their tops. That outline is the **shape** of the distribution, and it has four names worth knowing. **Symmetrical**: one hump, and the two sides mirror each other — most measurements near the middle, fewer as you go either way. **Skewed to the right** (also called positively skewed): one hump low down the scale

and a long thin tail stretching to the right — most values small, a few much larger. **Skewed to the left**: the mirror image, the tail running down to the low end. **Bimodal**: two separate humps with a dip between them, which almost always means two different groups have been mixed into one dataset. The name is only half the answer. The other half is the story: a right-skewed histogram of incomes says most people earn little and a few earn a great deal; a bimodal histogram of seedling heights says somebody planted two batches two weeks apart. And say the modal **class**, never a single value — a continuous measurement has no mode.

Shape gallery 1 · datasets A and B

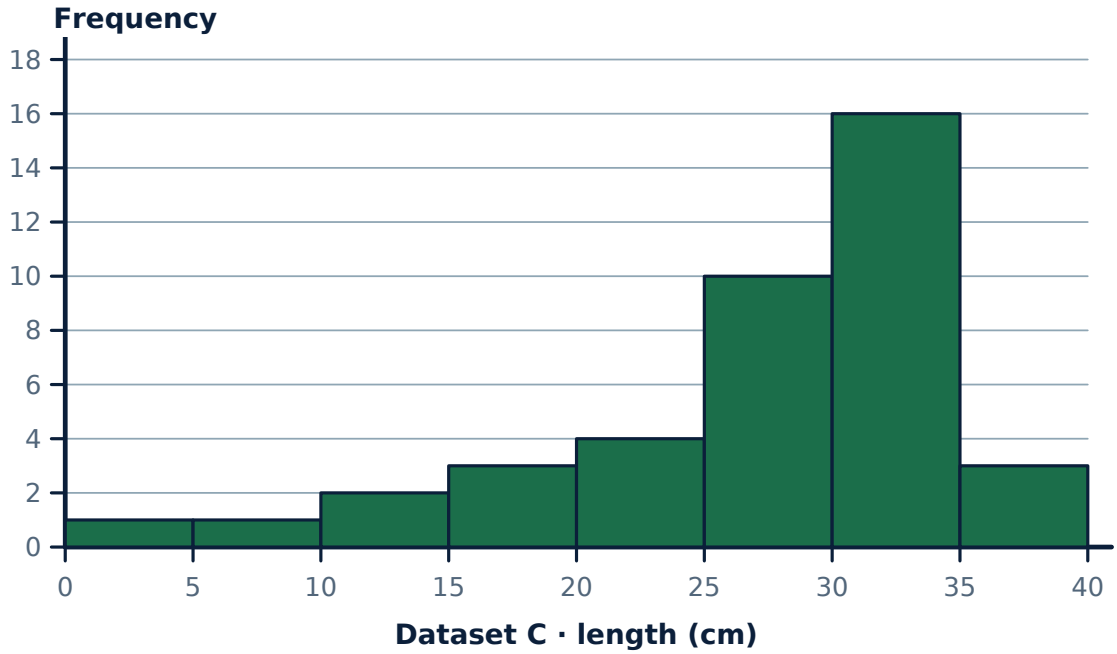


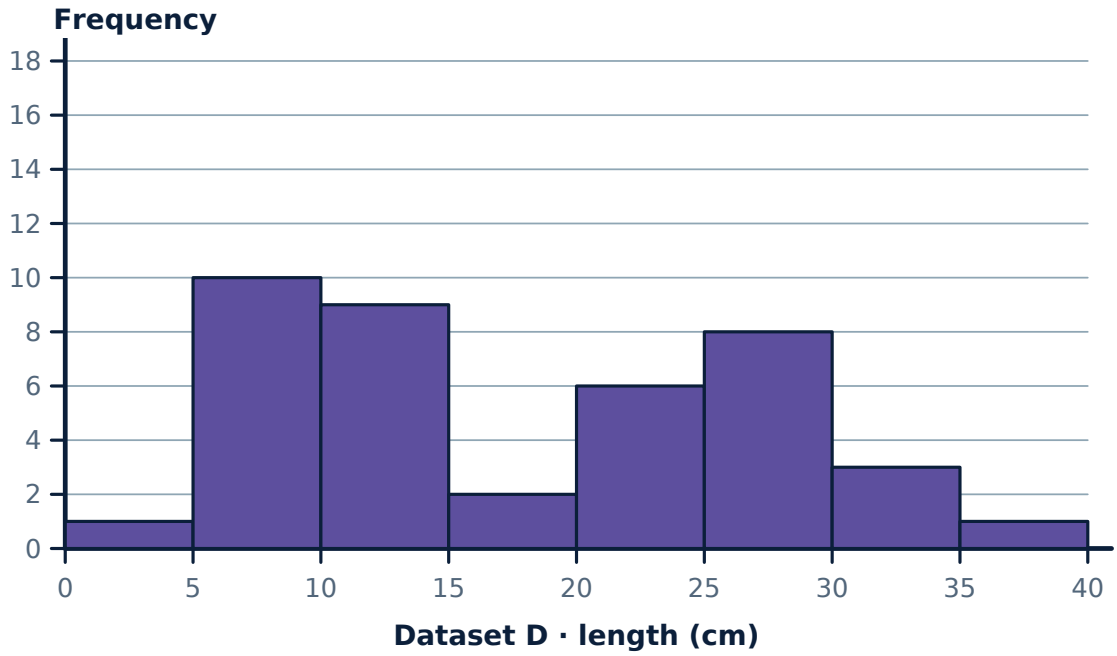


Same class width, same axis. Only the shape differs.

Two histograms on identical axes. Read the outline, not the bars: one of these rises and falls evenly, the other leans hard to one end and trails away.

Shape gallery 2 · datasets C and D





Name each shape before you read the worked example.

Two more. The upper one is the mirror image of dataset B; the lower one has something no single measurement process usually produces.

Worked example 3 · Name the shape

1	**Dataset A** — Dataset A, 40 lengths in centimetres: 0 ≤ x < 5: 1; 5 ≤ x < 10: 3; 10 ≤ x < 15: 6; 15 ≤ x < 20: 11; 20 ≤ x < 25: 9; 25 ≤ x < 30: 6; 30 ≤ x < 35: 3; 35 ≤ x < 40: 1.	**peaks: 1**
2	One peak, and the tails are 3 classes left and 4 right.	**symmetrical**
3	**Dataset B** — Dataset B, 40 lengths in centimetres: 0 ≤ x < 5: 6; 5 ≤ x < 10: 12; 10 ≤ x < 15: 8; 15 ≤ x < 20: 5; 20 ≤ x < 25: 4; 25 ≤ x < 30: 3; 30 ≤ x < 35: 1; 35 ≤ x < 40: 1.	**peaks: 1**
4	One peak low down, tails 1 classes left and 6 right.	**skewed to the right**
5	**Dataset C** — Dataset C, 40 lengths in centimetres: 0 ≤ x < 5: 1; 5 ≤ x < 10: 1; 10 ≤ x < 15: 2; 15 ≤ x < 20: 3; 20 ≤ x < 25: 4; 25 ≤ x < 30: 10; 30 ≤ x < 35: 16; 35 ≤ x < 40: 3.	**peaks: 1**
6	One peak high up, tails 6 classes left and 1 right.	**skewed to the left**
7	**Dataset D** — Dataset D, 40 lengths in centimetres: 0 ≤ x < 5: 1; 5 ≤ x < 10: 10; 10 ≤ x < 15: 9; 15 ≤ x < 20: 2; 20 ≤ x < 25: 6; 25 ≤ x < 30: 8; 30 ≤ x < 35: 3; 35 ≤ x < 40: 1.	**peaks: 2**
8	Two peaks, at 5 ≤ x < 10 and 25 ≤ x < 30, with a dip between.	**bimodal**

The name, then the story. *A is* ***symmetrical****: most lengths cluster near the middle and the two sides fall away together. B is* ***skewed to the right****: the peak sits at 5 ≤ x < 10 and the tail runs 6 classes to the right, so most of the lengths are small and a handful are much larger. C is* ***skewed to the left****, the mirror image, with 6 classes of tail on the low side. D is* ***bimodal****, and that one is a warning rather than a description: two humps in one histogram usually means two groups have been mixed, and the right response is to find the two groups and draw them separately. In every case the answer names a* ***class*** *— the modal class of B is 5 ≤ x < 10 — and never a single measurement, because these lengths are continuous and have no*

mode.

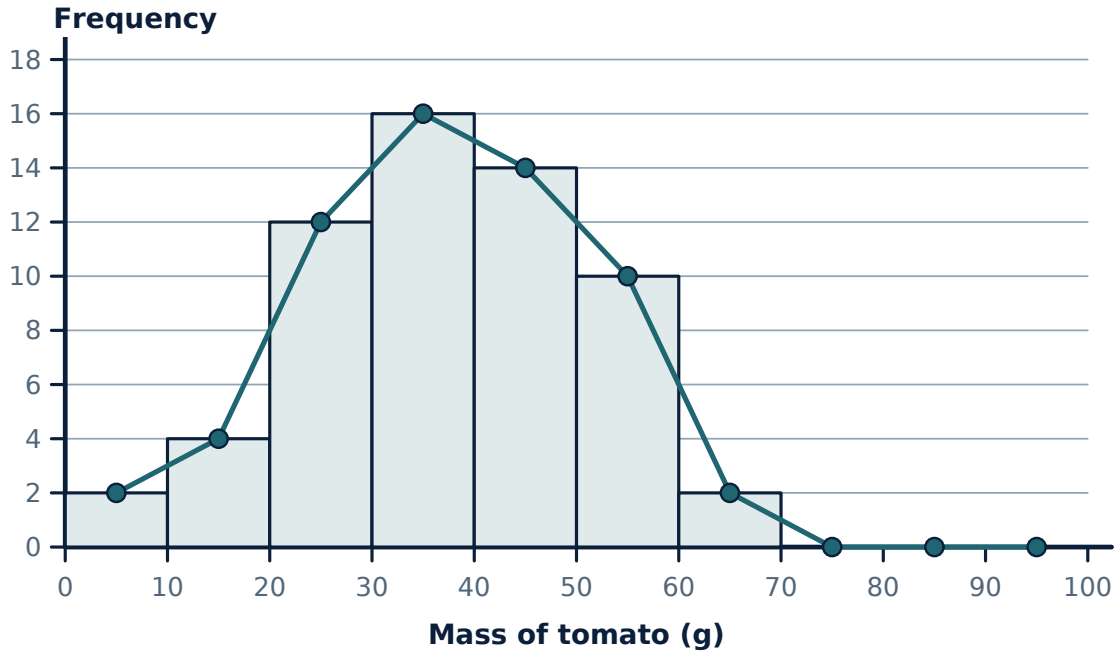
EXAMPLE. Dataset B has its tallest bar at $5 \le x < 10$ and then falls away class after class, with 6 classes of tail to the right of the peak and only 1 to the left. That is **skewed to the right**: most of the lengths are small and a few run far above them. Dataset D has two separate peaks, at $5 \le x < 10$ and at $25 \le x < 30$, with a dip between them, so it is **bimodal** — and the honest reading is not "the average is in the dip" but "there are probably two groups here".

CONCEPT 4 · JOINING THE MIDPOINTS

the frequency polygon, and two distributions on one axes

A **frequency polygon** is the same table drawn as a line instead of a block of bars. One point per class, placed at the **midpoint** of the class — S17's midpoint, (lower + upper) ÷ 2 — at a height equal to that class's frequency. Join the points in order with straight lines and stop at the first and last points you plotted. The midpoint is used because it is the one value in the class with nothing to favour, and because a line drawn through the midpoints of the bar tops keeps the same total area under it as the histogram does. What a polygon buys you is the whole point of this concept: two sets of bars laid over each other fight, but two lines do not. Draw both distributions as polygons on **one pair of axes** and the comparison — which one sits further right, which is more spread out, which has two peaks — is there in a single picture. One warning to carry into the next concept: a polygon must not be dragged down to the axis at each end unless there really are empty classes there, because a line reaching the axis claims a frequency of zero, and a claim about a class that does not exist is an invented measurement.

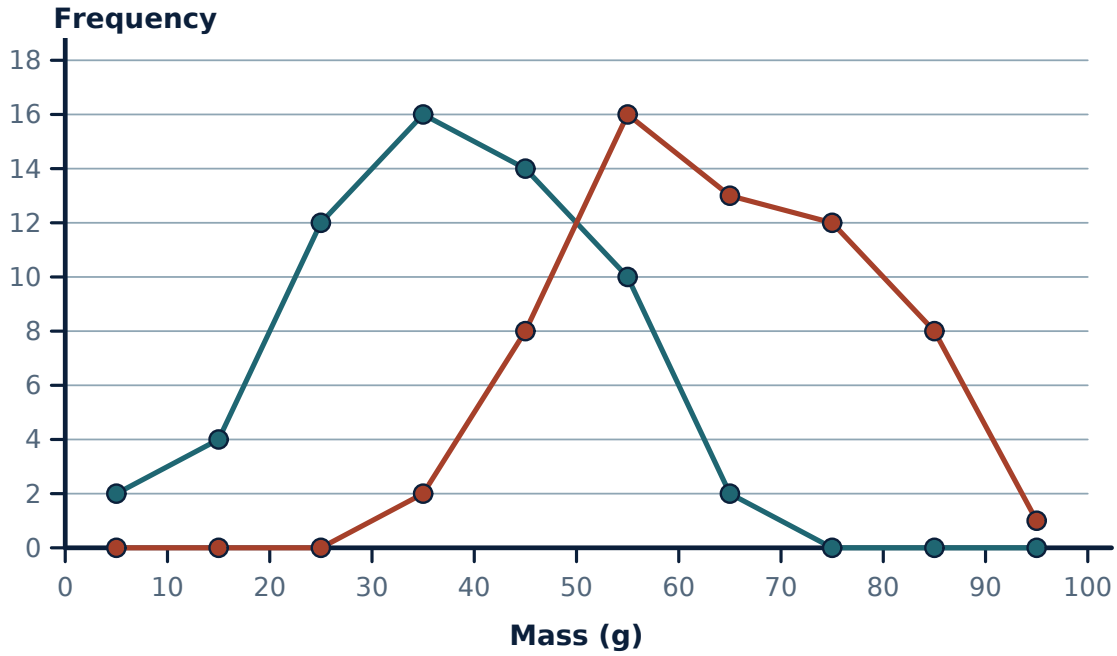
From bars to a polygon · the early planting



One point per class: at the midpoint, at the bar's height.

One point per class, at the class midpoint, at the height of the bar. Joined in order, and stopped at the last point the table gives.

Two plantings on one set of axes



Teal: early planting. Terracotta: late. Same classes, same axes.

Two distributions, one pair of axes, and the comparison readable at a glance. Two sets of bars could not have done this — the taller ones would have hidden the rest.

Worked example 4 · Two seasons, one pair of axes

1	Count both plantings into the SAME classes.	**10 classes of 10 g**
2	Early planting frequencies.	**2, 4, 12, 16, 14, 10, 2, 0, 0, 0**
3	Late planting frequencies.	**0, 0, 0, 2, 8, 16, 13, 12, 8, 1**
4	Write the midpoints once — both polygons use them.	**5, 15, 25, 35, 45, 55, 65, 75, 85, 95**
5	Plot the early points and join them.	**(5, 2), (15, 4), (25, 12), (35, 16), …**
6	Plot the late points on the same axes and join them.	**(5, 0), (15, 0), (25, 0), (35, 2), …**
7	Compare the centres.	**modal class $30 \le x < 40$ vs $50 \le x < 60$**
8	Compare the spreads — from the classes each table uses.	**70 g of scale vs 70 g**
9	Compare the shapes.	**symmetrical vs symmetrical**

Three sentences, each defended from the picture. *Centre: the late planting is heavier — its polygon sits to the right of the early one all the way along, and its modal class is $50 \le x < 60$ against $30 \le x < 40$. Spread: the two are equally spread, 70 g of the scale for the late planting and 70 g for the early one, so the late crop is not more variable, just heavier. Shape: both are single symmetrical humps, so the whole difference is a shift and nothing else. Two histograms drawn on top of each other could not have shown this — the bars would have hidden one another — and that is what the polygon is for. Note also where each line stops: at the first and last classes the table has, and not dragged down to the axis beyond them.*

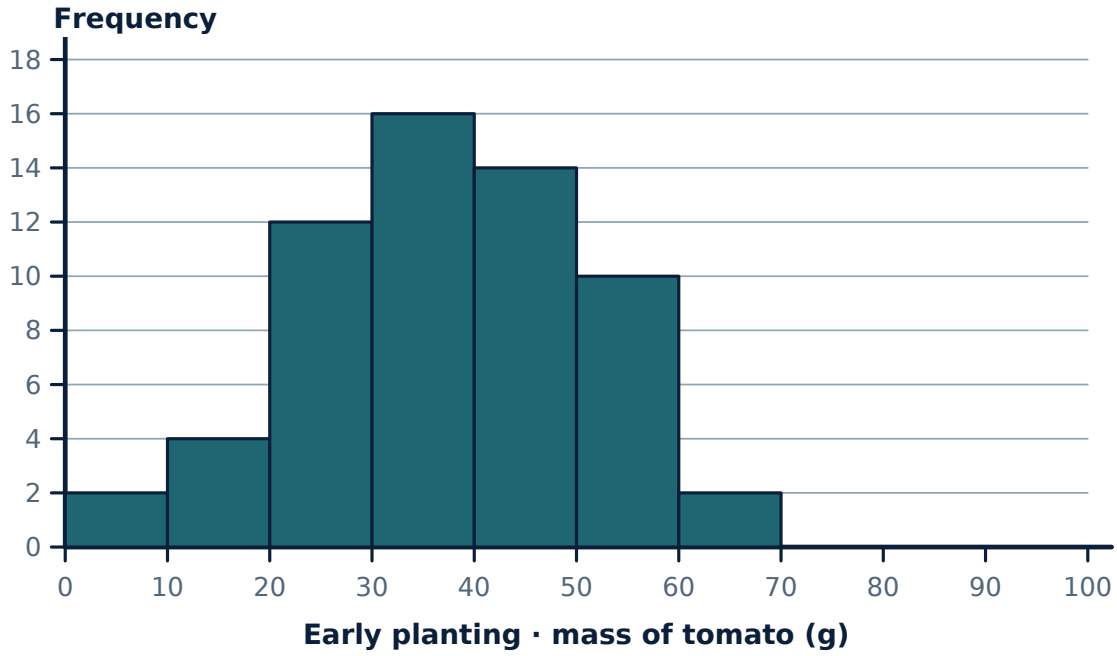
EXAMPLE. The early planting has 16 tomatoes in the class $30 \le x < 40$, so the polygon carries a point at (35, 16) — the midpoint, 35 g, at a height of 16. The full set of points is (5, 2), (15, 4), (25, 12), (35, 16), (45, 14), (55, 10), (65, 2), (75, 0), (85, 0), (95, 0). The line joins them in order and stops at the last one; it is not dragged down to the axis beyond the classes the table actually has.

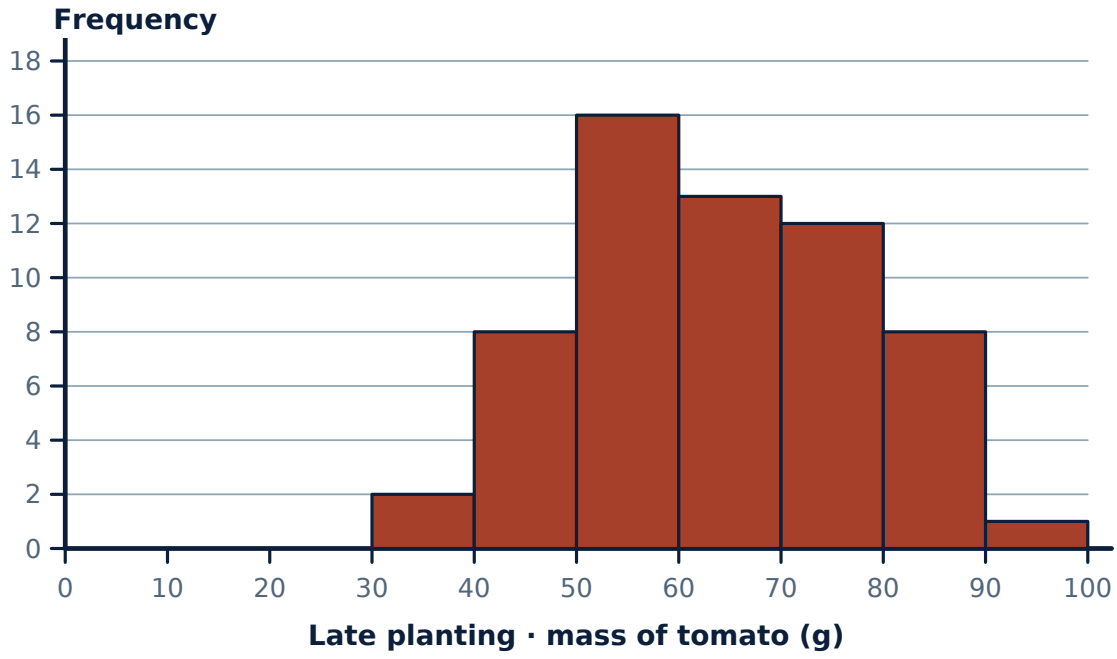
CONCEPT 5 · COMPARING HONESTLY

same classes, same axes, or no comparison at all

Two histograms may be compared only when they are drawn on the same classes and the same axes. That is not a style rule, it is the whole basis of the comparison: if one chart uses classes 5 wide and the other uses classes 10 wide, a bar on one is not the same kind of object as a bar on the other, and the taller bar means nothing. Frequency axes that stop at different heights do the same damage more quietly — both shapes are drawn correctly, and the eye still reads the one on the shorter axis as bigger. So the honest procedure is fixed: choose one set of classes that suits both datasets, count both into it, and draw both against one frequency scale, either as a pair of histograms or as two polygons on one axes. Then, and only then, write the comparison — and write it in three parts, **centre**, **spread** and **shape**, each backed by something you can point at on the chart. Where the two datasets are different sizes, say so, because a taller bar may just mean more measurements.

The honest comparison · one pair of axes for both





Identical classes, identical axis. Only then may these be compared.

The same two plantings as histograms. Identical classes and an identical frequency axis are what make the comparison legitimate; without them there is nothing to compare.

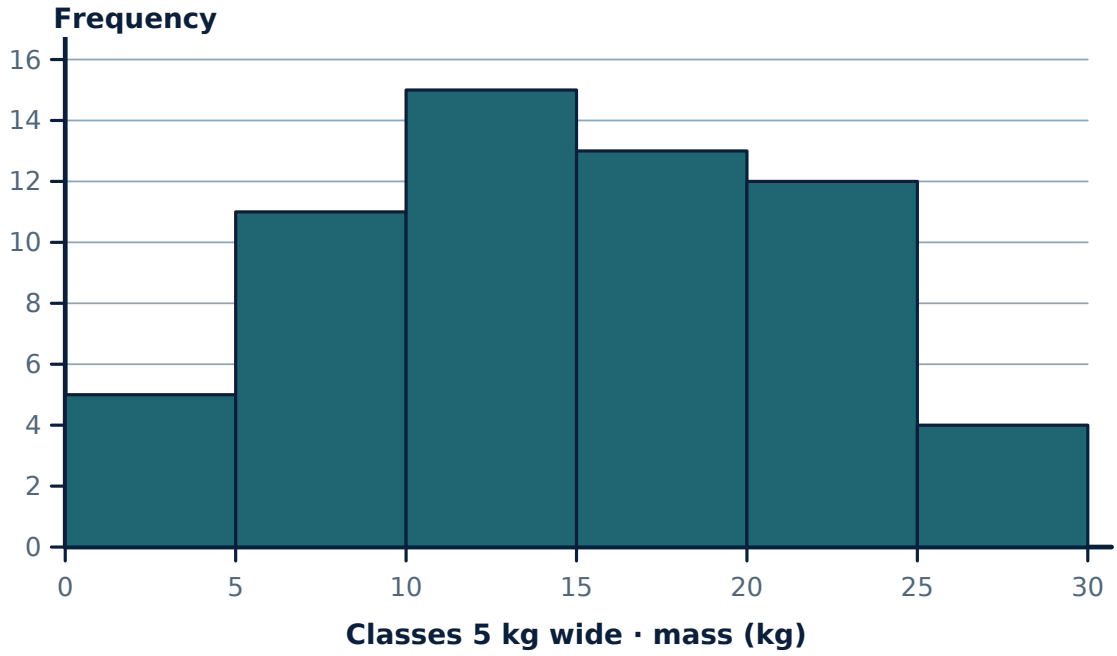
EXAMPLE. Both plantings were counted into the same 10 classes, each 10 g wide, and drawn against the same frequency axis. Now the comparison can be written and defended: the late planting is **heavier** — its modal class is $50 \le x < 60$ against $30 \le x < 40$ — while the two are equally **spread**, 70 g of the scale for the late planting and 70 g for the early one, and both have the same **shape**, a single symmetrical hump.

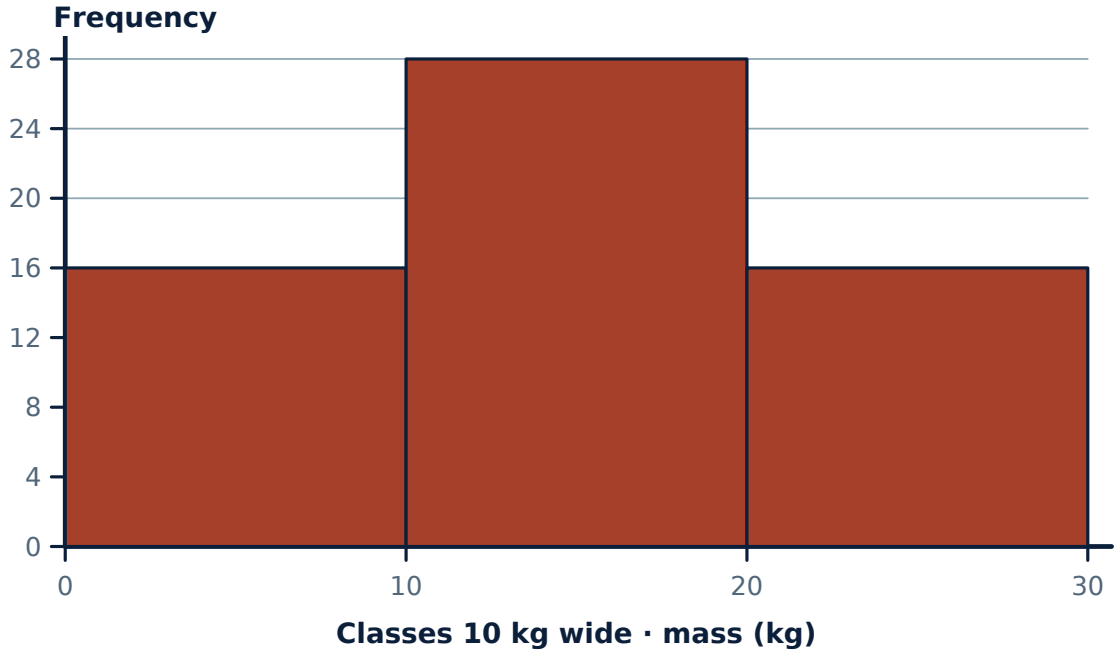
CONCEPT 6 · CHANGE THE WIDTH, CHANGE THE PICTURE

what survives a regrouping and what blurs away

You chose the class width, and the histogram you get is partly your decision. Group the 60 pumpkin masses in 5s and you get 6 bars with a clear single hump; group the very same masses in 10s and you get 3 bars, and the detail is gone — too few bars to show a shape. Go the other way and the opposite failure appears: group them in 2s and each of the 15 classes holds so few pumpkins that the tops jump up and down 6 times from pure chance, and the chart is now showing which classes happened to catch an extra pumpkin. Neither picture is wrong arithmetically; both are misleading to read. Somewhere between them is a width that shows the shape without showing the chance, and finding it is a judgement you have to make and then defend. Two habits follow. First, whatever width you choose, say what it is and why you chose it. Second, when a feature appears at only one width, treat it with suspicion until you have looked at the data another way — the Criterion B task below is exactly that investigation.

Same data, two class widths





The same 60 masses. 6 bars above, 3 below.

Nothing in the data changed between these two pictures. The class width did, and it took most of the shape with it.

The same 60 masses again, at width 2



15 classes, 6 peaks, and not one of them means anything.

And the failure at the other end. Sixty measurements cut into fifteen classes leave about four in each, and four is far too few for a bar height to be stable.

Worked example 5 · Same data, two widths

1	Start from the 60 raw masses, not from a table.	**60 values**
2	Count them into classes 5 kg wide.	**5, 11, 15, 13, 12, 4**
3	Read that histogram.	**6 bars, symmetrical**
4	State its modal class.	**10 ≤ x < 15**
5	Now count the same masses into classes 10 kg wide.	**16, 28, 16**
6	Read that histogram.	**3 bars, too few bars to show a shape**
7	State its modal class.	**10 ≤ x < 20**
8	Say what survived and what did not.	**centre survived, detail blurred**

What survives and what blurs. *The centre survives: both pictures put the bulk of the pumpkins in the middle of the scale, and the wider modal class $10 \le x < 20$ simply contains the narrower one $10 \le x < 15$. What blurs is everything the shape is made of. At 5 kg wide there are 6 bars, the outline rises and falls and the distribution reads as symmetrical; at 10 kg wide there are 3 bars, and with 3 of them a histogram cannot show a shape at all — the fall-away at the top of the scale, clearly visible in the 5-kg picture as the drop from 12 to 4, has been absorbed into one wide block. Neither chart is wrong. The 5-kg chart is more useful, and an answer that shows one of them has to say which width it used.*

EXAMPLE. The same 60 masses, three times over. At width 5 there are 6 bars, the shape reads as symmetrical and the modal class is $10 \le x < 15$. At width 10 there are only 3 bars: too few bars to show a shape, and the fall-away at the top of the scale — the drop from 12 to 4 in the wider picture — has been absorbed into one block. At width 2 there are 15 bars and 6 peaks, which is chance dressed up as detail. Same 60 pumpkins every time.

Beyond this course · one sentence only. If the classes are *not* all the same width, bar heights stop being comparable and the chart has to plot **frequency density** — frequency divided by class width — so that it is the *area* of each bar, not its height, that gives the frequency. That tool is named here so you know it exists and know its name; it is not part of this course, there is no practice on it in this unit, and every histogram you meet here has equal fixed intervals.

THE HISTOGRAM ROUTINE

1. Ask what kind of data it is, first. Measured on a scale means a histogram and touching bars. Categories or counted whole numbers mean a bar chart and gaps. Nothing else in the routine works if this step is wrong.

2. Fix equal classes, and label both axes. Equal fixed intervals that meet, boundaries marked along the measured scale, frequency up the side starting at 0, and a title. A bare axis is not a chart.

3. Read the outline, then name the shape. Count the peaks first: three or more is noise, two is bimodal, one sends you to the tails. Longer tail on the right means skewed to the right. Then say the modal class - a class, never a value.

4. Say the width you chose, and check the shape survives it. Regroup once at a different width. A feature that only exists at one width is not a feature of the data, and a comparison drawn on two different widths is not a comparison at all.

THE MISTAKES THAT COST THE MOST MARKS

Trap 1 · Leaving gaps between the bars of continuous data

Wrong. Drawing the 60 pumpkin masses as 6 separate bars with gaps between them, labelled “0–5”, “5–10”, “10–15” and so on, exactly as if the classes were subjects on a timetable.

Right. Draw the mass itself along the bottom axis, marked at the boundaries 0, 5, 10, 15, 20, 25, 30, and let each bar run from its lower boundary to its upper one, so that the bars touch because the classes touch.

The gap is not a decoration, it is a statement, and here the statement is false. A gap between $5 \le x < 10$ and $10 \le x < 15$ claims that no pumpkin weighs 10 kg — when the very table being drawn puts 15 pumpkins in the class that starts at 10 kg. The same slip has a twin worth naming: writing class *labels* such as “5–10” along the axis instead of the boundary numbers. Labels turn a measured scale into a row of categories, and once that has happened the gaps look natural and the chart has quietly become a bar chart.

Trap 2 · Reading the tallest bar of a skewed histogram as “the average”

Wrong. “Dataset B is skewed to the right, so the average length is about 7.5 cm.”

Right. “The modal class of dataset B is $5 \le x < 10$. Because the distribution is skewed to the right, the mean is pulled above that class by the 6 classes of tail on the right, and the table puts the mean of the 40 lengths at 13.4 cm.”

The tallest bar answers the question “where do most of the measurements sit?” and nothing else. In a symmetrical distribution that is close enough to the middle that the slip does no harm, which is exactly why it goes unnoticed — and then costs marks the moment the shape is skewed. Here the modal class runs from 5 to 10 cm while the mean the table gives is 13.4 cm, well outside it, because a long right tail drags the mean up and leaves the tallest bar behind. Say modal **class**, say it is a class and not an average, and if the question wants a typical value, name which average you are giving.

Trap 3 · Comparisons the two charts cannot support

Wrong. Putting the 5-kg-wide picture of the pumpkins beside the 10-kg-wide one and concluding “the second field has far more heavy pumpkins, look at the height of that bar” — and, on the polygon, running the line down to the axis at each end so that it closes neatly on the frequency scale.

Right. Count both datasets into **one** set of classes and draw them against **one** frequency axis before comparing anything — as the two plantings are, in 10 identical classes of 10 g. Stop each polygon at the first and last points the table actually gives: (5, 2) and (95, 0) for the early planting.

Both halves of this trap are the same mistake — the chart claiming something the data never said. A bar 10 kg wide holds about twice as many pumpkins as a bar 5 kg wide whatever the field is like, so comparing their heights measures the class width and not the harvest; the same 60 masses give a tallest bar of 28 at one width and 15 at the other. And a polygon dragged down to the axis beyond the last class invents a class, gives it a frequency of zero and joins it up, which is a measurement nobody made. If the end classes really are empty — as $70 \le x < 80$ and $80 \le x < 90$ honestly are for the early planting — then the line reaching the axis is true, and that is the only case in which it may.

PRACTICE · 60 QUESTIONS

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

Concept 1 · Bars that touch

P1 (Fluency)	**State** the one word that describes the kind of data a histogram is drawn for.
P2 (Fluency)	The 60 pumpkin masses, in kilograms, grouped in fives: $0 \le x < 5$: 5; $5 \le x < 10$: 11; $10 \le x < 15$: 15; $15 \le x < 20$: 13; $20 \le x < 25$: 12; $25 \le x < 30$: 4. **State** the class width of this table.
P3 (Fluency)	The 60 pumpkin masses, in kilograms, grouped in fives: $0 \le x < 5$: 5; $5 \le x < 10$: 11; $10 \le x < 15$: 15; $15 \le x < 20$: 13; $20 \le x < 25$: 12; $25 \le x < 30$: 4. **State** the height of the bar drawn over the class $15 \le x < 20$.
P4 (Fluency)	The 60 pumpkin masses, in kilograms, grouped in fives: $0 \le x < 5$: 5; $5 \le x < 10$: 11; $10 \le x < 15$: 15; $15 \le x < 20$: 13; $20 \le x < 25$: 12; $25 \le x < 30$: 4. **State** how many classes this histogram has.
P5 (Applied)	**Explain**, in one sentence, why the bars of a histogram of continuous data touch.
P6 (Applied)	The 60 pumpkin masses, in kilograms, grouped in fives: $0 \le x < 5$: 5; $5 \le x < 10$: 11; $10 \le x < 15$: 15; $15 \le x < 20$: 13; $20 \le x < 25$: 12; $25 \le x < 30$: 4. **Calculate** how many pumpkins weighed less than 10 kg.
P7 (Applied)	The 60 pumpkin masses, in kilograms, grouped in fives: $0 \le x < 5$: 5; $5 \le x < 10$: 11; $10 \le x < 15$: 15; $15 \le x < 20$: 13; $20 \le x < 25$: 12; $25 \le x < 30$: 4. **State** the numbers that should be written along the horizontal axis of the histogram.
P8 (Applied)	A learner draws the pumpkin histogram with the frequency axis starting at 4 instead of 0. **Explain** what goes wrong.

P9 (Challenge)	The 60 pumpkin masses, in kilograms, grouped in fives: $0 \le x < 5$: 5; $5 \le x < 10$: 11; $10 \le x < 15$: 15; $15 \le x < 20$: 13; $20 \le x < 25$: 12; $25 \le x < 30$: 4. A second table groups the same masses in classes 10 kg wide. **Justify** whether the two histograms may be drawn on the same frequency axis and compared bar by bar.
P10 (Challenge)	The 60 pumpkin masses, in kilograms, grouped in fives: $0 \le x < 5$: 5; $5 \le x < 10$: 11; $10 \le x < 15$: 15; $15 \le x < 20$: 13; $20 \le x < 25$: 12; $25 \le x < 30$: 4. **Determine** the smallest and the largest possible number of pumpkins weighing more than 22 kg that this table is consistent with.

Concept 2 · Bar chart or histogram?

P11 (Fluency)	**Identify** the chart that shows the favourite subject of 30 learners: Maths: 9; Setswana: 7; Science: 6; Art: 5; History: 3.
P12 (Fluency)	**Identify** the chart that shows the mass of 60 pumpkins measured to the nearest 0.1 kg.
P13 (Fluency)	**State** what the gaps in a bar chart tell the reader.
P14 (Fluency)	**State** the one question to ask of a variable before choosing between a bar chart and a histogram.
P15 (Applied)	The number of children in 30 households is 0: 4; 1: 7; 2: 9; 3: 6; 4: 3; 5: 1. **Justify** the choice of chart.
P16 (Applied)	The 60 early-planting tomato masses, in grams, grouped in tens: $0 \le x < 10$: 2; $10 \le x < 20$: 4; $20 \le x < 30$: 12; $30 \le x < 40$: 16; $40 \le x < 50$: 14; $50 \le x < 60$: 10; $60 \le x < 70$: 2; $70 \le x < 80$: 0; $80 \le x < 90$: 0; $90 \le x < 100$: 0. **Justify** the choice of chart for these masses.
P17 (Applied)	**Explain** why a bar chart of favourite subjects may have its bars rearranged, while a histogram may not.

P18 (Applied)	A learner says shoe size should be drawn as a histogram because it is a measurement. **Explain** whether they are right.
P19 (Challenge)	A survey records the exact time, in minutes, that each of 40 learners spent on homework, and also which of four subjects they worked on. **Justify** which chart each of the two variables needs.
P20 (Challenge)	**Justify** why a bar chart of 6 categories and a histogram of 6 classes can look almost identical on the page and still be different claims.

Concept 3 · Reading the shape

P21 (Fluency)	Dataset A, 40 lengths in centimetres: $0 \le x < 5$: 1; $5 \le x < 10$: 3; $10 \le x < 15$: 6; $15 \le x < 20$: 11; $20 \le x < 25$: 9; $25 \le x < 30$: 6; $30 \le x < 35$: 3; $35 \le x < 40$: 1. **State** the shape of this distribution.
P22 (Fluency)	Dataset B, 40 lengths in centimetres: $0 \le x < 5$: 6; $5 \le x < 10$: 12; $10 \le x < 15$: 8; $15 \le x < 20$: 5; $20 \le x < 25$: 4; $25 \le x < 30$: 3; $30 \le x < 35$: 1; $35 \le x < 40$: 1. **State** the shape of this distribution.
P23 (Fluency)	Dataset C, 40 lengths in centimetres: $0 \le x < 5$: 1; $5 \le x < 10$: 1; $10 \le x < 15$: 2; $15 \le x < 20$: 3; $20 \le x < 25$: 4; $25 \le x < 30$: 10; $30 \le x < 35$: 16; $35 \le x < 40$: 3. **State** the modal class.
P24 (Fluency)	**State** the word for a distribution whose histogram has two separate peaks.
P25 (Applied)	Dataset C, 40 lengths in centimetres: $0 \le x < 5$: 1; $5 \le x < 10$: 1; $10 \le x < 15$: 2; $15 \le x < 20$: 3; $20 \le x < 25$: 4; $25 \le x < 30$: 10; $30 \le x < 35$: 16; $35 \le x < 40$: 3. **Justify** the shape you would give this distribution.

P26 (Applied)	Dataset D, 40 lengths in centimetres: $0 \le x < 5$: 1; $5 \le x < 10$: 10; $10 \le x < 15$: 9; $15 \le x < 20$: 2; $20 \le x < 25$: 6; $25 \le x < 30$: 8; $30 \le x < 35$: 3; $35 \le x < 40$: 1. **Explain** what this shape suggests about where the 40 measurements came from.
P27 (Applied)	Dataset A, 40 lengths in centimetres: $0 \le x < 5$: 1; $5 \le x < 10$: 3; $10 \le x < 15$: 6; $15 \le x < 20$: 11; $20 \le x < 25$: 9; $25 \le x < 30$: 6; $30 \le x < 35$: 3; $35 \le x < 40$: 1. **Explain** why the answer to “what is the mode?” is a class and not a number.
P28 (Applied)	Dataset B, 40 lengths in centimetres: $0 \le x < 5$: 6; $5 \le x < 10$: 12; $10 \le x < 15$: 8; $15 \le x < 20$: 5; $20 \le x < 25$: 4; $25 \le x < 30$: 3; $30 \le x < 35$: 1; $35 \le x < 40$: 1. **Calculate** how many of the 40 lengths were 20 cm or more.
P29 (Challenge)	Dataset B, 40 lengths in centimetres: $0 \le x < 5$: 6; $5 \le x < 10$: 12; $10 \le x < 15$: 8; $15 \le x < 20$: 5; $20 \le x < 25$: 4; $25 \le x < 30$: 3; $30 \le x < 35$: 1; $35 \le x < 40$: 1. Taking each class at its midpoint, as in S17, this table estimates the mean of the 40 lengths at 13.4 cm. **Justify** why that mean lies outside the modal class.
P30 (Challenge)	Two histograms of 40 measurements each have the same modal class and the same range. **Justify** whether they must have the same shape.

Concept 4 · Joining the midpoints

P31 (Fluency)	**State** the value in each class that a frequency polygon is plotted at.
P32 (Fluency)	**Calculate** the midpoint of the class $40 \le x < 50$.

P33 (Fluency)	The 60 early-planting tomato masses, in grams, grouped in tens: $0 \le x < 10$: 2; $10 \le x < 20$: 4; $20 \le x < 30$: 12; $30 \le x < 40$: 16; $40 \le x < 50$: 14; $50 \le x < 60$: 10; $60 \le x < 70$: 2; $70 \le x < 80$: 0; $80 \le x < 90$: 0; $90 \le x < 100$: 0. **State** the coordinates of the polygon point for the class $20 \le x < 30$.
P34 (Fluency)	**State** how many points a frequency polygon for a table of 10 classes has.
P35 (Applied)	The 60 early-planting tomato masses, in grams, grouped in tens: $0 \le x < 10$: 2; $10 \le x < 20$: 4; $20 \le x < 30$: 12; $30 \le x < 40$: 16; $40 \le x < 50$: 14; $50 \le x < 60$: 10; $60 \le x < 70$: 2; $70 \le x < 80$: 0; $80 \le x < 90$: 0; $90 \le x < 100$: 0. **Calculate** the coordinates of all 10 points of the frequency polygon.
P36 (Applied)	**Explain** why a frequency polygon is easier than two histograms when two distributions are being compared.
P37 (Applied)	The 60 early-planting tomato masses, in grams, grouped in tens: $0 \le x < 10$: 2; $10 \le x < 20$: 4; $20 \le x < 30$: 12; $30 \le x < 40$: 16; $40 \le x < 50$: 14; $50 \le x < 60$: 10; $60 \le x < 70$: 2; $70 \le x < 80$: 0; $80 \le x < 90$: 0; $90 \le x < 100$: 0. **Explain** why the polygon for the early planting is not drawn down to the axis to the right of the point (95, 0).
P38 (Applied)	The 60 late-planting tomato masses, in grams, grouped in the same classes: $0 \le x < 10$: 0; $10 \le x < 20$: 0; $20 \le x < 30$: 0; $30 \le x < 40$: 2; $40 \le x < 50$: 8; $50 \le x < 60$: 16; $60 \le x < 70$: 13; $70 \le x < 80$: 12; $80 \le x < 90$: 8; $90 \le x < 100$: 1. **Calculate** the total frequency represented by the polygon and **state** what it must equal.

P39 (Challenge)	The 60 early-planting tomato masses, in grams, grouped in tens: 0 ≤ x < 10: 2; 10 ≤ x < 20: 4; 20 ≤ x < 30: 12; 30 ≤ x < 40: 16; 40 ≤ x < 50: 14; 50 ≤ x < 60: 10; 60 ≤ x < 70: 2; 70 ≤ x < 80: 0; 80 ≤ x < 90: 0; 90 ≤ x < 100: 0. The 60 late-planting tomato masses, in grams, grouped in the same classes: 0 ≤ x < 10: 0; 10 ≤ x < 20: 0; 20 ≤ x < 30: 0; 30 ≤ x < 40: 2; 40 ≤ x < 50: 8; 50 ≤ x < 60: 16; 60 ≤ x < 70: 13; 70 ≤ x < 80: 12; 80 ≤ x < 90: 8; 90 ≤ x < 100: 1. **Justify** a comparison of the two plantings in terms of centre, spread and shape.
P40 (Challenge)	A frequency polygon is drawn from a table of 6 classes of width 5. **Determine** what happens to the polygon if every class is split in two and the data regrouped, and **justify** your answer.

Concept 5 · Comparing honestly

P41 (Fluency)	**State** the two things that must be identical before two histograms may be compared.
P42 (Fluency)	The 60 early-planting tomato masses, in grams, grouped in tens: 0 ≤ x < 10: 2; 10 ≤ x < 20: 4; 20 ≤ x < 30: 12; 30 ≤ x < 40: 16; 40 ≤ x < 50: 14; 50 ≤ x < 60: 10; 60 ≤ x < 70: 2; 70 ≤ x < 80: 0; 80 ≤ x < 90: 0; 90 ≤ x < 100: 0. The 60 late-planting tomato masses, in grams, grouped in the same classes: 0 ≤ x < 10: 0; 10 ≤ x < 20: 0; 20 ≤ x < 30: 0; 30 ≤ x < 40: 2; 40 ≤ x < 50: 8; 50 ≤ x < 60: 16; 60 ≤ x < 70: 13; 70 ≤ x < 80: 12; 80 ≤ x < 90: 8; 90 ≤ x < 100: 1. **State** the modal class of the late planting.
P43 (Fluency)	**State** the three headings a written comparison of two distributions should cover.

P44 (Fluency)	The 60 early-planting tomato masses, in grams, grouped in tens: $0 \le x < 10$: 2; $10 \le x < 20$: 4; $20 \le x < 30$: 12; $30 \le x < 40$: 16; $40 \le x < 50$: 14; $50 \le x < 60$: 10; $60 \le x < 70$: 2; $70 \le x < 80$: 0; $80 \le x < 90$: 0; $90 \le x < 100$: 0. **Calculate** how many of the early-planting tomatoes weighed less than 30 g.
P45 (Applied)	The 60 early-planting tomato masses, in grams, grouped in tens: $0 \le x < 10$: 2; $10 \le x < 20$: 4; $20 \le x < 30$: 12; $30 \le x < 40$: 16; $40 \le x < 50$: 14; $50 \le x < 60$: 10; $60 \le x < 70$: 2; $70 \le x < 80$: 0; $80 \le x < 90$: 0; $90 \le x < 100$: 0. The 60 late-planting tomato masses, in grams, grouped in the same classes: $0 \le x < 10$: 0; $10 \le x < 20$: 0; $20 \le x < 30$: 0; $30 \le x < 40$: 2; $40 \le x < 50$: 8; $50 \le x < 60$: 16; $60 \le x < 70$: 13; $70 \le x < 80$: 12; $80 \le x < 90$: 8; $90 \le x < 100$: 1. **Compare** the centres of the two plantings, quoting evidence from the tables.
P46 (Applied)	A magazine prints the early planting as 10 bars of width 10 g beside the late planting as 5 bars of width 20 g. **Explain** why the two pictures cannot be compared.
P47 (Applied)	Two histograms use identical classes, but one frequency axis runs to 20 and the other to 40. **Explain** what a reader is likely to get wrong.

P48 (Applied)	The 60 early-planting tomato masses, in grams, grouped in tens: 0 ≤ x < 10: 2; 10 ≤ x < 20: 4; 20 ≤ x < 30: 12; 30 ≤ x < 40: 16; 40 ≤ x < 50: 14; 50 ≤ x < 60: 10; 60 ≤ x < 70: 2; 70 ≤ x < 80: 0; 80 ≤ x < 90: 0; 90 ≤ x < 100: 0. The 60 late-planting tomato masses, in grams, grouped in the same classes: 0 ≤ x < 10: 0; 10 ≤ x < 20: 0; 20 ≤ x < 30: 0; 30 ≤ x < 40: 2; 40 ≤ x < 50: 8; 50 ≤ x < 60: 16; 60 ≤ x < 70: 13; 70 ≤ x < 80: 12; 80 ≤ x < 90: 8; 90 ≤ x < 100: 1. Each table runs from the bottom of its lowest occupied class to the top of its highest, so it can support a range of 70 g for the early planting and 70 g for the late one. **Compare** the spreads of the two plantings.
P49 (Challenge)	The 60 early-planting tomato masses, in grams, grouped in tens: 0 ≤ x < 10: 2; 10 ≤ x < 20: 4; 20 ≤ x < 30: 12; 30 ≤ x < 40: 16; 40 ≤ x < 50: 14; 50 ≤ x < 60: 10; 60 ≤ x < 70: 2; 70 ≤ x < 80: 0; 80 ≤ x < 90: 0; 90 ≤ x < 100: 0. The 60 late-planting tomato masses, in grams, grouped in the same classes: 0 ≤ x < 10: 0; 10 ≤ x < 20: 0; 20 ≤ x < 30: 0; 30 ≤ x < 40: 2; 40 ≤ x < 50: 8; 50 ≤ x < 60: 16; 60 ≤ x < 70: 13; 70 ≤ x < 80: 12; 80 ≤ x < 90: 8; 90 ≤ x < 100: 1. **Justify** a full comparison of the two plantings under the headings centre, spread and shape.
P50 (Challenge)	One dataset has 40 measurements and another has 120, and they are drawn as two histograms on identical classes and one frequency axis. **Justify** what may and may not be concluded from the fact that every bar of the second is taller.

Concept 6 · Change the width, change the picture

P51 (Fluency)	The 60 pumpkin masses, in kilograms, grouped in fives: 0 ≤ x < 5: 5; 5 ≤ x < 10: 11; 10 ≤ x < 15: 15; 15 ≤ x < 20: 13; 20 ≤ x < 25: 12; 25 ≤ x < 30: 4. **State** the modal class.

P52 (Fluency)	The same 60 masses grouped in tens: $0 \le x < 10$: 16; $10 \le x < 20$: 28; $20 \le x < 30$: 16. **State** the modal class now.
P53 (Fluency)	The same 60 masses grouped in tens: $0 \le x < 10$: 16; $10 \le x < 20$: 28; $20 \le x < 30$: 16. **State** how many classes this table has.
P54 (Fluency)	**State** what happens to a histogram when the classes are made very narrow.
P55 (Applied)	The 60 pumpkin masses, in kilograms, grouped in fives: $0 \le x < 5$: 5; $5 \le x < 10$: 11; $10 \le x < 15$: 15; $15 \le x < 20$: 13; $20 \le x < 25$: 12; $25 \le x < 30$: 4. The same 60 masses grouped in tens: $0 \le x < 10$: 16; $10 \le x < 20$: 28; $20 \le x < 30$: 16. **Explain** how the same 60 masses give both of these tables.
P56 (Applied)	The same 60 masses grouped in tens: $0 \le x < 10$: 16; $10 \le x < 20$: 28; $20 \le x < 30$: 16. **Explain** why no shape can honestly be named from this table.
P57 (Applied)	The same 60 masses grouped in classes 2 kg wide give 15 classes with 6 peaks. **Explain** why those peaks should not be reported.
P58 (Applied)	The 60 pumpkin masses, in kilograms, grouped in fives: $0 \le x < 5$: 5; $5 \le x < 10$: 11; $10 \le x < 15$: 15; $15 \le x < 20$: 13; $20 \le x < 25$: 12; $25 \le x < 30$: 4. **Calculate** how many pumpkins weighed 20 kg or more, and **explain** why the width-10 table cannot answer the same question.
P59 (Challenge)	The 60 pumpkin masses, in kilograms, grouped in fives: $0 \le x < 5$: 5; $5 \le x < 10$: 11; $10 \le x < 15$: 15; $15 \le x < 20$: 13; $20 \le x < 25$: 12; $25 \le x < 30$: 4. The same 60 masses grouped in tens: $0 \le x < 10$: 16; $10 \le x < 20$: 28; $20 \le x < 30$: 16. **Justify** which of the two widths you would publish, including what each one costs.

P60 (Challenge) A learner regroups a dataset until a second peak appears and reports it as a discovery. **Justify** what they should do before publishing it.

CRITERION B · INVESTIGATING PATTERNS

Bin it again

Eighty sunflower seedlings were measured four weeks after planting, to the nearest tenth of a centimetre. Every one of the eighty heights is printed below, so nothing here has been decided for you: you group them yourself, as many ways as the task asks. The question is not what the average height is. It is this. The **shape** a histogram shows is supposed to be a property of the seedlings — but the class width is a property of your decision. How much of the shape you see is theirs, and how much is yours?

The 80 seedling heights, in centimetres, in full

17.6, 9.4, 20.3, 18.6, 10.5, 22.9, 25.0, 22.8, 20.8, 18.5, 34.2, 25.9, 20.7, 8.8, 27.6, 20.2, 31.0, 19.4, 13.1, 11.7, 20.5, 30.8, 10.4, 20.0, 16.0, 15.1, 27.3, 21.6, 17.8, 22.5, 27.1, 17.4, 3.4, 31.4, 27.9, 20.5, 22.5, 28.9, 5.6, 20.8, 28.7, 13.3, 28.4, 16.4, 22.5, 4.3, 22.9, 21.8, 35.6, 11.3, 27.9, 18.6, 8.7, 19.7, 12.8, 21.4, 26.1, 7.8, 32.9, 24.9, 16.7, 27.2, 16.8, 17.8, 31.7, 27.9, 25.0, 12.9, 27.2, 23.3, 12.0, 15.3, 18.0, 21.7, 29.1, 29.1, 13.8, 19.0, 35.9, 21.8.

The second dataset · 60 heights from a different bed, in full

31.8, 13.7, 11.5, 5.7, 31.8, 28.8, 14.5, 18.6, 32.5, 15.9, 10.8, 8.8, 8.9, 28.7, 9.9, 28.4, 9.5, 27.4, 15.0, 28.2, 11.1, 12.6, 32.2, 27.4, 30.0, 12.9, 11.5, 31.1, 7.0, 24.4, 16.6, 27.1, 9.1, 4.0, 11.7, 14.1, 13.9, 12.8, 13.3, 30.7, 23.6, 24.6, 9.6, 25.2, 36.2, 5.1, 10.4, 12.1, 30.6, 24.8, 7.8, 25.5, 13.4, 14.7, 28.4, 28.4, 28.3, 12.2, 30.5, 7.4.

Part 1 · the structured investigation

1. **Predict**, before you group anything, what you expect to happen to the shape of the histogram as the classes get narrower. Write the prediction down and do not change it afterwards.

2. **Tabulate** the eighty heights at class widths 20, 10, 5 and 2, all starting at 0. For each width record the number of classes, the frequencies, the **modal class**, the number of peaks in the outline, and a one-phrase **shape verdict**.

3. **Draw** the histogram at each of the four widths, on axes whose frequency scale starts at 0, so that the four shape verdicts can be seen

and not just asserted.

4. **Identify** the pattern in your table: say what improves as the classes narrow, at which width it stops improving, and what starts happening instead.

5. **Predict**, from that pattern and before you build it, what the table at width 1 would look like — how many classes, roughly what frequencies, and how many peaks. Then build it and say whether you were right.

6. **Determine** a general statement about choosing a class width, written as a sentence a classmate could apply to a dataset they have never seen. It must mention the number of measurements, because a rule that ignores *n* cannot be right.

7. **Test and verify** your general statement on the second dataset printed below — 60 heights from a different bed. Apply your rule to choose a width, group the data at that width and at one wider and one narrower, and say whether the width your rule chose is the one that shows the data best. Be honest if it is not.

8. **Justify** the trade-off from first principles: explain what a class width that is too wide destroys, explain what a class width that is too narrow manufactures, and explain why the second failure gets worse as the classes get narrower for a fixed *n*.

The table to fill in

Width	Classes	Frequencies	Modal class	Peaks	Shape verdict
20					
10					
5					
2					
1 - predict					

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

Part 2 · your own method choice

Part 2 · your own investigation. A general statement is only worth as much as the tests it survives. **Construct** your own rule for choosing a class width — it may be the one you determined above, sharpened, or a completely different one, and it may use n, the range, the smallest acceptable class frequency, or anything else you can defend. State it precisely enough that two people applying it to the same dataset would choose the same width. Then stress-test it. Find, or construct, a dataset on which your rule gives a width you would not defend, and say why it fails there. Finally, **justify** the rule's core claim by argument and not by example: show, in general, why a dataset of n measurements split into k classes puts on average n/k measurements in each class, and use that to argue why the number of classes cannot sensibly grow as fast as n does. You choose the rule, the tests and the way you present the argument — but a reader must be able to check every number and follow every step.

Criterion B · Investigating patterns. Levels 1–2: groups the eighty heights correctly at one or two widths and draws at least one correct histogram. **Levels 3–4:** all four tables complete and correct, each with its modal class and a shape verdict, the four histograms drawn on honest axes, and the pattern described in words. **Levels 5–6:** states a correct general statement about class width that refers to the number of measurements, predicts the width-1 table in advance, and tests the statement on the second dataset, reporting the result whichever way it falls. **Levels 7–8:** in addition, justifies the trade-off from first principles rather than from the two datasets — a wider class destroys information that no later step can recover, while a narrower class leaves about n/k measurements per class, so the chance wobble in each bar grows as k grows — and carries Part 2 through to a precisely stated rule, a dataset that breaks it, and a general argument rather than a worked example.

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

CRITERION D · APPLYING MATHEMATICS IN REAL-LIFE CONTEXTS

Was the second month really more changeable?

A school weather-watch group recorded the maximum temperature every day for two months. At the end of the second month one of them wrote in the report: *“The two months were about as hot as each other, but the second month was more variable.”* The group has the sheet below and has been asked to defend that sentence — or withdraw it — with a chart and a calculation, in a report the whole school will read.

Weather-watch record sheet

Entry	Recorded
Recorded by	the Form 4 weather-watch group, one reading a day
Instrument	a maximum-minimum thermometer in a screened box
Reading taken	17:00 each day, to the nearest 0.1 °C
Month 1	30 days, no readings missed
Month 2	31 days, no readings missed
Also recorded each day	minimum temperature, rainfall, wind direction, cloud cover
Rainfall, month 1	4 mm on 2 days
Rainfall, month 2	61 mm on 9 days
Thermometer checked against a second one	start of month 1 only

Month 1 · daily maximum temperature (°C), day 1 onwards

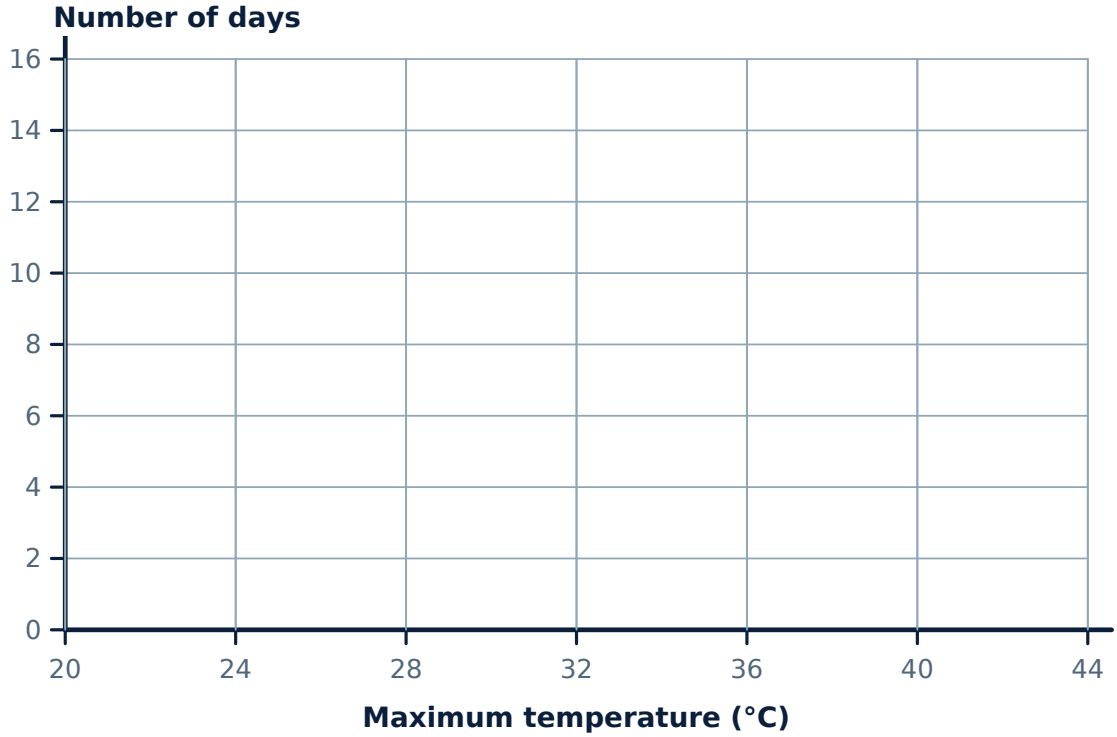
27.2, 32.9, 26.0, 30.6, 27.2, 29.0, 36.2, 31.4, 31.7, 33.7, 23.8, 30.2, 31.8, 33.0, 34.0, 31.3, 33.5, 34.1, 29.6, 31.8, 32.6, 31.0, 33.2, 29.1, 25.3, 34.2, 30.5, 27.1, 30.7, 29.8.

Month 2 · daily maximum temperature (°C), day 1 onwards

27.0, 29.2, 28.1, 36.1, 29.9, 26.0, 28.8, 34.4, 26.6, 23.8, 30.0, 26.3, 23.8, 31.6, 31.2, 33.4, 31.0, 41.8, 37.6, 26.9, 35.9, 22.6, 24.3, 30.9, 28.8, 27.0, 30.4, 35.1, 27.5, 33.9, 37.8.

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 · draw your histograms here



Blank axes for the Criterion D histograms. Choose your class interval first, use the same one for both months, and start the frequency axis at 0.

1. **Interpreting the situation. Justify** which entries on the record sheet the claim 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, and **justify** why one of them — the note about the thermometer check — still matters even though it is not used in any calculation.

2. **Selecting the mathematics. Justify** the choice of a histogram for each month and a frequency polygon overlay for the comparison, rather than two bar charts or a single list of numbers. **Justify** the class interval you will use, saying why it must be the same for both months and what a wider or a narrower interval would cost you.

3. **Applying the mathematics. Draw** a histogram of each month's maximum temperatures on the printed axes, using your chosen interval and a frequency axis starting at 0, with the bars touching. Then **draw** both frequency polygons on one pair of axes. **State** the modal class and the shape of each month, and **calculate** the range and the interquartile range of each month from the full lists of daily maxima.

4. **Accuracy. Justify** how far your answers can be trusted: say which of them come exactly from the raw lists and which depend on the interval you chose, **explain** what the readings being taken to the nearest 0.1 °C at one fixed time of day does and does not guarantee, and **discuss** what the unchecked thermometer means for a comparison of two months rather than for a single reading.

5. **Sense in context. Justify** a verdict on the sentence in the report. Does the evidence support “about as hot as each other”, and does it support “more variable”? Say which measure carries most of the weight of your verdict and why, **discuss** whether two months of readings from one thermometer at one school can support a sentence the whole school will read as a statement about the weather, and rewrite the sentence so that it claims exactly what the data supports and no more.

Criterion D · Applying mathematics in real-life contexts. Levels 1–2: identifies some relevant information and begins a grouped table for at least one month. **Levels 3–4:** separates the needed data from the surplus, groups both months into the same classes, and draws at least one correct histogram with touching bars and a frequency axis starting at 0. **Levels 5–6:** both histograms and both polygons correct on shared axes, modal classes and shapes stated, range and interquartile range correct for both months, the interval choice explained, and a defensible verdict on the report sentence. **Levels 7–8:** in addition, justifies the interval choice against what a wider and a narrower one would cost, separates the answers that are exact from those that depend on the grouping, reaches a verdict that treats “as hot as” and “more variable” as two separate claims with separate evidence, weighs the unchecked thermometer as a threat to a *comparison* rather than to a single reading, and rewrites the sentence so that it claims exactly what two months of one school's readings can support.

ANSWERS

Every answer here is recomputed from the datasets printed in the questions before this file is written. Shape words are the four this unit names, and a shape is never asserted without the peaks and the tails that decide it.

P1	continuous	**P31**	the midpoint of the class
P2	5 kg	**P32**	45
P3	13	**P33**	(25, 12)
P4	6	**P34**	10
P11	bar chart	**P41**	the classes and the frequency axis
P12	histogram	**P42**	$50 \le x < 60$
P13	that the categories are separate, so no value can fall between them	**P43**	centre, spread and shape
P14	can a value fall between two of the labels?	**P44**	18
P21	symmetrical	**P51**	$10 \le x < 15$
P22	skewed to the right	**P52**	$10 \le x < 20$
P23	$30 \le x < 35$	**P53**	3
P24	bimodal	**P54**	each class holds very few values, so the tops rise and fall from chance and the chart shows noise instead of shape

Applied and Challenge · fully worked

P5 because the top boundary of one class is the bottom boundary of the next, so the bars touch because the classes touch and a gap would claim that no measurement can land there

Every class ends exactly where the next one begins: $5 \le x < 10$ stops at 10 and $10 \le x < 15$ starts at 10. There is no stretch of the mass axis that belongs to neither class, so there is nowhere on the axis for a gap to sit. Put one there and you have made a claim about the data — that nothing was measured at 10 kg — which the table itself contradicts, since 15 pumpkins fall in the class beginning at 10 kg.

P6 16

Two bars lie below 10 kg, so add their heights: 5 + 11 = 16. Nothing has to be estimated here — the classes meet, so every pumpkin under 10 kg is in one of those two classes and none is counted twice. This is the kind of question a histogram answers exactly, and it is why the frequency axis has to start at 0: the heights are the counts themselves.

P7 0, 5, 10, 15, 20, 25, 30

The horizontal axis of a histogram carries the measured scale itself, so it is marked at the class **boundaries**: 0, 5, 10, 15, 20, 25, 30. It is not marked with class labels such as “5–10” sitting under the middle of each bar, because that turns a continuous scale into a row of names, and once the axis reads as names the gaps look natural and the chart has become a bar chart.

P8 the bar heights stop being the frequencies, so the differences between the classes are exaggerated

With the axis starting at 4 the bar for the class holding 4 pumpkins is drawn only 0 units tall and the bar for the class holding 15 is drawn 11 units tall, so a real ratio of 4 to 15 is shown as 0 to 11. The chart still looks like a histogram, which is what makes it dangerous. A frequency axis begins at 0 for the same reason the bars touch: the picture has to say what the data says and nothing more.

P9 no — the classes are different widths, so a bar on one covers twice as much of the mass scale as a bar on the other and the heights are not comparable

A bar’s height is the number of measurements in the stretch of scale it covers. A class 10 kg wide catches roughly twice as many as a class 5 kg wide from the same data, whatever the data is like: here the tallest bar goes from 15 to 28 simply because the classes doubled. Comparing those two heights measures the class width, not the pumpkins. The comparison becomes legitimate only when both pictures use one set of classes and one frequency axis.

P10 4 and 16

The class $25 \le x < 30$ is entirely above 22 kg, so its 4 pumpkins certainly count. The class $20 \le x < 25$ straddles 22 kg: none of its 12 pumpkins need be above 22 kg, and all of them could be. So the answer lies between 4 and 16. This is the honest limit of a grouped table — it can answer exactly at a boundary and only within a range in between, which is precisely what was traded away when the raw masses were grouped.

P15 a bar chart, because the number of children is discrete — counted in whole numbers — and no household can have a value between them

The variable is **discrete**: it is counted, and its possible values are 0, 1, 2, 3, 4 or 5 and nothing in between. A household with 1.5 children does not exist, so the stretch of scale between 1 and 2 is genuinely empty and the gap on the chart is true information. The numbers looking numeric is not the test — the test is whether the in-between values exist, and here they do not.

P16 a histogram, because mass is measured on a continuous scale where every value in between is possible, so the bars touch because the classes touch

Mass is **continuous**: a tomato can weigh anything at all between the lightest and the heaviest recorded, and rounding the record to the nearest gram does not change that. So the values are grouped into 10 equal classes of 10 g that meet, and the bars are drawn from boundary to boundary with no gaps. The fact that the recorded numbers are whole is a rounding decision, not a property of mass.

P17 because the subjects have no order, but the classes of a histogram sit on a measured scale that fixes their positions

The categories of a bar chart are names. Put History first or last and no information is lost, because there is no scale saying where History belongs. The classes of a histogram are stretches of a measured scale: $10 \le x < 15$ comes after $5 \le x < 10$ because 10 kg comes after 5 kg. Move a bar and the axis becomes nonsense, which is another way of seeing why one chart has gaps and the other cannot.

P18 no — shoe size is recorded in fixed steps, so it is discrete and belongs on a bar chart with gaps

Feet are measured, but shoe *size* is not: it is issued in fixed steps, and there is no size between 7 and 8 that anybody can buy. The variable that reaches the chart is a counted label, so the bars stay apart. The general test survives: ask whether a value can fall between two labels. For foot length in centimetres, yes — histogram. For shoe size, no — bar chart.

P19 time needs a histogram because it is continuous; subject needs a bar chart because the four subjects are separate categories

One dataset, two variables, two different charts — and the decision is made per variable, never per dataset. Time is measured on a continuous scale: any value between the shortest and longest is possible, so the times are grouped into equal fixed intervals and drawn with touching bars. Subject is a set of four names with no order and nothing in between, so it is drawn as four separate bars with gaps that mean something. Drawing both the same way would misrepresent one of them.

P20 because the histogram's horizontal axis is a measured scale whose bars must touch, while the bar chart's is a list of separate categories whose gaps are information

Both pictures are 6 rectangles whose heights are frequencies, so the ink is nearly the same. What differs is what the horizontal axis *is*. On the histogram it is the mass scale itself, running unbroken from 0 to 30 kg; every point on it is a possible measurement, so every point must be covered and the bars touch. On the bar chart it is a shelf of names; the space between two names corresponds to no possible value at all, so it is left empty on purpose. Read the axis, not the rectangles.

P25 skewed to the left, because the single peak is at $30 \le x < 35$ and the tail runs 6 classes to the left against 1 to the right

Count the peaks first: the frequencies 1, 1, 2, 3, 4, 10, 16, 3 rise once and fall once, so there is 1 peak and the distribution is not bimodal. Now look at the tails either side of the peak at $30 \le x < 35$: 6 classes hold values below it and only 1 above it, so the long thin tail runs down the scale. That is **skewed to the left** — most of the lengths are large, and a few are much smaller.

P26 that two different groups have been mixed into one dataset, because the histogram is bimodal with peaks at $5 \le x < 10$ and $25 \le x < 30$

The frequencies 1, 10, 9, 2, 6, 8, 3, 1 climb to 10 at $5 \le x < 10$, fall back to 2, and climb again to 8 at $25 \le x < 30$. Two humps with a dip between them is **bimodal**, and a single population of one thing measured one way very rarely produces it. The usual explanation is that two groups have been counted together — two plantings, two age groups, two machines — and the useful response is not to average across the dip but to split the data and draw the two groups separately.

P27 because the lengths are continuous, so almost no value occurs twice; the table gives the modal class 15 ≤ x < 20 instead

A mode is a value that occurs most often. These lengths were measured to the nearest tenth of a centimetre, so repeats are accidents and mean nothing about the data. What the table can give is the class with the highest frequency — here 15 ≤ x < 20, holding 11 of the 40 lengths — and that is a statement about an interval of the scale, not about any single leaf, pod or learner.

P28 9

Add the heights of every bar from 20 ≤ x < 25 upwards: 4 + 3 + 1 + 1 = 9. That is 9 of the 40 lengths, which is the thin right-hand tail this distribution is named for — a small number of values spread over a long stretch of the scale, while 31 of the lengths sit below 20 cm.

P29 because the distribution is skewed to the right: the 6 classes of tail on the right pull the mean up, while the tallest bar stays at 5 ≤ x < 10

First check the figure you have been given, because every number in it is on the table: Σfx = 535 over 40 lengths, so the estimated mean is 535 ÷ 40 = 13.375, which is 13.4 cm to 3 significant figures. Now the argument. The modal class 5 ≤ x < 10 answers “where do most of the lengths sit?”. The mean answers “where is the balance point?” and every value pulls on it in proportion to how far out it is. Here 9 lengths lie above 20 cm, some of them far above, and each of those drags the balance point up the scale; nothing on the left can pull back as hard, because the scale stops at 0. So the mean, 13.4 cm, ends up above the modal class. In a symmetrical distribution the two would very nearly coincide, which is exactly why this slip is invisible until the shape is skewed.

P30 no — one could be symmetrical and the other bimodal, with the same tallest class and the same range

The modal class fixes one bar and the range fixes the two ends; everything between is still free. Datasets A and D make the point: both are 40 lengths in 8 classes of 5 cm, but A is symmetrical while D is bimodal, with peaks at 5 ≤ x < 10 and 25 ≤ x < 30. Shape is a statement about the whole outline, and it cannot be recovered from any one summary number — which is the case for drawing the histogram rather than quoting statistics about it.

P35 (5, 2), (15, 4), (25, 12), (35, 16), (45, 14), (55, 10), (65, 2), (75, 0), (85, 0), (95, 0)

Take the midpoints first — (lower + upper) ÷ 2 for each class, giving 5, 15, 25, 35, 45, 55, 65, 75, 85, 95 — and pair each one with its own frequency, 2, 4, 12, 16, 14, 10, 2, 0, 0, 0. That gives (5, 2), (15, 4), (25, 12), (35, 16), (45, 14), (55, 10), (65, 2), (75, 0), (85, 0), (95, 0). The points are joined in order and the line stops at the last one; the three zero heights at the top of the scale are real, because those classes genuinely hold no tomatoes.

P36 because two lines can be drawn on one pair of axes without hiding each other, while two sets of bars overlap

Bars are solid. Draw one histogram over another and the taller bars hide the shorter ones, so the reader has to look at two pictures and hold one in their head while reading the other. A polygon is a single line, so both distributions fit on one pair of axes and the comparison — which sits further right, which is more spread out, which has two peaks — is visible in one glance. The information is identical; only the reading is easier.

P37 because there is no class beyond it, so a line drawn there would invent a frequency for measurements nobody made

The polygon plots one point per class in the table, and the table stops at $90 \le x < 100$. Continuing the line past the last point means plotting a point for a class that does not exist, giving it a height, and joining it up — three claims about data that was never collected. Note the honest exception here: this polygon does reach the axis, at (75, 0), (85, 0) and (95, 0), and it is entitled to, because those classes really are in the table and really do hold nothing.

P38 60, the number of tomatoes in the dataset

Add the heights of all 10 points: $0 + 0 + 0 + 2 + 8 + 16 + 13 + 12 + 8 + 1 = 60$. A frequency polygon is the same table as the histogram, drawn differently, so the heights must add to n exactly as the bar heights do. It is a free check — if the points do not add to 60, either a class has been missed or a height has been misread.

P39 the late planting is heavier — modal class 50 ≤ x < 60 against 30 ≤ x < 40 — the two are equally spread, 70 g of the scale for the late planting and 70 g for the early one — and both are symmetrical

Centre. The late polygon lies to the right of the early one along its whole length; the modal class moves from 30 ≤ x < 40 to 50 ≤ x < 60, a shift of about 20 g. **Spread.** The early masses run over 70 g of the scale and the late ones over 70 g, so the two are equally variable — the late crop is heavier, not more inconsistent. **Shape.** Both polygons are single symmetrical humps, so nothing has changed in the way the masses are distributed; the whole difference is a shift along the scale. All three statements are legitimate only because both plantings were counted into the same 10 classes and drawn against the same frequency axis.

P40 it gains points and follows the data more closely, but each height is based on fewer measurements, so the line becomes less smooth and eventually shows chance rather than shape

Halving the width doubles the number of points, so the line has twice as many places to bend and can follow a genuine feature the coarser table smoothed over. The price is that each height now counts about half as many measurements, so each is more affected by chance. The 80 seedling heights show both ends of that trade: at width 5 the polygon has 8 points and one clean hump, and at width 2 it has 20 points and 6 peaks, none of which is a feature of the seedlings. Splitting classes is worth doing until the wobble stops being data.

P45 the late planting is heavier: its modal class is 50 ≤ x < 60 against 30 ≤ x < 40 for the early planting

The tallest bar of the early planting stands over 30 ≤ x < 40 and the tallest bar of the late planting stands over 50 ≤ x < 60, so the whole distribution has moved about 20 g up the scale. The supporting evidence is the tails: 2 early tomatoes weigh 60 g or more, against 34 late ones. Because both tables use the same 10 classes, those two counts mean the same thing and may be set beside each other.

P46 because a bar covering 20 g of the scale collects far more tomatoes than a bar covering 10 g, so the heights measure the class width and not the crop

Each bar height is a count of everything inside its own stretch of the mass scale. Double the stretch and you roughly double the count, whatever the crop is like — the same effect that turns the 60 pumpkin masses from a tallest bar of 15 at width 5 into a tallest bar of 28 at width 10. Set the two pictures side by side and a reader compares heights that were never measuring the same thing. The fix is not a caption; it is to recount both datasets into one set of classes.

P47 the distribution on the shorter axis looks taller and more concentrated than it is, because the eye compares the heights on the page and not the numbers

Both charts are drawn correctly, which is what makes this harder to spot than a missing gap. A frequency of 10 fills half the height of an axis that stops at 20 and a quarter of one that stops at 40, so the same count looks twice as impressive on one page as on the other. A reader who is comparing shapes rather than reading numbers off the axis will draw the wrong conclusion every time. One axis, one scale, both charts — and if the two datasets are different sizes, say so, because a taller bar may only mean more measurements.

P48 they are equally spread: 70 g of the scale for the early planting and 70 g for the late one

The two tables span 70 g and 70 g of the mass scale respectively, so the late crop is not more variable — it is simply heavier. The counts say the same thing: each planting spreads its 60 tomatoes over 7 non-empty classes of the same width. Separating *centre* from *spread* matters here, because “bigger” and “more variable” are different claims and only one of them is true.

P49 centre: the late planting is heavier, modal class 50 ≤ x < 60 against 30 ≤ x < 40; spread: equal, 70 g of the scale for the late planting and 70 g for the early one; shape: both are symmetrical

Centre. The modal class moves from 30 ≤ x < 40 to 50 ≤ x < 60 and every one of the 34 late tomatoes weighing 60 g or more has only 2 early counterparts, so the late crop sits higher up the scale by about 20 g. **Spread.** The tables span 70 g and 70 g of the scale, so the two are equally variable; the shift is not accompanied by any extra inconsistency. **Shape.** Each table has 1 peak and reads as symmetrical, so the distributions have the same form. **The licence for all three.** Both plantings were counted into the same 10 classes of 10 g and drawn against the same frequency axis; without that, none of the three sentences could be defended.

P50 nothing about shape or centre — the second dataset has three times as many measurements, so its bars are taller everywhere; only the proportions in each class may be compared

Frequency is a count, so a dataset three times the size produces bars about three times as tall in every class, whatever its shape. Reading “taller everywhere” as “bigger measurements” confuses how many with how much. Two honest routes are open. Compare the shapes, which are unaffected by n — where the peak sits, which way the tail runs, how many humps. Or convert each frequency to a proportion of its own total, so that 11 out of 40 and 33 out of 120 become directly comparable. Both routes require what this concept insists on: identical classes.

P55 the classes of the second table are pairs of the first, so its frequencies are the first table's added in twos: 5 + 11 = 16 and 15 + 13 = 28 and 12 + 4 = 16

Nothing was measured again. The class 0 ≤ x < 10 is exactly 0 ≤ x < 5 and 5 ≤ x < 10 joined together, so its frequency is 5 + 11 = 16, and the same happens down the table: 16, 28, 16. Σf is unchanged at 60, because no pumpkin was added or removed. What changed is how much of the mass scale each bar is responsible for, and that is enough to change the picture completely.

P56 because 3 classes is too few bars to show a shape: an outline needs several bars before it can rise, fall and have tails

A shape is a statement about how the frequencies change across the scale. With 3 bars, too few bars to show a shape, there is one rise and one fall and nothing else to see: the same three heights 16, 28, 16 would be produced by a great many different distributions, including the clearly symmetrical one the width-5 table shows. Naming a shape from it would be claiming detail the chart does not carry.

P57 because each class holds only a handful of pumpkins, so one or two extra values move a bar noticeably and the ups and downs are chance, not a feature of the field

With 60 pumpkins spread over 15 classes the average class holds about 4, and the frequencies run 2, 2, 3, 1, 8, 6, 6, 4, 8, 4, 7, 3, 4, 0, 2. At that size a single extra pumpkin is a visible step. Weigh a second field of 60 pumpkins from the same crop and the peaks would land in different places, which is the test of whether a feature is real. Reporting 6 peaks here would be reporting which classes happened to catch an extra value.

P58 16 from the width-5 table; the width-10 table has no boundary at 20 kg

In the width-5 table 20 kg is a boundary, so add the bars above it: 12 + 4 = 16, exactly. In the width-10 table 20 kg falls in the middle of the class $10 \le x < 20$, which holds 28 pumpkins and says nothing about how they sit inside it — the answer could be anything from 16 to 44. Wider classes do not just blur the shape; they destroy the questions the table can answer exactly.

P59 width 5, because 6 bars show the single symmetrical hump and the fall-away at the top of the scale, while width 10 gives only 3 bars and too few bars to show a shape

Width 5 buys detail at the cost of smaller counts per class: 6 bars, frequencies 5, 11, 15, 13, 12, 4, every one of them big enough to be stable, and the outline reads as symmetrical with a clear drop from 12 to 4 at the top of the scale. Width 10 buys stability at the cost of nearly everything else: 3 bars, frequencies 16, 28, 16, and too few bars to show a shape — the drop at the top has disappeared inside the last block. Neither is arithmetically wrong; the width-5 chart answers more questions, so it is the one to publish, and the sentence that publishes it must say which width was used.

P60 test whether the peak survives at other widths and in other samples; a feature that exists at only one width is chance, not shape

Searching through widths until something appears is a way of guaranteeing that something appears: with enough narrow classes every dataset grows extra peaks, as the 60 pumpkin masses do at width 2, where 6 peaks turn up in 15 classes and none of them mean anything. The honest test has two parts. Redraw at two or three other widths and see whether the second peak is still there — a real one, like the two humps of dataset D, survives regrouping. Then ask whether there is a reason for it in the world: two plantings, two batches, two age groups. A peak that survives no width but its own and has no story behind it should be reported as what it is, an artefact of the class width.