●●●●○ Level 4 · Mathematics M4.7

Statistics

Take fair samples, sum up data with numbers and charts, and compare groups.

Learn why some samples mislead and how to take a fair one. Find averages from frequency tables, draw cumulative frequency graphs, box plots and histograms, read scatter graphs and time series, and spot a graph or an average that does not tell the truth.

YOU ARE READY FOR THIS BOOK IF

You can find the mean, median and range of 8 numbers.

One step up	M5.7 Probability distributions and inference
This book	**M4.7 Statistics**
One step down	M3.7 Statistics and chance

Contents

M4.7.1 Samples and bias

M4.7.2 Averages from frequency tables

M4.7.3 Cumulative frequency

M4.7.4 Box plots

M4.7.5 Histograms with unequal widths

M4.7.6 Scatter graphs and correlation

M4.7.7 Data over time

M4.7.8 Telling the truth with data

Key words, A to Z

Written and drawn by attnom. Free to read, print and share.

M4.7.1 CHAPTER 1

Samples and bias

> **Ravi asks 20 people leaving the sports hall if they exercise every day. Do his answers tell us about the whole town?**

YOUR GUESS

Pick one before you read on. It is fine to be wrong.

- ☐ **A** Yes. Twenty people are enough to tell us about the town.
- ☐ **B** No. People at a sports hall are more likely to exercise than other people.
- ☐ **C** Yes, as long as he asks 20 more people at the same place.

IF THIS FEELS HARD, READ THIS FIRST

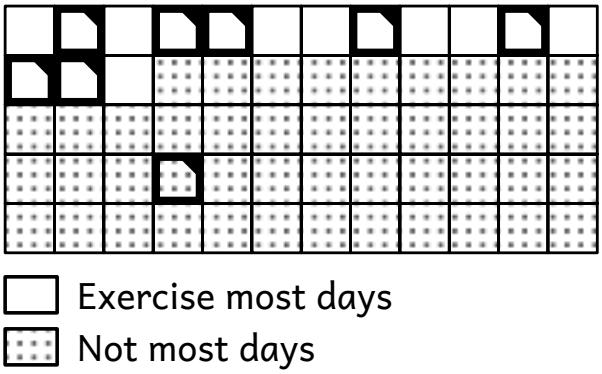
- Level 3, Statistics and chance, M3.7.1. If planning a survey and asking a clear question looks new, read this first.

BY THE END YOU CAN

- Explain how a sample can be biased
- Take a random sample using random numbers
- Choose a sample that matches the groups in a population

The **population** is everything we want to know about: all the plants in a field, or all the people in a town. Asking every member takes too long, so we look at a few. Those few are the **sample**.

Ravi's town has a street with 60 people. 15 of them exercise most days. Ravi stands at the sports hall and asks 8 people leaving. 7 of them say yes.



Framed: the 8 people Ravi asked. Almost all of them are from the small white group.

In the sample, 7 of 8 people exercise. In the street, only 15 of 60 do, which is 1 in 4. Ravi's sample gives the wrong picture. Here the cause is the place: people at a sports hall are more likely to exercise.

A sample is **biased** when some members of the population are more likely to be chosen than others. The results then lean one way. A bias comes from how members are chosen. A bigger sample taken the same way is just as biased.

A **random sample** gives every member of the population the same chance of being chosen. A fair way to make one is with random numbers. Number the members from 1 to the size of the population. Then let random numbers choose.

DRAW IT FIRST

A random sample of 5 plants from a bed of 40

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40

1 Number every plant in the bed, from 1 to 40.

1	2	3	**4**	5	6	7	8	9	10
11	12	13	14	15	16	**17**	18	19	20
21	22	23	24	25	26	27	28	**29**	30
31	32	33	34	35	**36**	37	**38**	39	40

sample of 5

2 Read random numbers two digits at a time. Skip 00, any number above 40, and any repeat.

Suppose the random pairs are 82, 17, 36, 17, 04, 55, 29, 38. Plant 82 and plant 55 do not exist, so skip them. The second 17 is a repeat, so skip it. The sample is plants 17, 36, 4, 29 and 38.

WORKED EXAMPLE

The pairs 52, 09, 17, 40, 09, 31 are random. A bed has 40 plants. Choose a sample of 4.

1. Read the pairs in order. 52 is above 40, so skip it.
2. 09 is plant 9. 17 is plant 17. 40 is plant 40.
3. The second 09 is a repeat, so skip it.
4. 31 is plant 31. The sample is plants 9, 17, 40 and 31.

Answer: Plants 9, 17, 40 and 31

A random sample can still be odd by luck. That is not bias. Bias is built into the way members are chosen. Luck changes from one sample to the next.

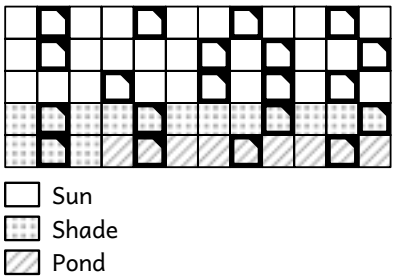
Sometimes a population has groups that differ. A **stratified sample** takes from each group in proportion to its size. A group that is half the population gives half the sample. The share from a group is group size ÷ whole population × sample size.

SEE IT THREE WAYS

With things

A field has 60 plants: 36 in the sun, 15 in the shade and 9 by the pond. You want a sample of 20.

As a picture



In symbols

Sun: 36 ÷ 60 × 20 = 12. Shade: 15 ÷ 60 × 20 = 5. Pond: 9 ÷ 60 × 20 = 3. Together 12 + 5 + 3 = 20.

Inside each group, choose the members at random. Stratifying keeps the groups in proportion. Choosing at random keeps each member equally likely. If a share is not

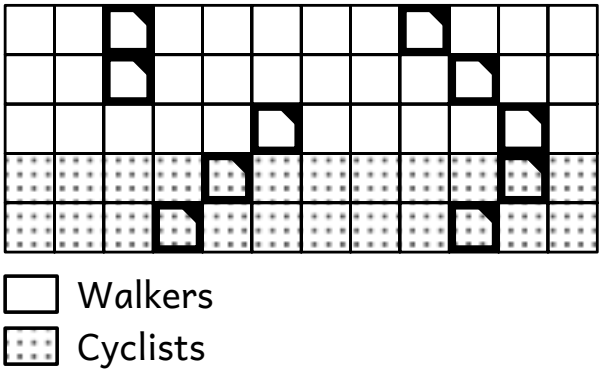
a whole number, round it and check that the total is still the sample size.

WORKED EXAMPLE

A club has 60 members: 36 walkers and 24 cyclists. Choose a stratified sample of 10.

1. Walkers: 36 ÷ 60 × 10 = 6.
2. Cyclists: 24 ÷ 60 × 10 = 4.
3. Check the total: 6 + 4 = 10, the size of the sample.
4. Choose 6 walkers at random and 4 cyclists at random.

Answer: 6 walkers and 4 cyclists



Of the 10 chosen, 6 are walkers and 4 are cyclists, the same shares as the club.

TRY IT

Take a random sample with paper slips

You need: 20 small pieces of paper, a pencil, a bag or a bowl.

1. Write the numbers 1 to 20 on the slips, one number on each. Fold them.
2. Put the slips in the bag and shake well.
3. Without looking, take out 5 slips one at a time and keep them out. Write the numbers.

What you will see: Every slip had the same chance. Nobody chose by looking. If you do it again you will get different numbers, and that is fine.

If you cannot try it, this is what happens.

1	2	**3**	4	5	6	**7**	8	9	10
11	**12**	13	14	15	**16**	17	18	**19**	20

A WRONG IDEA MANY PEOPLE HAVE

~~A big sample cannot be biased.~~

Why it feels right: More people sounds like a better picture of the population.

The test that breaks it: Ravi asks 500 people at the sports hall. Most of them exercise, as before. More people from the same place repeat the same lean.

What is true: Bias comes from how members are chosen, not from how many. A smaller random sample can be fairer than a big biased one.

KEY WORDS

population
everything we want to know about.

sample
a part of the population that we look at.

biased
leaning one way, because some members are more likely to be chosen.

random sample
a sample where every member has the same chance of being chosen.

stratified sample
a sample that takes from each group in proportion to its size.

BACK TO YOUR GUESS

The answer is B: No. People at a sports hall are more likely to exercise than other people.

People at a sports hall are more likely to exercise than the rest of the town, so his sample leans one way. Asking more people at the same place does not cure that. A fair sample gives every person in the town the same chance of being asked.

CHECK YOURSELF

Try each one before you look. A wrong answer tells you where to go next.

1. Amina wants to know how many hours people in a town sleep. Which way of choosing 50 people is the best?

   A Ask 50 of her friends.

   B Ask the first 50 people to buy bread at a bakery that opens before dawn.

   C Choose names at random from a list of everyone in the town.

   Answer at the back of the book.

2. Ravi asks 500 people at the sports hall. He says his sample is too big to be biased. What do you say?

   A A big sample can still be biased if the way of choosing favours some people.

   B He is right. A sample of more than 100 cannot be biased.

   C He is wrong. A sample is fair only if it is more than half the population.

   Answer at the back of the book.

3. What makes a sample random?

   A The person picking chooses members who look typical.

   B Every member of the population has the same chance of being chosen.

   C The members are chosen quickly, without thinking much.

   D The sample is large.

   Answer at the back of the book.

4. Plants in a bed are numbered 1 to 40. Random pairs are 47, 12, 38, 12, 05, 26. Which plants are in a sample of 4?

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40

   A Plants 47, 12, 38 and 12

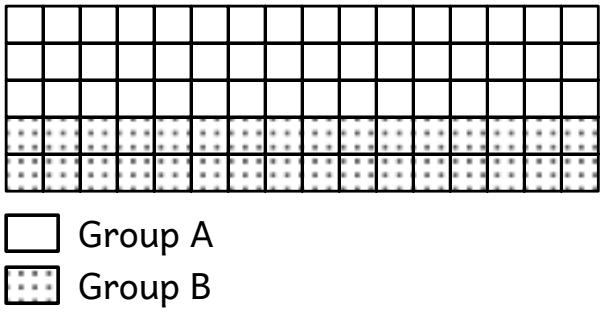
   B Plants 12, 38, 12 and 5

   C Plants 4, 7, 1 and 2

   D Plants 12, 38, 5 and 26

   Answer at the back of the book.

5. A population of 80 has two groups, A with 48 members and B with 32. A stratified sample of 10 is wanted. How many come from group A?



A 5

B 6

C 4

Answer at the back of the book.

6. A radio show asks listeners to phone in a vote on a new bus route. Why is the result likely to be biased?

A Only people who listen and care enough to phone take part.

B The radio show might count the votes wrongly.

C Phoning in is a random way to choose people.

Answer at the back of the book.

7. A garden has 90 plants: 60 bean plants and 30 pea plants. A stratified sample of 15 is wanted. How many bean plants are in it?

Answer at the back of the book.

8. A field has 100 plants: 40 beans, 35 peas and 25 maize. A stratified sample of 20 is wanted. How many maize plants are in it?

Answer at the back of the book.

KEEP IT

Copy this into your exercise book.

A sample is biased if some members are more likely to be chosen. A random sample gives every member the same chance. A stratified sample takes group ÷ whole × sample size from each group.

Draw: A grid of squares with two groups and a few squares framed, and the share sum under it.

M4.7.2 CHAPTER 2

Averages from frequency tables

> **Mei opens 25 pea pods and counts the peas in each. How can she find the mean number of peas without adding 25 numbers?**

YOUR GUESS

Pick one before you read on. It is fine to be wrong.

- ☐ **A** Add 3, 4, 5, 6 and 7 from the table, then divide by 5.
- ☐ **B** Pick the number of peas that most pods have.
- ☐ **C** Multiply each number of peas by how many pods have it, add these up, then divide by the number of pods.

IF THIS FEELS HARD, READ THIS FIRST

- Level 3, Statistics and chance, M3.7.3. If finding the mean, median and mode of a short list looks new, read this first.
- Level 2, Tables and charts, M2.7.1. If reading a frequency table looks new, read this first.

BY THE END YOU CAN

- Find the mean from a frequency table
- Estimate the mean from grouped data
- Find the group that holds the median

Mei opens 25 pods. Instead of writing 25 numbers, she writes a **frequency table**. It lists each number of peas, and the **frequency**, f, which is how many pods have that number.

peas x	pods f
3	2
4	5
5	9
6	6
7	3

Two pods have 3 peas. Five pods have 4 peas. Nine pods have 5 peas.

To find the mean we need two totals: the number of peas, and the number of pods. A row of the table is a group of equal values. Nine pods with 5 peas each hold 9 × 5 = 45 peas. So multiply x by f for every row.

DRAW IT FIRST

Build the mean from the table

peas x	pods f
3	2
4	5
5	9
6	6
7	3

1 Write each number of peas, x, and its frequency, f.

peas x	pods f	f × x
3	2	6
4	5	20
5	9	45
6	6	36
7	3	21
total	25	128

2 Add a column f × x. Add up the f column and the f × x column. The totals are 25 pods and 128 peas.

The mean is the total of all the values divided by how many values there are. In symbols, mean = Σfx ÷ Σf. The sign Σ means "add up". So Σfx = 128 is the total number of peas, and Σf = 25 is the number of pods.

WORKED EXAMPLE

Find the mean number of peas in a pod, from the table of 25 pods.

1. Multiply each x by f: 3 × 2 = 6, 4 × 5 = 20, 5 × 9 = 45, 6 × 6 = 36 and 7 × 3 = 21.
2. Add the f × x column: 6 + 20 + 45 + 36 + 21 = 128. That is the number of peas.
3. Add the frequencies: 2 + 5 + 9 + 6 + 3 = 25. That is the number of pods.
4. Mean = 128 ÷ 25 = 5.12 peas in a pod.

Answer: 5.12 peas in a pod

A WRONG IDEA MANY PEOPLE HAVE

~~**The mean of a frequency table is the mean of the x column.**~~

Why it feels right: The x column lists the values, so it looks like the data.

The test that breaks it: The x values 3, 4, 5, 6 and 7 have mean 5. But 9 pods have more than 5 peas and only 7 have fewer, so the real mean is a little above 5. It is 5.12.

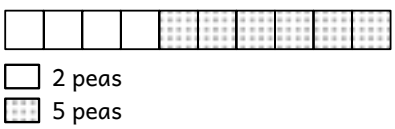
What is true: Each x counts f times. Multiply x by f first, then divide by the total frequency, not by the number of rows.

SEE IT THREE WAYS

With things

A bowl holds 10 pods. Four pods have 2 peas each and six pods have 5 peas each.

As a picture



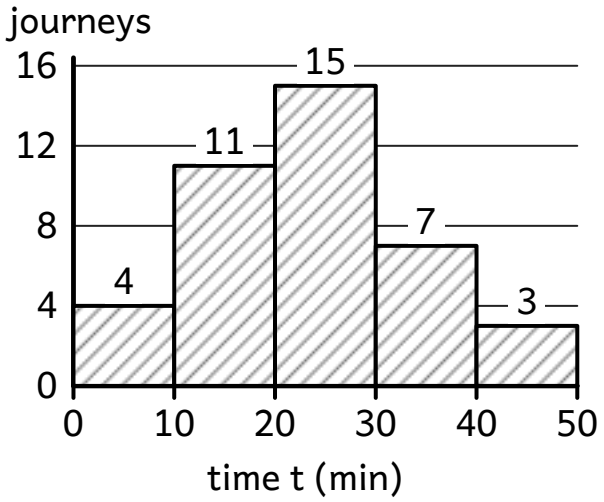
In symbols

f × x: 4 × 2 = 8 and 6 × 5 = 30. Σfx = 38 peas. Σf = 10 pods. Mean = 38 ÷ 10 = 3.8.

Now measured data. Times and heights are often grouped in classes. The table of 40 journey times below says 4 journeys took 0 or more minutes but under 10. We no longer know each time, only its class.

t (min)	f	midpoint
$0 \le t < 10$	4	5
$10 \le t < 20$	11	15
$20 \le t < 30$	15	25
$30 \le t < 40$	7	35
$40 \le t < 50$	3	45

$0 \le t < 10$ means 0 or more, but under 10 minutes.



The classes have equal widths, so the height of a bar is the frequency.

To find a mean we need one value for each class. We use the **midpoint**, the number halfway between the two ends of the class. For $10 \le t < 20$ the midpoint is $(10 + 20) \div 2 = 15$. We pretend every journey in the class took 15 minutes.

This is a guess, so the answer is only an **estimate**. Some journeys took less than 15 minutes and some more. We hope the errors cancel.

WORKED EXAMPLE

Estimate the mean journey time from the table of 40 journeys.

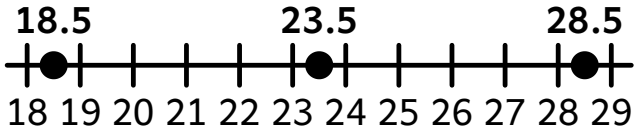
1. The midpoints are 5, 15, 25, 35 and 45.
2. Multiply each midpoint by f: 5 × 4 = 20, 15 × 11 = 165, 25 × 15 = 375, 35 × 7 = 245 and 45 × 3 = 135.
3. Add them: 20 + 165 + 375 + 245 + 135 = 940.
4. Estimate of the mean = 940 ÷ 40 = 23.5 minutes.

Answer: 23.5 minutes

x	f	f × x
5	4	20
15	11	165
25	15	375
35	7	245
45	3	135

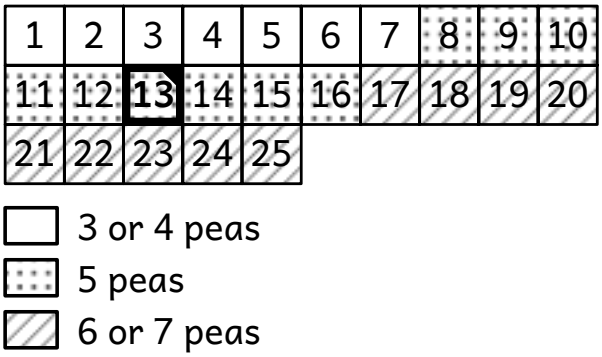
The f × x column adds up to 940. The f column adds up to 40.

How far can the estimate be wrong? If every journey took the least its class allows, the mean would be 740 ÷ 40 = 18.5. If every journey took almost the most, it would be about 1 140 ÷ 40 = 28.5. The true mean lies between 18.5 and 28.5, and 23.5 sits in the middle.



Times in minutes. The dots are the least the mean could be, the estimate, and the most it could be.

Now the median. With n values in order, the median is the (n + 1) ÷ 2 th value. For 25 pods it is the 13th pod. From a table we cannot see the 13th value, but we can see which row holds it. Add up the frequencies until the running total reaches 13.



Pod 13 is the middle one of 25. It lies in the group with 5 peas.

WORKED EXAMPLE

Find the class that holds the median of the 40 journey times.

1. There are 40 values. The median is the (40 + 1) ÷ 2 = 20.5th value.
2. The 20.5th value is not in the first two classes, which hold 4 + 11 = 15 values.
3. It is in the third class, because the first three classes hold 15 + 15 = 30 values.
4. So the median is in the class $20 \leq t < 30$.

Answer: The class $20 \leq t < 30$

We can name the class that holds the median, but not the median itself. The most common class is the **modal class**, the one with the greatest frequency. Here that is also $20 \leq t < 30$.

TRY IT

Throw *a* die 20 times

You need: *a* die from *a* game, paper and *a* pencil.

1. Throw the die 20 times. After each throw, put *a* tally mark by that number.
2. Write the frequency table: x is the face, f is the number of tallies.
3. Work out f × x for each row, add them up, and divide by 20.

What you will see: The mean is close to 3.5, the mean of the six faces. Your 20 throws may not give exactly 3.5, and that is fine.

If you cannot try it, this is what happens.

face x	throws f	f × x
1	2	2
2	4	8
3	3	9
4	5	20
5	3	15
6	3	18
total	20	72

KEY WORDS

frequency table
a table that lists each value and how many times it happens.

midpoint
the number halfway between the two ends of *a* class.

estimate
an answer that is close, not exact, because some information is missing.

modal class
the class with the greatest frequency.

BACK TO YOUR GUESS

The answer is C: Multiply each number of peas by how many pods have it, add these up, then divide by the number of pods.

The table says 2 pods have 3 peas, 5 pods have 4 peas, and so on. Multiply each number of peas by its number of pods, and add: 128 peas in all. Then divide by the 25 pods. The mean is 128 ÷ 25 = 5.12 peas in a pod.

CHECK YOURSELF

Try each one before you look. A wrong answer tells you where to go next.

1. The table shows the seeds in 20 pods. What is the mean number of seeds in a pod?

seeds x	pods f
1	5
2	8
3	4
4	3

A 2.5

B 45

C 5

D 2.25

Answer at the back of the book.

2. The table shows the heights of 30 plants in classes. Which is the best estimate of the mean height?

height h (cm)	f
$0 \leq h < 10$	3
$10 \leq h < 20$	9
$20 \leq h < 30$	12
$30 \leq h < 40$	6

A 20 cm

B 22 cm

C 660 cm

D 17 cm

Answer at the back of the book.

3. Waiting times of 45 visitors are in classes. Which class holds the median?

wait t (min)	f
$0 \leq t < 10$	12
$10 \leq t < 20$	20
$20 \leq t < 30$	6
$30 \leq t < 40$	4
$40 \leq t < 50$	3

A $10 \leq t < 20$

B $20 \leq t < 30$

C $0 \leq t < 10$

Answer at the back of the book.

4. What is the midpoint of the class $40 \leq h < 60$?

   A 40

   B 20

   C 50

   D 60

   Answer at the back of the book.

5. Why is a mean worked out from grouped data only an estimate?

   A Grouped data have fewer values than the original data.

   B We know only which class each value is in, so we use the midpoint as its value.

   C The midpoint is always lower than the real mean.

   Answer at the back of the book.

6. In a frequency table, what does $f \times x$ give for one row?

   A The number of values in that row.

   B The mean of the values in that row.

   C The total of all the values in that row.

   Answer at the back of the book.

7. **From an earlier chapter**

   A garden has 80 plants and 30 of them are herbs. A stratified sample of 16 is wanted. How many herbs are in it?

   Answer at the back of the book.

8. Pods hold 1, 2, 3, 4 or 5 peas. The numbers of pods are 3, 5, 6, 4 and 2. What is the mean number of peas in a pod?

   Answer at the back of the book.

9. Plant heights, in four classes of width 10 cm from 0 to 40 cm, have frequencies 2, 6, 8 and 4. Estimate the mean height.

   Answer at the back of the book.

KEEP IT

Copy this into your exercise book.

Mean = Σfx ÷ Σf. For classes, use midpoints and say estimate. Median place = (n + 1) ÷ 2: add the frequencies until the running total reaches it.

Draw: A table with columns x, f and f × x, and the two totals under it.

Cumulative frequency

> **A bus company times 80 journeys and sorts the times into classes. How can one graph show the middle time, and how many journeys took under 25 minutes?**

YOUR GUESS

Pick one before you read on. It is fine to be wrong.

- ☐ **A** Pick the middle class from the table and call its midpoint the median.
- ☐ **B** Find the class with the highest frequency.
- ☐ **C** Plot the running total of the frequencies, then read across from half of 80, which is 40.

IF THIS FEELS HARD, READ THIS FIRST

- Level 3, Statistics and chance, M3.7.3. If finding the median of a list looks new, read this first.
- Level 4, Statistics, M4.7.2. If finding the class that holds the median looks new, read this first.

BY THE END YOU CAN

- Draw a cumulative frequency graph
- Read the median and quartiles from the graph
- Find the interquartile range

The table shows the times of 80 bus journeys. A **cumulative frequency** is a running total. It tells how many values are below the top of a class.

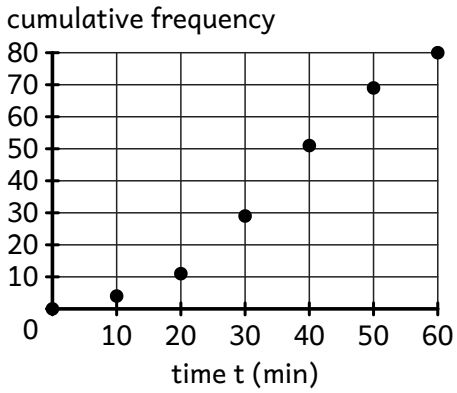
t (min)	f	cf
0 ≤ t < 10	4	4
10 ≤ t < 20	7	11
20 ≤ t < 30	18	29
30 ≤ t < 40	22	51
40 ≤ t < 50	18	69
50 ≤ t < 60	11	80

4 + 7 = 11, then 11 + 18 = 29, and so on. The last total is all 80 journeys.

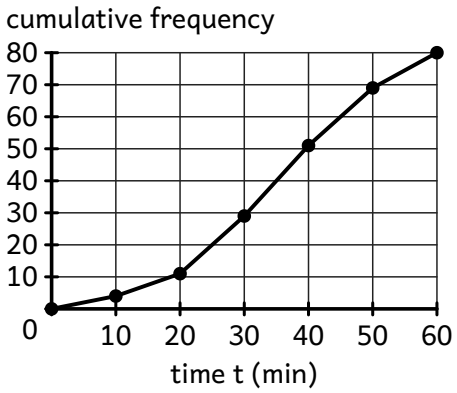
Look at the class 10 ≤ t < 20. Every journey in it took under 20 minutes. So the running total of 11 is the number of journeys under 20 minutes. It belongs at 20, the **upper boundary** of the class, not at its middle.

DRAW IT FIRST

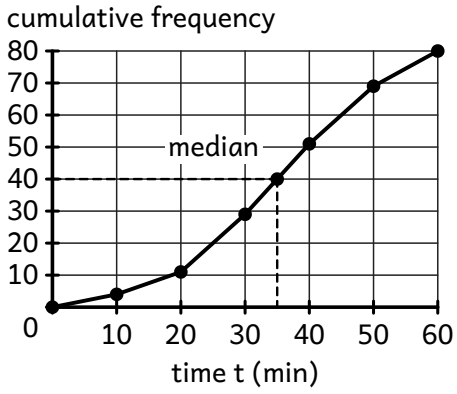
Draw the cumulative frequency graph



1 Plot (upper boundary, cumulative frequency) for each class. Start at (0, 0), because no journey took under 0 minutes.



2 Join the points with straight lines. The graph never goes down, and ends at 80.

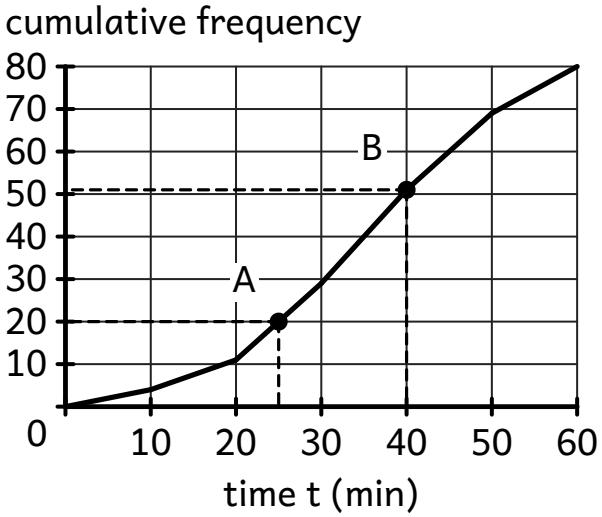


3 To find the median, go across from 40, half of 80, to the graph. Then go down to the time axis.

The graph lets us read other values. How many journeys took under 25 minutes? Go up from 25 on the time axis to the graph, then across to the cumulative frequency axis. The reading is 20.

Joining the points with straight lines treats the journeys in each class as spread evenly along it. So a reading between two plotted points is an estimate. Some books

draw a smooth curve instead, and its readings differ a little.



A: about 20 journeys took under 25 minutes. B: 51 journeys took under 40 minutes.

WORKED EXAMPLE

How many of the 80 journeys took 40 minutes or more?

1. Read up from 40 minutes on the time axis. The graph shows 51 journeys.
2. Check with the table of cumulative frequencies: 4 + 7 + 18 + 22 = 51.
3. The others took 40 minutes or more. There are 80 journeys in all.
4. 80 − 51 = 29 journeys.

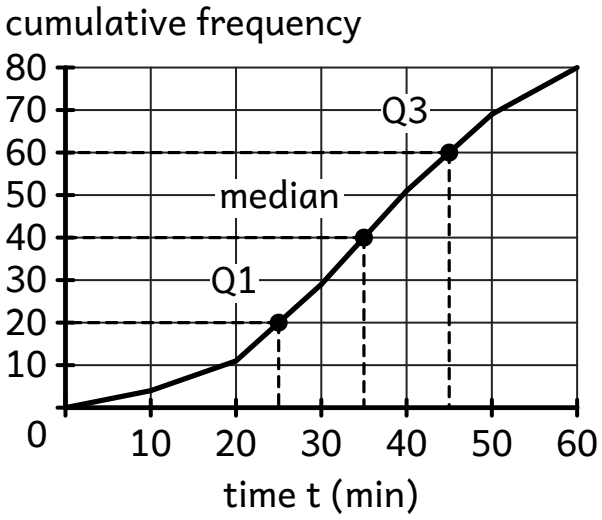
Answer: 29 journeys

Now the median and the quartiles. The **median** is the middle value. The **lower quartile**, Q1, has a quarter of the values below it. The **upper quartile**, Q3, has three quarters of the values below it.

For 80 values we read the graph at 40 for the median, 20 for Q1 and 60 for Q3. Why 40, and not the 40.5th value of chapter 2? The graph does not show 80 separate journeys. Its height at a time is how many journeys took less than that time. The median is the time with half the journeys below it. Half of 80 is 40, so we read across from 40.

In the same way Q1 has a quarter below it, so we read at 80 ÷ 4 = 20. The graph is an estimate made from classes. Reading at 40.5 instead of 40 gives about 35.2

minutes, not 35, and that small gap does not matter here. For a short list of separate values we count instead, as the next chapter shows.



Q1 is read at 20, the median at 40 and Q3 at 60.

WORKED EXAMPLE

Read the median and the two quartiles of the 80 journey times from the graph.

1. The median is read at 80 ÷ 2 = 40. Across to the graph and down gives an estimate of 35 minutes.
2. Q1 is read at 80 ÷ 4 = 20. Across to the graph and down gives an estimate of 25 minutes.
3. Q3 is read at 3 × 80 ÷ 4 = 60. Across to the graph and down gives an estimate of 45 minutes.

Answer: The median is 35 minutes. The lower quartile is 25 minutes and the upper quartile is 45 minutes.

The **interquartile range**, IQR, is Q3 − Q1. It is the length of the middle half of the data. The range, highest minus lowest, depends on two extreme values. A single odd value can change it a lot. The IQR leaves out the lowest quarter and the highest quarter, so one odd value cannot move it much.

WORKED EXAMPLE

Find the interquartile range of the journey times.

1. From the graph, the upper quartile is 45 minutes.
2. The lower quartile is 25 minutes.
3. IQR = Q3 − Q1 = 45 − 25 = 20 minutes.
4. The middle half of the journeys fit within a spread of 20 minutes.

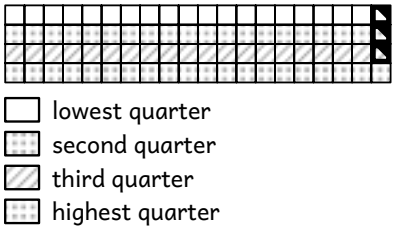
Answer: 20 minutes

SEE IT THREE WAYS

With things

Line up 80 journeys from the shortest to the longest. Split the line into four equal groups. Walk to the end of the first group, the second group, and the third group.

As a picture



In symbols

80 ÷ 4 = 20. 80 ÷ 2 = 40. 3 × 80 ÷ 4 = 60. Read the graph at 20, 40 and 60.

TRY IT

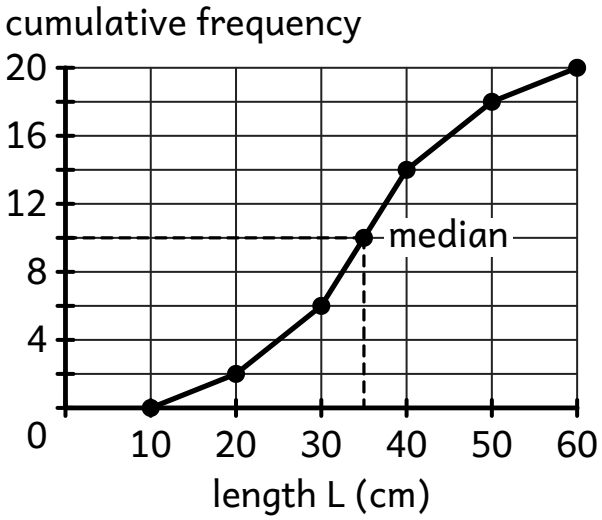
Graph the lengths of 20 sticks

You need: 20 sticks between 10 cm and 60 cm long, a ruler, squared paper.

1. Measure each stick. Sort them into classes: $10 \le L < 20$, $20 \le L < 30$ and so on up to 60.
2. Write the frequencies, then the running totals.
3. Plot (upper boundary, running total), starting at (10, 0). Join the points with straight lines. Read the median at 10.

What you will see: The line never goes down, because a running total never goes down. It ends at 20, the number of sticks. Here, the median is about 35 cm.

If you cannot try it, this is what happens.



A WRONG IDEA MANY PEOPLE HAVE

~~**Plot each cumulative frequency at the middle of its class.**~~

Why it feels right: Bar charts and mean calculations use the middle of a class, so it seems natural here too.

The test that breaks it: The running total of 11 for $10 \leq t < 20$ counts journeys that took under 20 minutes. If it were plotted at 15, the graph would say that 11 journeys took under 15 minutes. That is not known.

What is true: A running total up to the end of a class belongs at the upper boundary of that class.

KEY WORDS

cumulative frequency
a running total of the frequencies.

upper boundary
the top end of a class, where its cumulative frequency is plotted.

quartile
a value that splits ordered data into quarters. Q1 has a quarter below it, Q3 a quarter above.

interquartile range
the upper quartile minus the lower quartile, the length of the middle half of the data.

BACK TO YOUR GUESS

The answer is C: Plot the running total of the frequencies, then read across from half of 80, which is 40.

A running total of the frequencies is plotted, and the graph is read at 40, half of 80. Reading across from 40 to the graph and down to the time axis gives the median. The middle class by position can be the wrong class. The class with the highest frequency is the modal class, not the median.

CHECK YOURSELF

Try each one before you look. A wrong answer tells you where to go next.

1. Heights of 60 plants are in classes. Which point do you plot for the class $20 \leq h < 30$?

h (cm)	f
$0 \leq h < 10$	4
$10 \leq h < 20$	5
$20 \leq h < 30$	15
$30 \leq h < 40$	15
$40 \leq h < 50$	12
$50 \leq h < 60$	9

A (25, 24)

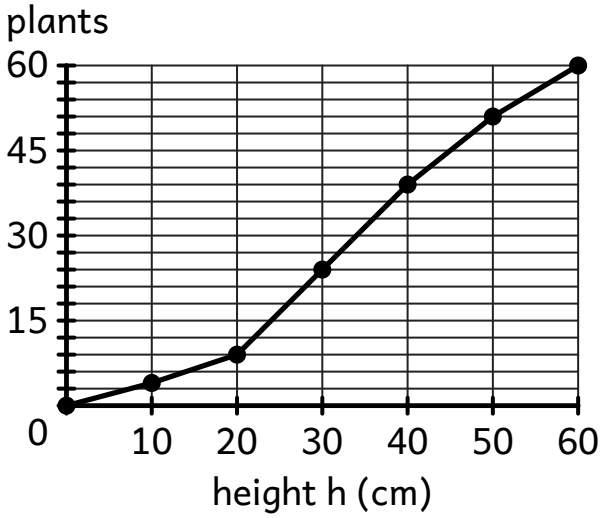
B (30, 15)

C (20, 24)

D (30, 24)

Answer at the back of the book.

2. This graph shows the heights of the 60 plants. What is the median height?



A 34 cm

B 30 cm

C 24 cm

D 45 cm

Answer at the back of the book.

3. For some data, Q1 = 22 and Q3 = 41. What is the interquartile range?

A 63

B 41

C 19

Answer at the back of the book.

4. Why is the interquartile range often a better measure of spread than the range?

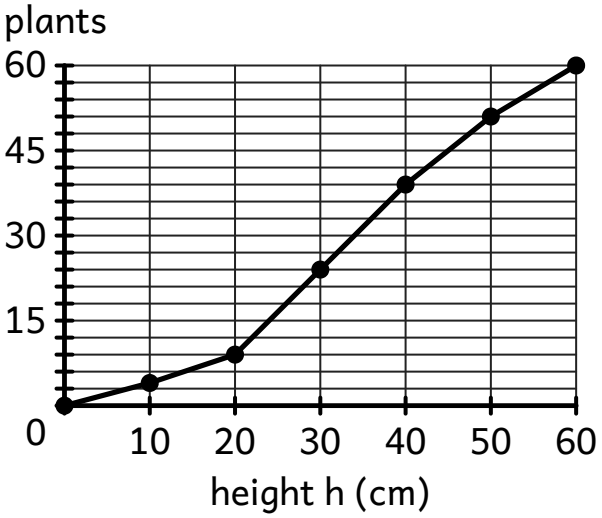
A It uses every value in the data, so it is more accurate.

B It leaves out the lowest and highest quarters, so one odd value cannot change it much.

C It is always bigger than the range.

Answer at the back of the book.

5. Use the graph of the 60 plants. How many plants are shorter than 40 cm?



A 15

B 21

C 39

D 24

Answer at the back of the book.

6. A graph shows 80 values. At which cumulative frequency do you read the upper quartile?

A 20

B 40

C 75

D 60

Answer at the back of the book.

7. **From an earlier chapter**

Using the table of 80 bus journeys, estimate the mean journey time. Use the midpoints of the classes.

Answer at the back of the book.

8. The lower quartile of some data is 31 and the upper quartile is 52. What is the interquartile range?

Answer at the back of the book.

9. A cumulative frequency graph of 60 values reads 42 at a height of 40 cm. How many values are 40 cm or more?

   Answer at the back of the book.

KEEP IT

Copy this into your exercise book.

Cumulative frequency is a running total, plotted at the upper boundary. Read the median at n ÷ 2, Q1 at n ÷ 4, Q3 at 3n ÷ 4. IQR = Q3 − Q1.

Draw: A rising line through points, with dashed lines from 20, 40 and 60 on the upright axis down to the other axis.

M4.7.4 CHAPTER 4

Box plots

> **Two groups of plants have the same median height. How can I show in one small picture that one group is much more spread out?**

YOUR GUESS

Pick one before you read on. It is fine to be wrong.

- ☐ **A** Draw a box plot of each group, on the same scale.
- ☐ **B** Draw a bar chart of the two medians.
- ☐ **C** Add up the heights in each group.

IF THIS FEELS HARD, READ THIS FIRST

- Level 3, Statistics and chance, M3.7.4. If comparing two sets of data with an average and the range looks new, read this first.
- Level 4, Statistics, M4.7.3. If the median, the quartiles and the interquartile range look new, read this first.

BY THE END YOU CAN

- Draw a box plot from five key values
- Compare two groups using box plots

Amina grows 11 bean plants in the sun. After two weeks their heights in centimetres, in order, are 12, 15, 17, 18, 20, 22, 23, 25, 27, 30, 31.

Five values sum up the list: the smallest, the lower quartile Q1, the median, the upper quartile Q3 and the largest. A **box plot** draws these five values on one scale.

For a short list, we find them like this. This is not the n ÷ 4 reading of the last chapter. A graph reads a running total. A short list has separate values, so we count them. The median is the middle value. Split the ordered list into a lower half

and an upper half. If the list has an odd number of values, leave the middle value out of both halves. Q1 is the median of the lower half. Q3 is the median of the upper half.

WORKED EXAMPLE

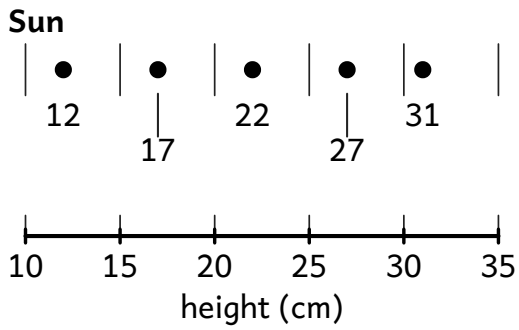
Find the five key values of Amina's 11 heights: 12, 15, 17, 18, 20, 22, 23, 25, 27, 30, 31.

1. There are 11 values, so the median is the (11 + 1) ÷ 2 = 6th value, which is 22.
2. Leave 22 out. The lower half is 12, 15, 17, 18, 20. Its middle value is the 3rd, so Q1 = 17.
3. The upper half is 23, 25, 27, 30, 31. Its middle value is the 3rd, so Q3 = 27.
4. The smallest value is 12 and the largest is 31.

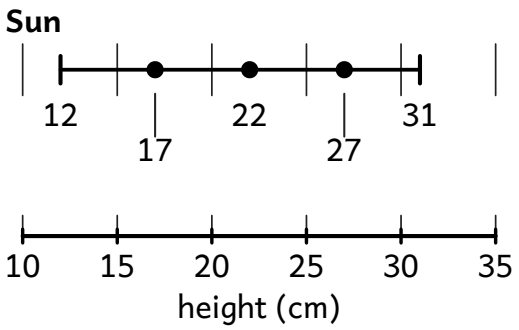
Answer: Smallest 12, lower quartile 17, median 22, upper quartile 27, largest 31

DRAW IT FIRST

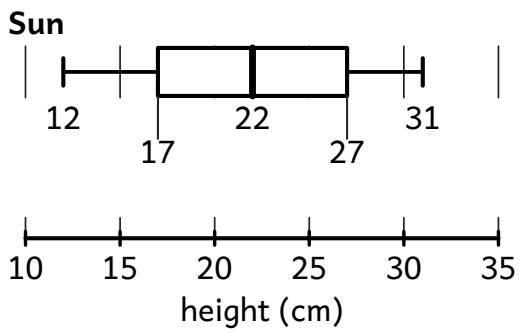
Draw the box plot



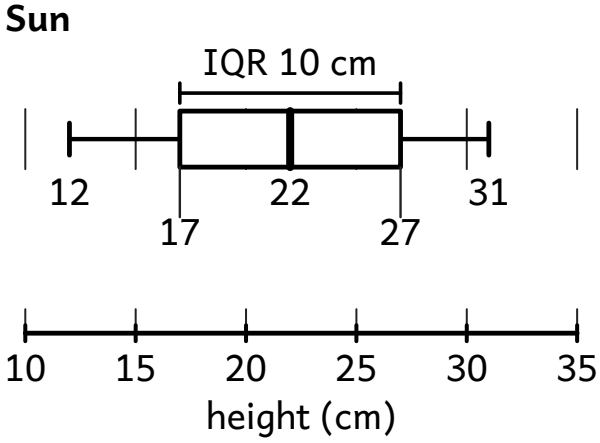
1 Draw a scale and mark the five key values: 12, 17, 22, 27 and 31.



2 Draw a line from the smallest value to the largest.



3 Draw a box from Q1 to Q3. Draw a thick line across the box at the median. The parts of the line outside the box are the **whiskers**.



The box is the interquartile range: 27 − 17 = 10 cm. The whole line is the range: 31 − 12 = 19 cm.

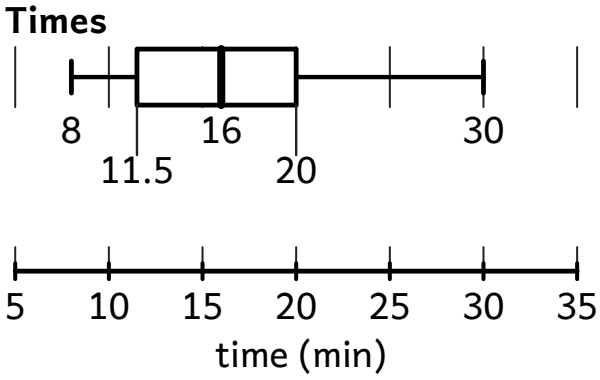
A box plot splits the data into four parts. These are the two whiskers and the two halves of the box. About a quarter of the values lie in each part. The box holds about half of them.

WORKED EXAMPLE

Find the five key values of 12 journey times (min): 8, 9, 11, 12, 14, 15, 17, 18, 19, 21, 24, 30.

1. There are 12 values, so the median is halfway between the 6th and 7th values, 15 and 17. (15 + 17) ÷ 2 = 16.
2. The lower half is the first 6 values: 8, 9, 11, 12, 14, 15. Its median is halfway between 11 and 12, which is 11.5.
3. The upper half is the last 6 values: 17, 18, 19, 21, 24, 30. Its median is halfway between 19 and 21, which is 20.
4. The smallest is 8 and the largest is 30.

Answer: Smallest 8, lower quartile 11.5, median 16, upper quartile 20, largest 30



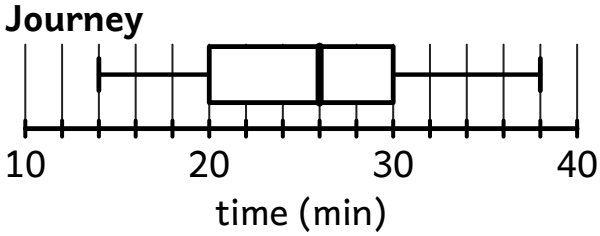
The median 16 is not in the middle of the box. The box runs from 11.5 to 20, so its middle is 15.75.

WORKED EXAMPLE

Read the box plot of journey times below. Find the median, the interquartile range and the range.

1. The whiskers run from 14 to 38. The thick line in the box is at 26, so the median is 26.
2. The box runs from 20 to 30.
3. IQR = 30 − 20 = 10 minutes.
4. Range = 38 − 14 = 24 minutes.

Answer: The median is 26 minutes, the IQR is 10 minutes and the range is 24 minutes



Read each end of the box and each whisker against the scale.

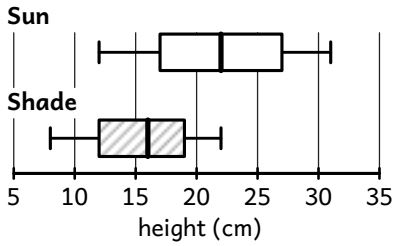
To compare two groups, put their box plots on the same scale. Compare the medians to compare the middle. Compare the IQRs to compare the spread. A bigger IQR means more spread. A group with a smaller spread is more consistent.

SEE IT THREE WAYS

With things

Amina also grows 11 bean plants in the shade. Their heights in order are 8, 10, 12, 13, 15, 16, 17, 18, 19, 20, 22.

As a picture



In symbols

Sun: median 22, IQR 27 − 17 = 10.
Shade: median 16, IQR 19 − 12 = 7.

value	Sun	Shade
smallest	12	8
lower quartile	17	12
median	22	16
upper quartile	27	19
largest	31	22

The same numbers as the box plots.

WORKED EXAMPLE

Compare the heights of the plants in the sun and in the shade, using the median and the IQR.

1. The medians are 22 cm in the sun and 16 cm in the shade.
2. The IQRs are 10 cm in the sun and 7 cm in the shade.
3. So the plants in the sun are taller in the middle, with a median 6 cm higher.
4. The sun plants are also more spread out, because their IQR is 3 cm bigger.

Answer: The sun plants have the higher median, 22 cm against 16 cm, and the larger IQR, 10 cm against 7 cm

TRY IT

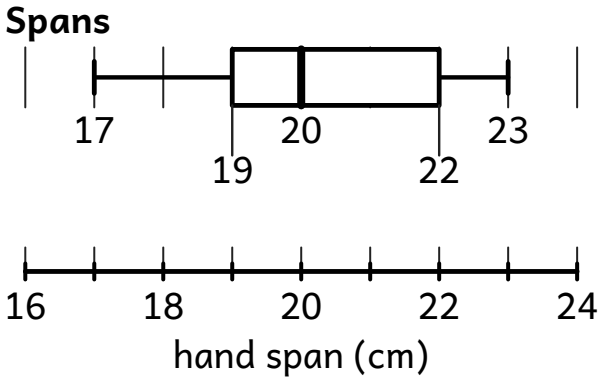
Draw a box plot of 11 hand spans

You need: 11 people to measure, a ruler, squared paper.

1. Measure each hand span, from the tip of the thumb to the tip of the little finger, with the hand spread flat.
2. Write the 11 spans in order. Find the smallest, Q1, the median, Q3 and the largest.
3. Draw a scale on squared paper. Draw the line, the box and the thick median line.

What you will see: The box holds about the middle half of the spans. Compare your box plot with a friend's. Which is more spread out?

If you cannot try it, this is what happens.



A WRONG IDEA MANY PEOPLE HAVE

~~The longer whisker holds more of the values.~~

Why it feels right: A longer line looks like more of something.

The test that breaks it: In Amina's sun list the lower whisker is 5 cm long and the upper whisker is 4 cm long. But each holds two values beyond the box: 12 and 15 below it, 30 and 31 above it.

What is true: Each of the four parts holds about a quarter of the values, however long it is. A long part means the values in it are more spread out.

A WRONG IDEA MANY PEOPLE HAVE

~~The median is always in the middle of the box.~~

Why it feels right: The median is the middle value, and the box looks like the middle.

The test that breaks it: The shade plants have Q1 = 12 and Q3 = 19. The middle of the box is 15.5. But the median is 16.

What is true: The thick line is where the median is. It can lie anywhere inside the box. If it is nearer one end, the values are more bunched up on that side.

KEY WORDS

box plot
a picture of the smallest value, Q1, the median, Q3 and the largest value on one scale.

whisker
the line from the box to the smallest or the largest value.

consistent
not spread out: the values stay close together.

BACK TO YOUR GUESS

The answer is A: Draw a box plot of each group, on the same scale.

A box plot shows the middle, the spread of the middle half, and the smallest and largest values. Two box plots on one scale can be compared at a glance. A bar chart of medians hides the spread, and a total depends on how many plants there are.

CHECK YOURSELF

Try each one before you look. A wrong answer tells you where to go next.

1. Find the lower quartile of this list: 3, 5, 8, 9, 12, 14, 15.

   A 6.5

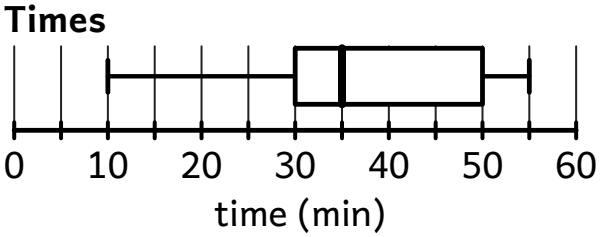
   B 3

   C 5

   D 8

   Answer at the back of the book.

2. What is the median shown by this box plot?



   A 30 minutes

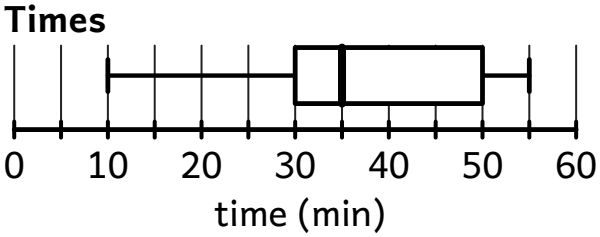
   B 35 minutes

   C 40 minutes

   D 50 minutes

   Answer at the back of the book.

3. What is the interquartile range of the same box plot?



   A 45 minutes

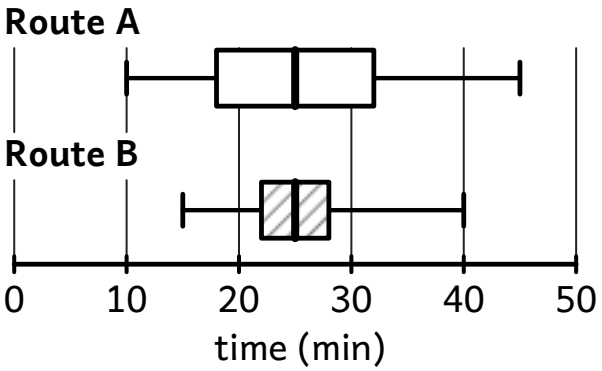
   B 35 minutes

   C 50 minutes

   D 20 minutes

   Answer at the back of the book.

4. The box plots show the journey times on two bus routes. Which statement is true?



A Route B is more consistent, because its box is shorter.

B The routes are the same, because their medians are equal.

C Route A is quicker, because its box is longer.

Answer at the back of the book.

5. About what share of the values lies inside the box of a box plot?

A About a quarter

B About half

C About three quarters

Answer at the back of the book.

6. The upper whisker of a box plot is much longer than the lower whisker. What does this tell us?

A More values lie in the upper whisker.

B The top quarter of the values is more spread out than the bottom quarter.

C More than half of the values lie above the median.

Answer at the back of the book.

7. **From an earlier chapter**

A cumulative frequency graph shows 120 values. At what cumulative frequency do you read the lower quartile?

Answer at the back of the book.

8. Find the interquartile range of these 11 values: 4, 6, 7, 10, 12, 13, 15, 18, 19, 22, 25.

   Answer at the back of the book.

KEEP IT

Copy this into your exercise book.

Box plot: smallest, Q1, median, Q3, largest on one scale. The box is about half the data. Compare the medians for the middle, the IQRs for the spread.

Draw: Two box plots on one scale, one above the other, with the five values marked under the first.

M4.7.5 CHAPTER 5

Histograms with unequal widths

> **The heights of plants are grouped in classes of different widths. How can one chart show them fairly?**

YOUR GUESS

Pick one before you read on. It is fine to be wrong.

- ☐ **A** Make each bar as tall as the frequency of its class.
- ☐ **B** Make the area of each bar match the frequency of its class.
- ☐ **C** Draw every bar the same width and leave out the scale.

IF THIS FEELS HARD, READ THIS FIRST

- Level 3, Statistics and chance, M3.7.2. If drawing a bar chart for grouped data looks new, read this first.
- Level 3, Perimeter, area and volume, M3.6.3. If area as length times width looks new, read this first.

BY THE END YOU CAN

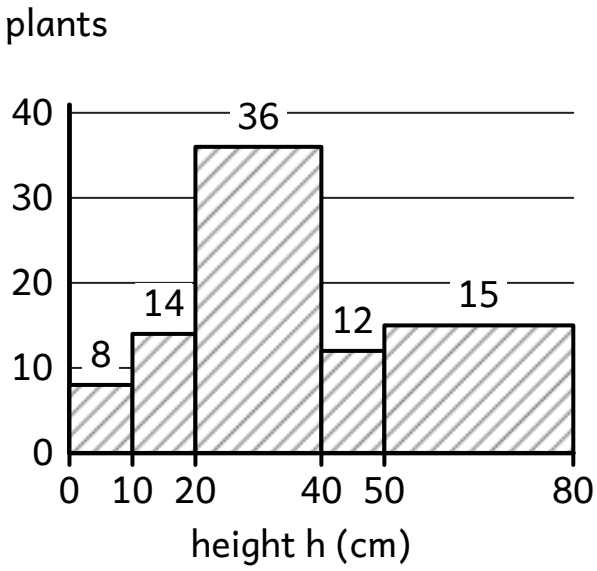
- Draw a histogram using frequency density
- Read frequencies from a histogram

A gardener measures 85 bean plants. The heights h are grouped in classes of different widths, such as $40 \le h < 50$ and $50 \le h < 80$. A bar chart with equal widths would not fit this table.

h (cm)	f
0 ≤ h < 10	8
10 ≤ h < 20	14
20 ≤ h < 40	36
40 ≤ h < 50	12
50 ≤ h < 80	15

The classes are 10, 10, 20, 10 and 30 cm wide.

Suppose we draw each bar as tall as its frequency. The picture below does that. Look at the last two bars. The class 50 ≤ h < 80 is three times as wide as the class 40 ≤ h < 50. Its frequency is only a little bigger. Yet its bar looks about four times as big.



A misleading chart. The wide last bar holds 15 plants, but looks much bigger than the bar before it, which holds 12.

The fix is to let the area of a bar show the frequency. The area of a bar is width × height. So the height must be frequency ÷ width. This height is called the **frequency density**. A chart drawn this way is a **histogram**.

WORKED EXAMPLE

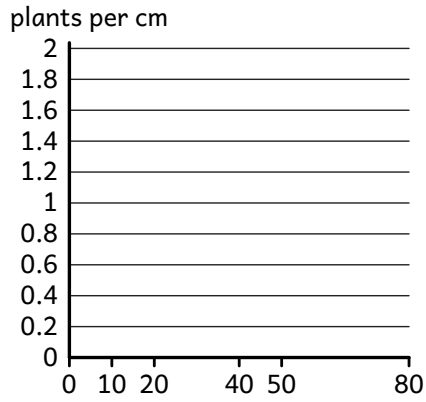
Find the frequency density of the classes $40 \leq h < 50$ and $50 \leq h < 80$.

1. The width of $40 \leq h < 50$ is $50 - 40 = 10$ cm.
2. Frequency density = frequency ÷ width = 12 ÷ 10 = 1.2 plants per cm.
3. The width of $50 \leq h < 80$ is $80 - 50 = 30$ cm. So the density is 15 ÷ 30 = 0.5 plants per cm.

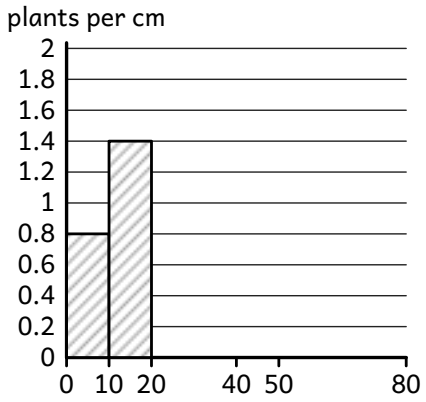
Answer: 1.2 plants per cm and 0.5 plants per cm

DRAW IT FIRST

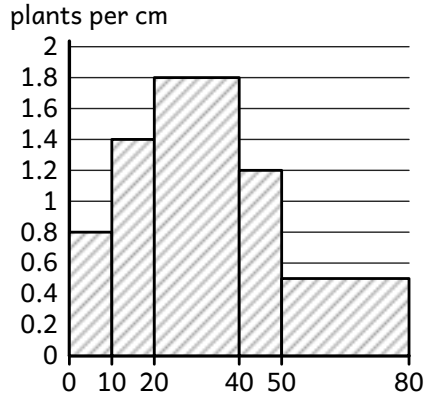
Draw the histogram



1 Draw the height axis with the true class edges: 0, 10, 20, 40, 50 and 80. Draw the frequency density axis up the side.



2 For each class, work out the density. The first two classes give 8 ÷ 10 = 0.8 and 14 ÷ 10 = 1.4. Draw their bars.

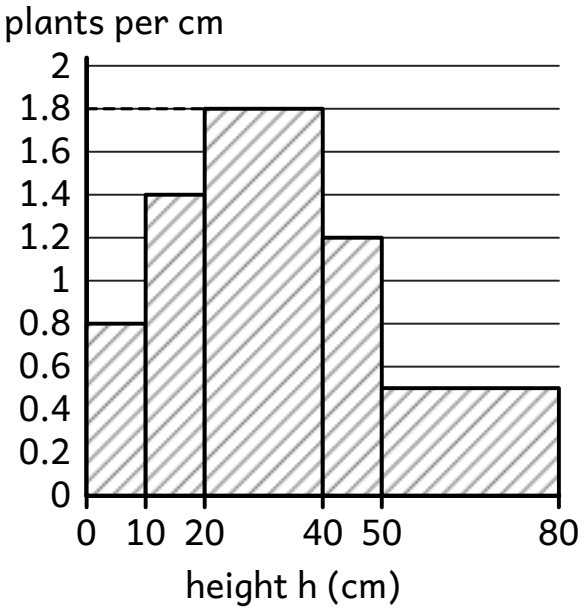


3 Draw the other three bars. 36 ÷ 20 = 1.8, 12 ÷ 10 = 1.2 and 15 ÷ 30 = 0.5.

h (cm)	f	density
0 ≤ h < 10	8	0.8
10 ≤ h < 20	14	1.4
20 ≤ h < 40	36	1.8
40 ≤ h < 50	12	1.2
50 ≤ h < 80	15	0.5

Frequency density = frequency ÷ class width.

On the histogram, the wide last bar is now low. Its area is 30 × 0.5 = 15 plants, which matches the table. The area of a bar is the frequency. This is how to read a frequency from a histogram.



The third bar reaches 1.8. A dashed line carries that height across to the axis.

WORKED EXAMPLE

The bar for 20 ≤ h < 40 reaches a height of 1.8 plants per cm. How many plants are in this class?

1. The class is from 20 to 40, so its width is 40 − 20 = 20 cm.
2. The frequency is the area of the bar: width × height = 20 × 1.8.
3. 20 × 1.8 = 36 plants.

Answer: 36 plants

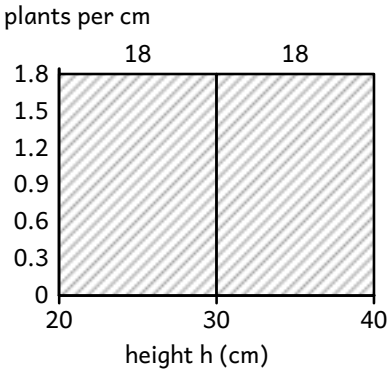
We can also find part of a class. Suppose the 36 plants in 20 ≤ h < 40 are spread evenly along the class. Then the bar has the same height all along it. A piece of the class that is half as wide holds half as many plants.

SEE IT THREE WAYS

With things

Cut the class 20 ≤ h < 40 into two pieces at 30 cm. Each piece is half the width, so it holds half of the 36 plants.

As a picture



In symbols

36 ÷ 20 = 1.8 plants per cm. A piece 10 cm wide: 10 × 1.8 = 18 plants.

WORKED EXAMPLE

Estimate how many plants are between 20 cm and 30 cm tall.

1. The plants in $20 \le h < 40$ are spread evenly, at $36 \div 20 = 1.8$ plants per cm.
2. The part from 20 to 30 is $30 - 20 = 10$ cm wide.
3. Number of plants = width × density = $10 \times 1.8 = 18$.
4. So there are about 18 plants between 20 cm and 30 cm tall.

Answer: About 18 plants

This is an estimate, because the plants are probably not spread perfectly evenly. We use it when we know nothing better. It is exact only if the spread is even.

TRY IT

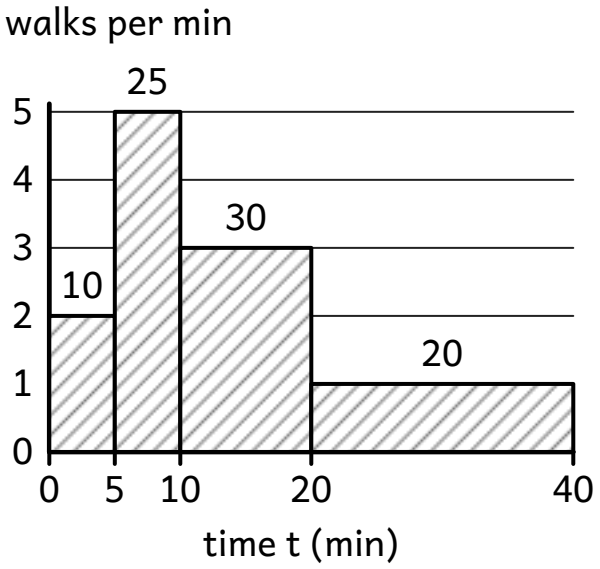
Draw a histogram on squared paper

You need: squared paper, a ruler, a pencil.

1. Use 85 walks. The classes of time t, in minutes, are $0 \le t < 5$, $5 \le t < 10$, $10 \le t < 20$ and $20 \le t < 40$.
2. The frequencies of the classes are 10, 25, 30 and 20.
3. Work out each density: frequency ÷ width. They are 2, 5, 3 and 1 walks per minute.
4. Draw the bars. Make the widths true to the scale of the time axis. Check each area: width × height should give 10, 25, 30 and 20.

What you will see: The tallest bar is for $5 \le t < 10$, but the biggest class is $10 \le t < 20$, which has the largest area, 30.

If you cannot try it, this is what happens.



A WRONG IDEA MANY PEOPLE HAVE

~~The height of a bar in a histogram is the frequency.~~

Why it feels right: In a bar chart the height of a bar is the frequency, so the same seems true here.

The test that breaks it: In the plant chart the class $50 \le h < 80$ has frequency 15, but its bar is only 0.5 high. It is wide, so its area is $30 \times 0.5 = 15$.

What is true: In a histogram with unequal widths the area of a bar is the frequency. The height is the frequency density.

A WRONG IDEA MANY PEOPLE HAVE

~~The tallest bar is always the class with the most values.~~

Why it feels right: We are used to reading the tallest bar as the biggest.

The test that breaks it: In the walks histogram the tallest bar is 5 high, but it is only 5 minutes wide, so it holds 25 walks. The bar 3 high and 10 wide holds 30.

What is true: The class with the most values has the largest area, not necessarily the largest height.

KEY WORDS

histogram
a chart for measured data in which the area of each bar is the frequency.

frequency density
frequency divided by class width. It is the height of a bar in a histogram.

BACK TO YOUR GUESS

The answer is B: Make the area of each bar match the frequency of its class.

If the classes have different widths, a bar as tall as its frequency makes wide classes look too big. Letting the area of each bar show the frequency is fair.

The height of a bar is then the frequency divided by the class width, called the frequency density.

CHECK YOURSELF

Try each one before you look. A wrong answer tells you where to go next.

1. A class $20 \leq h < 40$ has 36 plants. What is its frequency density?

   A 0.56 plants per cm

   B 720 plants per cm

   C 1.8 plants per cm

   D 36 plants per cm

   Answer at the back of the book.

2. A bar of a histogram is 6 cm wide and 2.5 plants per cm high. How many plants does it stand for?

   A 8.5

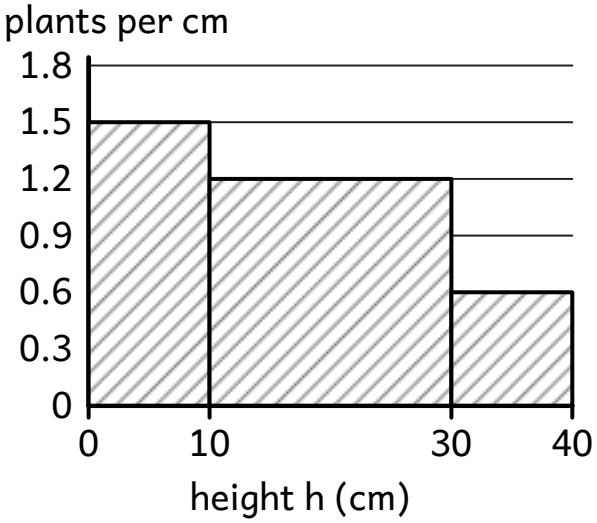
   B 2.4

   C 0.42

   D 15

   Answer at the back of the book.

3. This histogram shows the heights of some plants. Which class has the most plants?



A $0 \leq h < 10$, because it has the tallest bar

B $10 \leq h < 30$

C You cannot tell without a table of frequencies

Answer at the back of the book.

4. What does the vertical axis of a histogram with unequal class widths show?

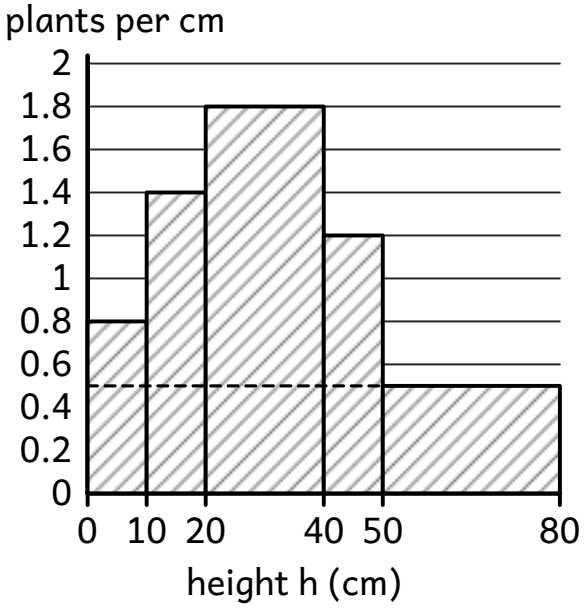
A The frequency of each class.

B The width of each class.

C The frequency density: the frequency divided by the class width.

Answer at the back of the book.

5. Plants in the class $50 \le h < 80$ are spread evenly. The bar is 0.5 high. Estimate how many plants are between 50 cm and 60 cm.



A 15

B 0.5

C 7.5

D 5

Answer at the back of the book.

6. When must the height of a bar in a histogram be the frequency density?

A When the classes have different widths.

B When there are more than 50 values.

C When the classes all have the same width.

Answer at the back of the book.

7. **From an earlier chapter**

The upper quartile of a list is the median of its upper half. Find the upper quartile of 9, 14, 16, 20, 23, 25, 31.

Answer at the back of the book.

8. A class 10 cm wide has frequency 14. What is the frequency density?

Answer at the back of the book.

9. A bar is 20 cm wide and 1.5 plants per cm high. How many plants does it stand for?

   Answer at the back of the book.

KEEP IT

Copy this into your exercise book.

Frequency density = frequency ÷ class width. The height of a bar is the density. The area of a bar is the frequency: width × height.

Draw: A histogram with three bars of different widths, and the area of the widest bar worked out beside it.

M4.7.6 CHAPTER 6

Scatter graphs and correlation

> **Mei measures the height and the arm span of 12 friends. Can a graph help her guess a friend's arm span from their height?**

YOUR GUESS

Pick one before you read on. It is fine to be wrong.

- ☐ **A** Add up all the heights and all the arm spans.
- ☐ **B** Plot each friend as one point, height across and arm span up, and look for a pattern.
- ☐ **C** Draw one bar for each friend.

IF THIS FEELS HARD, READ THIS FIRST

- Level 3, Coordinates and straight lines, M3.9.1. If plotting a point such as (160, 161) looks new, read this first.
- Level 3, Coordinates and straight lines, M3.9.7. If a line written as y = mx + c looks new, read this first.

BY THE END YOU CAN

- Draw a scatter graph and a line of best fit
- Describe positive, negative or no correlation
- Explain why correlation does not prove one thing causes another

Mei measures 12 friends. For each friend she writes the height and the arm span, from fingertip to fingertip, in centimetres. Each friend gives a pair of numbers.

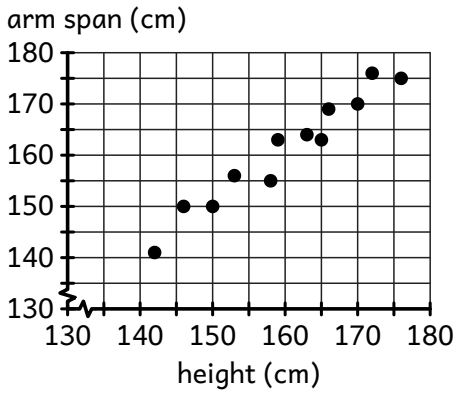
height (cm)	span (cm)
142	141
146	150
150	150
153	156
158	155
159	163
163	164
165	163
166	169
170	170
172	176
176	175

The first friend is 142 cm tall with an arm span of 141 cm.

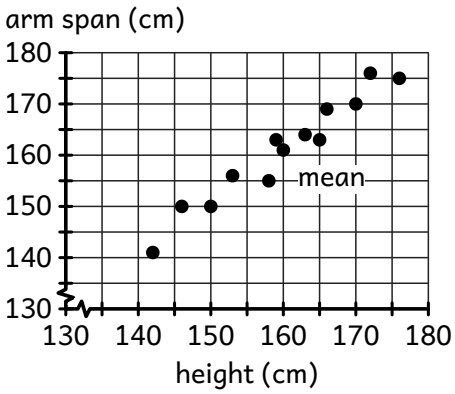
A **scatter graph** plots each pair as one point. The height is the x value and the arm span is the y value. Neither axis starts at 0 here, so each axis shows a zigzag break near its corner.

DRAW IT FIRST

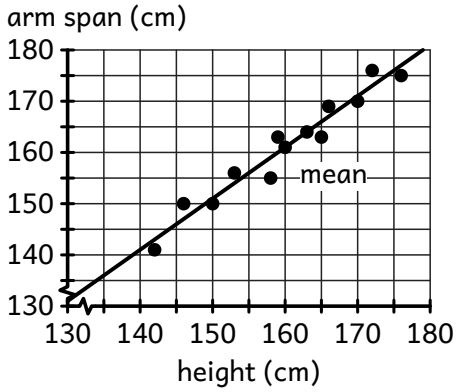
Draw the scatter graph and the line of best fit



1 Plot one dot for each friend: height across, arm span up.



2 Find the mean height and the mean arm span. Plot that mean point.



3 Draw one straight line through the middle of the dots and through the mean point. Aim for about as many dots on each side.

The straight line is the **line of best fit**. It is drawn by eye, so two people will draw slightly different lines. A good line follows the direction of the dots. It has about as many dots above it as below, and it goes through the mean point.

The line drawn here passes through (140, 141) and (170, 171). It goes up 1 for each 1 across, so its rule is $y = x + 1$.

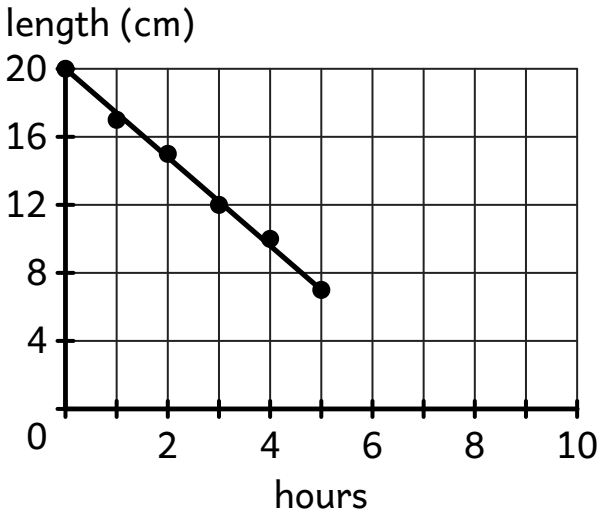
WORKED EXAMPLE

The 12 heights add up to 1 920 cm and the 12 arm spans to 1 932 cm. Find the mean point.

1. Mean height = 1 920 ÷ 12 = 160 cm.
2. Mean arm span = 1 932 ÷ 12 = 161 cm.
3. The mean point is (160, 161). The line of best fit should pass through it.
4. The line y = x + 1 passes through the mean point, because 160 + 1 = 161.

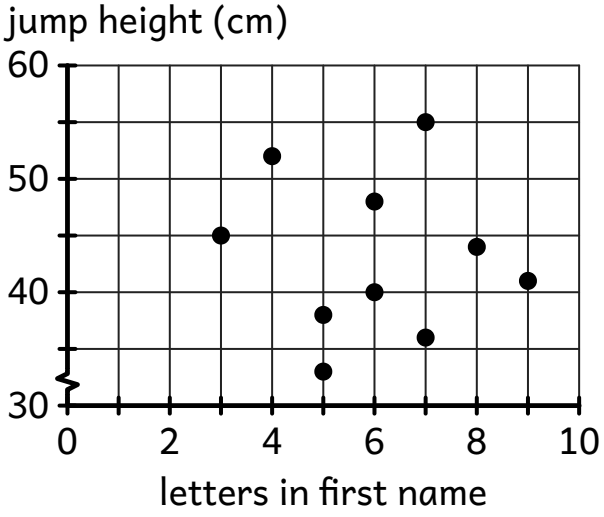
Answer: The mean point is (160, 161)

The pattern in the dots is called **correlation**. Here the dots rise from left to right. Taller friends tend to have longer arm spans. This is **positive correlation**: as one value rises, the other tends to rise.



A candle burns. The dots fall from left to right: negative correlation.

When the dots fall from left to right, as one value rises the other tends to fall. This is **negative correlation**. A burning candle gets shorter as the hours go by.



The number of letters in a name has nothing to do with jumping. The dots show no pattern: no correlation.

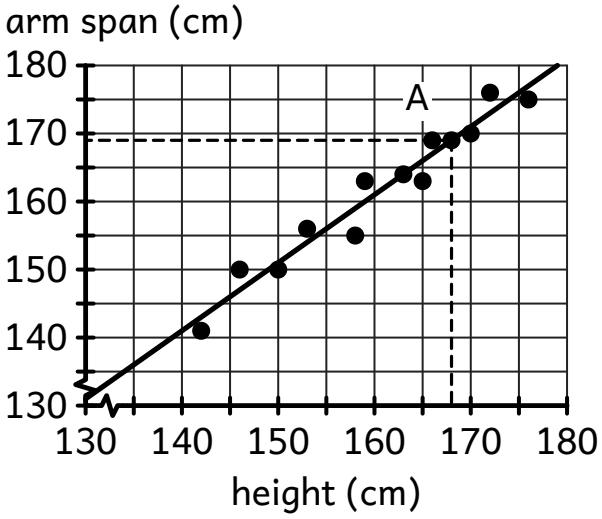
The correlation is strong when the dots lie close to a straight line, and weak when they are more spread out. The steepness of the line does not matter. A gentle slope can still have dots tight around it, as long as the line is not level.

WORKED EXAMPLE

Use the line y = x + 1 to estimate the arm spans of friends 168 cm and 150 cm tall.

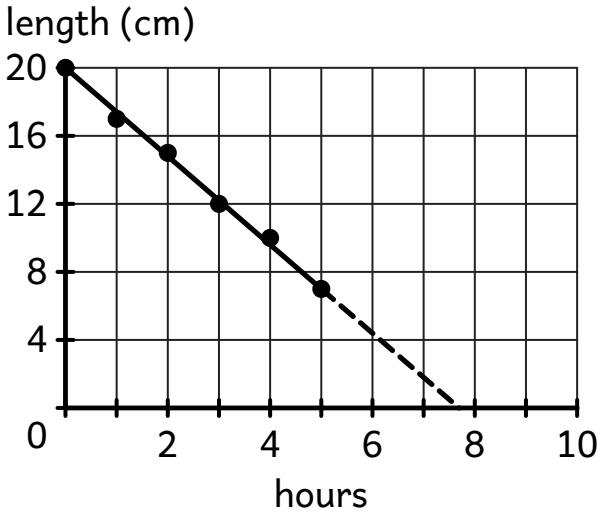
1. For a height of 168 cm, the line gives y = 168 + 1 = 169.
2. For a height of 150 cm, the line gives y = 150 + 1 = 151.
3. Both heights are inside the data, which runs from 142 cm to 176 cm. So the estimates are reasonable.

Answer: About 169 cm and about 151 cm



Go up from 168 cm to the line, then across to the arm span axis: about 169 cm.

A line of best fit is for estimates inside the range of the data. This is **interpolation**. Going beyond the data is **extrapolation**. It is unreliable, because the pattern may not carry on.



Solid: inside the data. Dashed: extended too far. The dashed line reaches 0 cm at about 8 hours. Beyond that it would give a candle with a negative length.

WORKED EXAMPLE

The candle line is y = 20 − 2.6x, with y the length in cm and x the hours. Find the length at 3.5 hours. Is the line fair to use at 10 hours?

1. At 3.5 hours the line says 2.6 × 3.5 = 9.1 cm have burned.
2. The length is 20 − 9.1 = 10.9 cm.
3. At 10 hours the line says 2.6 × 10 = 26 cm have burned.
4. That is more than the 20 cm of candle, so the length would be negative. Ten hours is outside the data, so the line cannot be used there.

Answer: About 10.9 cm. No: the line gives a negative length

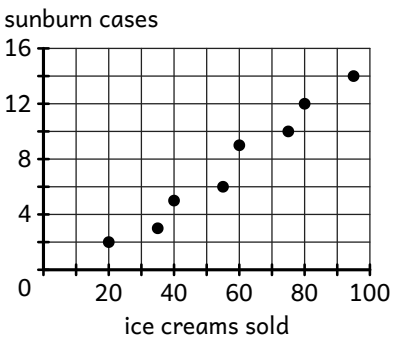
Now a warning. Two things can be correlated without one causing the other. On a beach, the number of ice creams sold each day and the number of cases of sunburn rise together.

SEE IT THREE WAYS

With things

On 8 days a beach counts the ice creams sold and the cases of sunburn.

As a picture



In words

Strong positive correlation. But ice cream does not burn skin, and sunburn does not make people buy ice cream. On hot sunny days both go up. The hot sunny weather is the hidden third factor.

In symbols

Correlation does not prove cause.

A WRONG IDEA MANY PEOPLE HAVE

~~If two things are correlated, one must cause the other.~~

Why it feels right: When two things rise together it feels as if one must be pulling the other.

The test that breaks it: Ice creams sold and sunburn cases are strongly correlated. Neither causes the other. Hot sunny weather makes both rise.

What is true: Correlation shows that two things go together. To say one causes the other we need other evidence, and we should look for a hidden third factor.

TRY IT

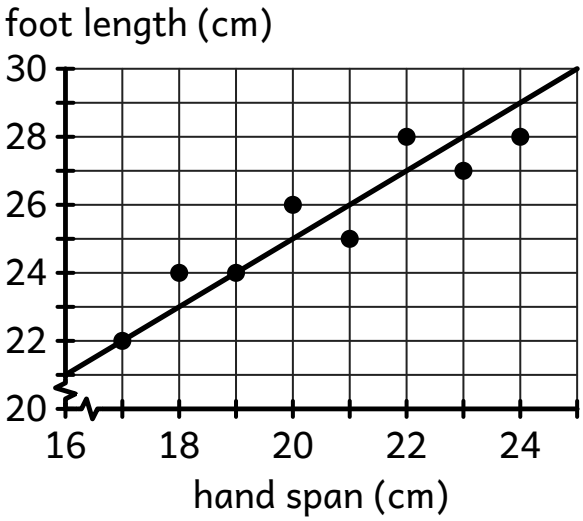
Measure hand span and foot length

You need: 8 people to measure, a ruler, squared paper.

1. For each person, measure the hand span and the foot length in centimetres.
2. Plot a dot for each person: hand span across, foot length up.
3. Find the mean point. Draw a line of best fit through it with about as many dots on each side.

What you will see: The dots rise from left to right. People with a bigger hand span tend to have a longer foot. Your own dots may be more or less spread out.

If you cannot try it, this is what happens.



A WRONG IDEA MANY PEOPLE HAVE

~~A steeper line means stronger correlation.~~

Why it feels right: A steep line looks more definite than a flat one.

The test that breaks it: Draw two scatter graphs. In one, the dots hug a flat line. In the other, the dots are spread loosely around a steep line. The first has stronger correlation.

What is true: Strength depends on how close the dots are to the line. The steepness depends on the scales chosen for the axes.

KEY WORDS

scatter graph
a graph with one dot for each pair of measurements.

correlation
the pattern between two measurements: positive, negative or none.

line of best fit
a straight line drawn through the middle of the dots on a scatter graph.

extrapolation
using a line beyond the range of the data, which is unreliable.

BACK TO YOUR GUESS

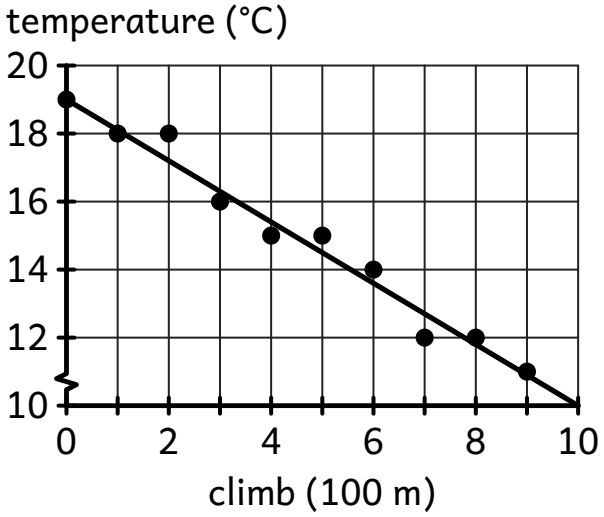
The answer is B: Plot each friend as one point, height across and arm span up, and look for a pattern.

One point for each friend, with the height across and the arm span up, makes a scatter graph. If the points fall along a line, the pattern can be used to estimate. Totals and bars lose the link between a friend's two measurements.

CHECK YOURSELF

Try each one before you look. A wrong answer tells you where to go next.

1. A walker records the temperature at different heights climbed. What does the scatter graph show?



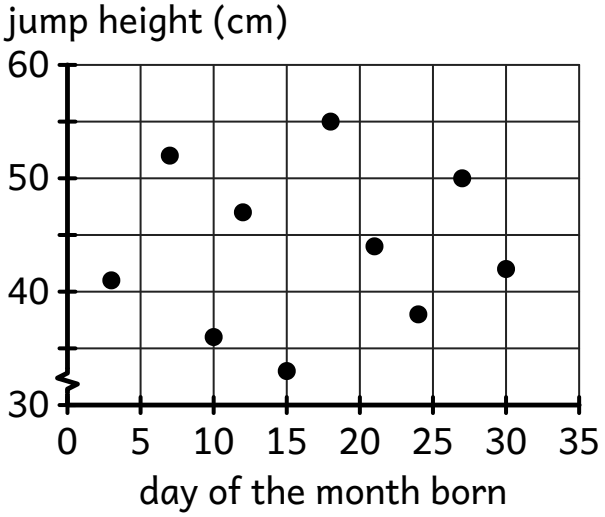
A Positive correlation

B No correlation

C Negative correlation

Answer at the back of the book.

2. Jump heights of 10 people are plotted against the day of the month on which each was born. What does this scatter graph show?



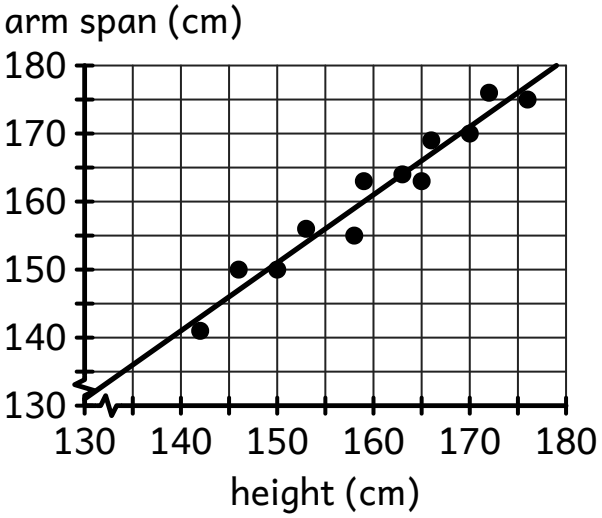
A No correlation

B Positive correlation

C Negative correlation

Answer at the back of the book.

3. The line of best fit for the arm spans is y = x + 1. Estimate the arm span of a new friend, not in the table, who is 165 cm tall.



A 165 cm

B 164 cm

C 163 cm

D 166 cm

Answer at the back of the book.

4. The number of umbrellas sold and the number of road accidents in a town rise together. What is the best conclusion?

A Umbrellas cause road accidents.

B Road accidents make people buy umbrellas.

C Rainy weather probably makes both rise.

Answer at the back of the book.

5. A line of best fit was drawn from friends whose heights run from 142 cm to 176 cm. For which height is an estimate of the arm span the least reliable?

A 150 cm

B 190 cm

C 160 cm

Answer at the back of the book.

6. What makes the correlation in a scatter graph strong?

   A The line of best fit is steep.

   B The dots lie close to a straight line.

   C There are many dots on the graph.

   Answer at the back of the book.

7. **From an earlier chapter**

   A bar of a histogram is 5 cm wide and 2.4 plants per cm high. How many plants does it stand for?

   Answer at the back of the book.

8. A line of best fit has the rule y = 2x + 3. Estimate y when x = 7.

   Answer at the back of the book.

KEEP IT

Copy this into your exercise book.

Scatter graph: one dot for each pair. Correlation: positive, negative or none. The line of best fit passes through the mean point. Use it inside the data. Correlation is not cause.

Draw: A scatter graph with a line of best fit, and under it the words positive, negative and none beside small sketches.

M4.7.7 CHAPTER 7

Data over time

> **The rain in a town rises and falls every year. How can I see a slow change over the years, and the ups and downs that repeat?**

YOUR GUESS

Pick one before you read on. It is fine to be wrong.

- ☐ **A** Find the mean of all the readings.
- ☐ **B** Draw a bar chart and look at the biggest bar.
- ☐ **C** Plot each reading against time, join the points, and look for the general direction and for ups and downs that repeat.

IF THIS FEELS HARD, READ THIS FIRST

- Level 2, Tables and charts, M2.7.6. If reading a line graph of change over time looks new, read this first.
- Level 3, Statistics and chance, M3.7.5. If drawing a line graph for data over time looks new, read this first.

BY THE END YOU CAN

- Plot a time series and describe the trend
- Spot ups and downs that repeat in a time series

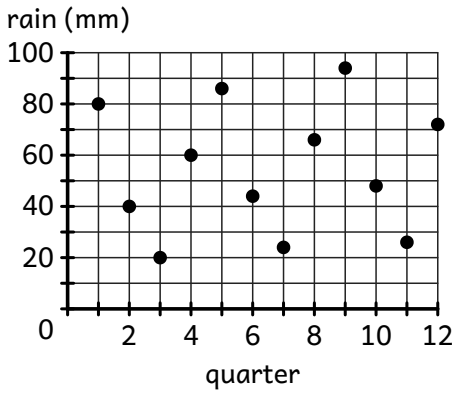
A **time series** is a set of readings taken at regular times, such as each day or each quarter of a year. The order matters. A town records its rain in millimetres for each of 12 quarters, three years in all.

quarter	rain (mm)
1	80
2	40
3	20
4	60
5	86
6	44
7	24
8	66
9	94
10	48
11	26
12	72

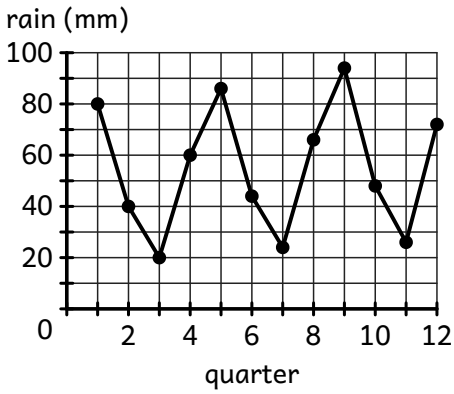
Quarters 1 to 4 are year 1, 5 to 8 are year 2, and 9 to 12 are year 3.

DRAW IT FIRST

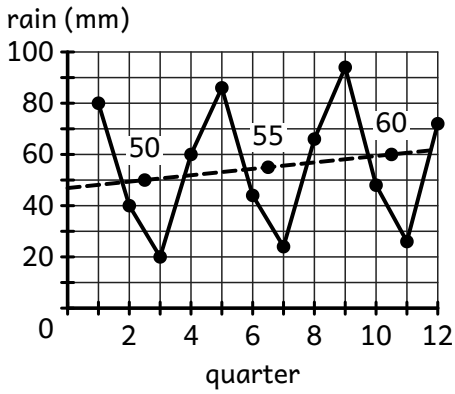
Plot the time series



1 Put time across and the reading up. Plot one dot for each quarter.



2 Join the dots in order with straight lines. The joining lines show the order, not new readings.



3 Draw a dashed line through the middle of the readings, to show the general direction.

Two things stand out. First, the readings go up and down every year: high at the start, low in the third quarter. Second, each year as a whole is a little wetter than the one before.

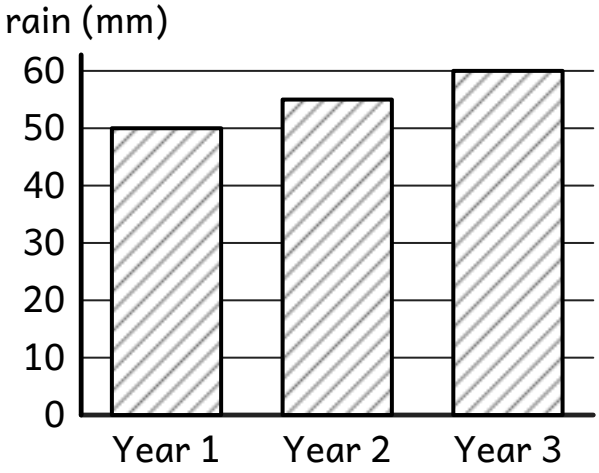
The **trend** is the general direction of the readings over a long time. Here it is a slow rise. A **seasonal pattern** is a rise and fall that repeats after the same number of readings. Here it repeats every 4 quarters.

WORKED EXAMPLE

Find the mean rain in each of the three years.

1. Year 1: 80 + 40 + 20 + 60 = 200 mm. The mean is 200 ÷ 4 = 50 mm.
2. Year 2: 86 + 44 + 24 + 66 = 220 mm. The mean is 220 ÷ 4 = 55 mm.
3. Year 3: 94 + 48 + 26 + 72 = 240 mm. The mean is 240 ÷ 4 = 60 mm.

Answer: The yearly means are 50 mm, 55 mm and 60 mm



The yearly means rise by 5 mm each year.

A mean over a whole year averages out the seasonal ups and downs. What is left is the trend. The yearly means are 50, 55 and 60 mm. They are the dots on the dashed line in the first graph, at the middle of each year.

yr	1	2	3	4
1	80	40	20	60
2	86	44	24	66
3	94	48	26	72

Read down a column. Quarter 1 is always the wettest. Quarter 3 is always the driest.

WORKED EXAMPLE

Which quarter has the lowest rain in every year? How big is the gap between the highest and the lowest quarter in year 1?

1. Year 1 readings are 80, 40, 20 and 60. The lowest, 20, is in quarter 3.
2. Year 2 readings are 86, 44, 24 and 66, and year 3 readings are 94, 48, 26 and 72. Again the lowest is in quarter 3.
3. So the pattern repeats: quarter 1 is the wettest and quarter 3 is the driest each year.
4. In year 1 the gap is 80 − 20 = 60 mm.

Answer: Quarter 3, and the gap is 60 mm

We can use the trend to look a little way ahead. The means rise by 5 mm each year. If that carries on, year 4 has a mean near 65 mm. This is a step beyond the data, so it is a cautious estimate. The further we go, the less we should trust it.

WORKED EXAMPLE

Use the yearly means 50, 55 and 60 mm to estimate the mean of year 4.

1. The first rise is 55 − 50 = 5 mm.
2. The second rise is 60 − 55 = 5 mm too. The trend is a rise of 5 mm a year.
3. If the trend carries on, year 4 has a mean of 60 + 5 = 65 mm.
4. Year 4 is only one year beyond the data, so this is a cautious estimate, not a promise.

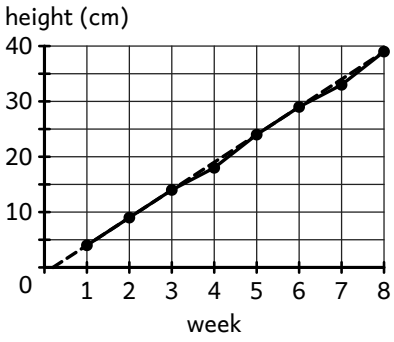
Answer: About 65 mm

SEE IT THREE WAYS

With things

A seedling is measured every week for 8 weeks. Its heights in centimetres are 4, 9, 14, 18, 24, 29, 33 and 39.

As a picture



In words

The trend is a steady rise of about 5 cm each week. There is no seasonal pattern: nothing goes up and down in a repeating way.

In symbols

The dashed line is $y = 5x - 1$. Each week adds about 5 cm.

TRY IT

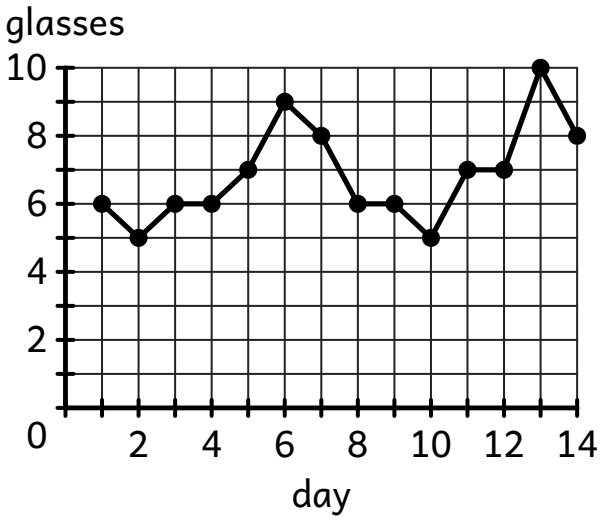
Count something each day for 14 days

You need: paper and a pencil, something you can count each day, such as glasses of water.

1. Each evening, write the day number and the count. Do this for 14 days.
2. Draw a graph with the day across and the count up. Plot a dot for each day and join them.
3. Look for a trend. Then look for a pattern that repeats, such as a high count at the end of each week.

What you will see: A pattern that repeats every 7 days is common, because many daily habits follow the week. Here, days 6 and 13 are the highest.

If you cannot try it, this is what happens.



A WRONG IDEA MANY PEOPLE HAVE

~~**If the readings go up and down, there is no trend.**~~

Why it feels right: A line that zigzags does not look like it is going anywhere.

The test that breaks it: The rain readings zigzag every year. But the yearly means are 50, 55 and 60. Each year is wetter than the last.

What is true: A trend is the general direction. Ups and downs that repeat sit on top of the trend. Averaging over a whole cycle shows the trend.

KEY WORDS

time series
readings taken at regular times, plotted in order.

trend
the general direction of a time series over a long time.

seasonal pattern
a rise and fall that repeats after the same number of readings.

BACK TO YOUR GUESS

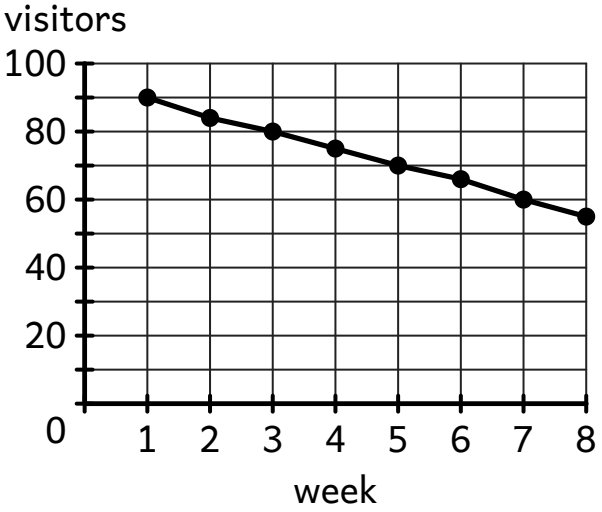
The answer is C: Plot each reading against time, join the points, and look for the general direction and for ups and downs that repeat.

A mean of all the readings hides the order in which they happened. Plotting each reading against time keeps the order. Then we can see the general direction, called the trend, and the ups and downs that repeat, called a seasonal pattern.

CHECK YOURSELF

Try each one before you look. A wrong answer tells you where to go next.

1. The graph shows the number of visitors to a park each week. Which describes the trend?



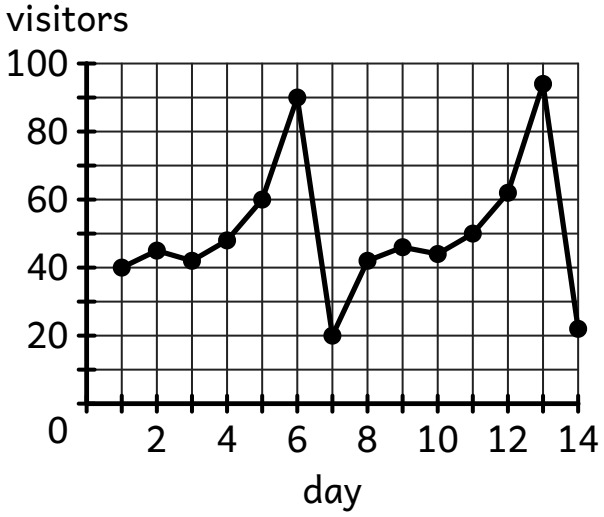
A The trend is a rise.

B There is no trend, because the numbers change.

C The trend is a fall.

Answer at the back of the book.

2. This graph shows the visitors to a library on each of 14 days. After how many days does the pattern repeat?



A 6 days

B 7 days

C 14 days

D 2 days

Answer at the back of the book.

3. Which time series is most likely to show a pattern that repeats?

A The number of visitors to an outdoor swimming pool each month for several years.

B The height of a tree each year as it grows.

C The distance a ball has rolled each second after a push.

Answer at the back of the book.

4. The rain in the four quarters of a year is 88, 52, 28, 72 mm. What is the mean rain for the year?

A 240 mm

B 62 mm

C 60 mm

Answer at the back of the book.

5. The yearly means of a time series are 40, 44, 48. If the trend carries on, what is the estimate for the next year?

   A 4

   B 56

   C 48

   D 52

   Answer at the back of the book.

6. What does the trend of a time series show?

   A The highest reading in the series.

   B The ups and downs that repeat.

   C The general direction of the readings over a long time.

   Answer at the back of the book.

7. **From an earlier chapter**

   A line of best fit has the rule y = 3x + 2. Estimate y when x = 9.

   Answer at the back of the book.

8. The four quarterly readings of one year are 12, 30, 22 and 8. What is the mean for the year?

   Answer at the back of the book.

KEEP IT

Copy this into your exercise book.

Time series: readings in order, time across. Trend: the general direction. Seasonal pattern: a rise and fall that repeats. The mean over a whole cycle shows the trend.

Draw: A zigzag line that rises slowly, with a dashed line through the middle and two matching peaks marked.

M4.7.8 CHAPTER 8

Telling the truth with data

> **A chart shows one tall bar and one short bar, but the two numbers are almost the same. Is the chart telling the truth?**

YOUR GUESS

Pick one before you read on. It is fine to be wrong.

- ☐ **A** Not always. A chart can cut off its axis, which makes a small difference look big.
- ☐ **B** Yes. If the bars are drawn from the numbers, the chart must be true.
- ☐ **C** Never. A chart cannot be trusted at all.

IF THIS FEELS HARD, READ THIS FIRST

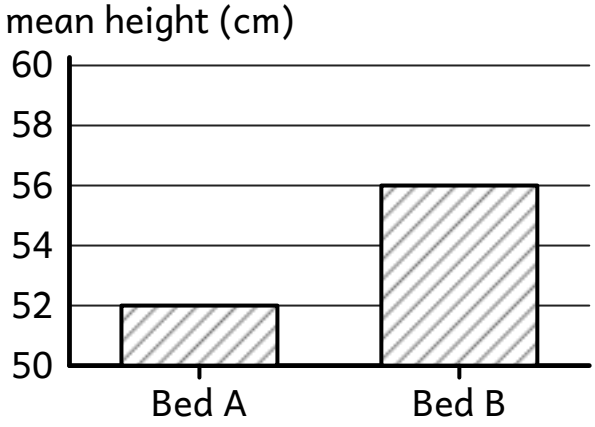
- Level 3, Statistics and chance, M3.7.5. If spotting a chart that does not start at zero looks new, read this first.
- Level 3, Statistics and chance, M3.7.4. If comparing two groups with an average and the range looks new, read this first.

BY THE END YOU CAN

- Spot a misleading graph or a misleading average
- Write a short report that compares two groups

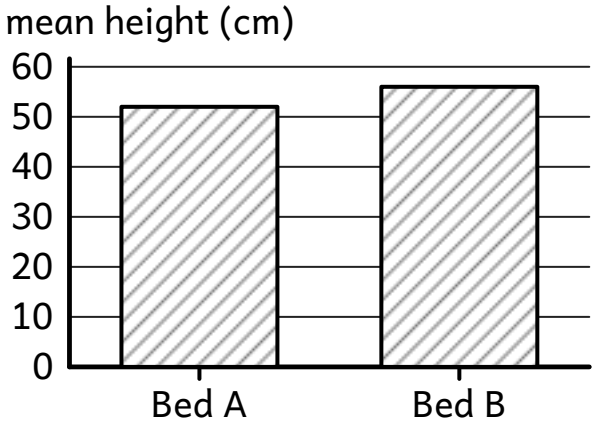
Numbers do not lie, but the way we show them can mislead. A reader who knows the common tricks is hard to fool. This chapter shows three: a cut axis, an average that is not typical, and a comparison that tells only half the story.

Amina measures the mean height of the bean plants in two beds. Bed A has a mean of 52 cm and Bed B has a mean of 56 cm. Here is one chart of these numbers.



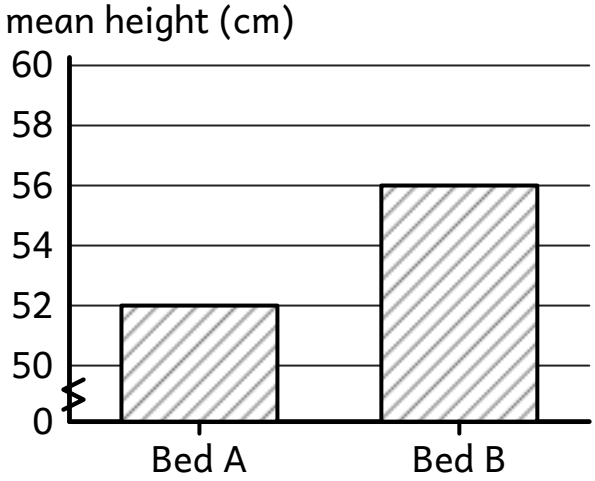
Bed B looks three times as tall as Bed A. The axis starts at 50.

The bar for Bed A is only 2 units tall above the floor. The bar for Bed B is 6 units tall. So Bed B looks three times as tall. But 56 is only a little more than 52. The difference is 4 cm, which is about 8% of 52.



The same numbers, with the axis from 0. The bars are nearly equal.

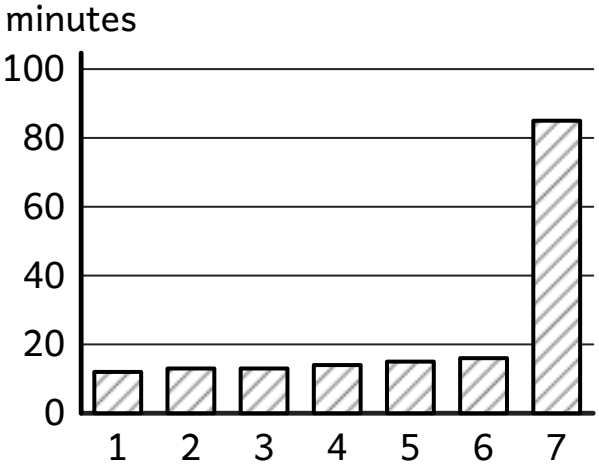
In a bar chart the length of a bar shows the size. A cut axis breaks that link. A fair bar chart starts at 0. If an axis has to be cut, a zigzag break mark on the axis warns the reader.



A cut axis with a zigzag break. The break says that this axis does not start at 0.

Other tricks are common. Two charts side by side may use different scales. A chart may show a short stretch of time, to hide a longer story. A bar chart with classes of different widths, drawn with height as the frequency, makes wide classes look big. Always ask: what is the scale, and what is left out?

Now a trick with an average. A bus company times 7 journeys in minutes: 12, 13, 13, 14, 15, 16 and 85. The 85 minute journey was held up by a flood. A leaflet says: "the average journey is 24 minutes".



One journey took 85 minutes. The other six took 16 minutes or less. The mean is 24 minutes.

WORKED EXAMPLE

Find the mean, the median and the mode of the journey times 12, 13, 13, 14, 15, 16 and 85.

1. Mean: 12 + 13 + 13 + 14 + 15 + 16 + 85 = 168. Then 168 ÷ 7 = 24.
2. Median: the values are in order. The 4th of the 7 values is 14.
3. Mode: 13 is the only value that appears twice.

Answer: The mean is 24 minutes, the median is 14 minutes and the mode is 13 minutes

A value far from the rest of the data is an **outlier**. The 85 is an outlier. It pulls the mean up. The median does not care how far away the largest value is.

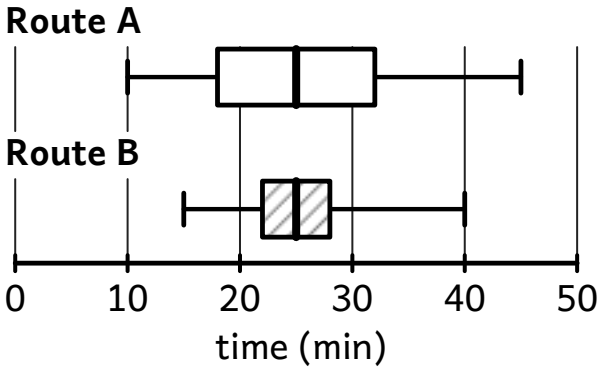
WORKED EXAMPLE

Which average best describes a typical journey: the mean or the median?

1. Count the journeys shorter than the mean of 24 minutes. There are 6.
2. So the mean is higher than six of the seven journeys. It is not typical.
3. The 85 minute journey is an outlier. The median, the 4th value, is 14 minutes.
4. Here, with an outlier in the data, the median describes a typical journey better.

Answer: The median, 14 minutes

Now the second half of the story. To compare two groups, one average is not enough. We need a number for the middle, the median, and a number for the spread, the IQR or the range. The two bus routes from the box plots chapter show this.



The same median. A much smaller box for Route B.

A short report has three parts. First, compare the middles with the medians. Second, compare the spreads with the IQRs or the ranges. Third, say what it means in plain words, with the numbers and the units.

WORKED EXAMPLE

Write a short report comparing the journey times on Route A and Route B.

1. The medians: both routes have a median of 25 minutes.
2. The IQRs: Route A has 32 − 18 = 14 minutes. Route B has 28 − 22 = 6 minutes.
3. The ranges: Route A has 45 − 10 = 35 minutes. Route B has 40 − 15 = 25 minutes.
4. Report: both routes have a median of 25 minutes. Route B is more consistent, with an IQR of 6 minutes against 14.

Answer: Both routes have a median of 25 minutes, but Route B is more consistent: its IQR is 6 minutes against 14

TRY IT

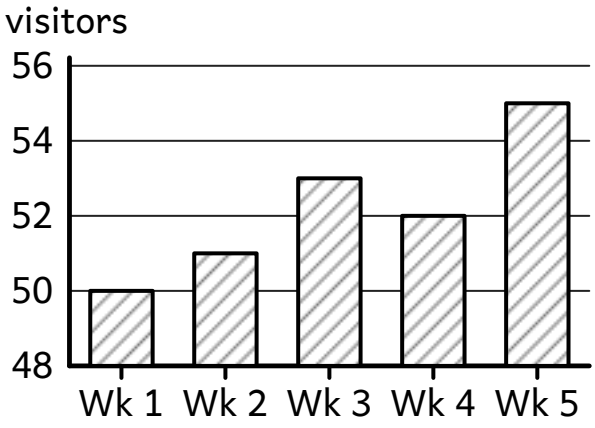
Make one chart tell two stories

You need: squared paper, a ruler, a pencil.

1. Take the counts of visitors in 5 weeks: 50, 51, 53, 52 and 55. Draw a bar chart with the axis from 0 to 60.
2. Draw the same bar chart again with the axis from 48 to 56. Make both charts the same height on the page.
3. Compare the two charts. Which one would you use to say "visitors are soaring"? Which is the fair one?

What you will see: The cut chart makes week 5 look more than three times week 1. The fair chart shows a gentle rise of 5 visitors in 4 weeks.

If you cannot try it, this is what happens.



A WRONG IDEA MANY PEOPLE HAVE

~~The mean is always the best average.~~

Why it feels right: The mean uses every value, so it seems the fairest.

The test that breaks it: In the journey times, the mean is 24 minutes, though six of the seven journeys took 16 minutes or less.

What is true: With an outlier, the median usually describes a typical value better. Say which average you used, and why.

A WRONG IDEA MANY PEOPLE HAVE

~~Any chart whose axis does not start at 0 is dishonest.~~

Why it feels right: A cut axis is a favourite trick, so it is tempting to blame it every time.

The test that breaks it: A line graph of the temperature of a room, from 18 °C to 24 °C, needs no 0. Readers look at the changes along the line, not at the size of a bar.

What is true: A cut axis misleads most in a bar chart, where length means size. A cut axis should be marked with a break, and the numbers on it read.

KEY WORDS

outlier
a value that is far from the rest of the data.

misleading
giving a wrong idea, even if the numbers are true.

BACK TO YOUR GUESS

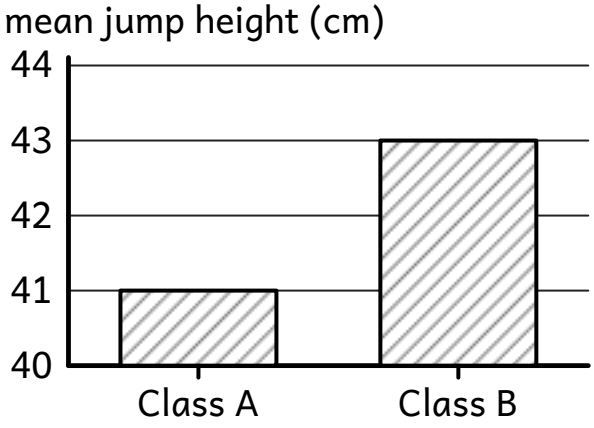
The answer is A: Not always. A chart can cut off its axis, which makes a small difference look big.

A chart can be drawn from correct numbers and still mislead. If the axis starts above zero, a small difference can look large. Always read the scale. Charts can be trusted when they are drawn fairly, and we can learn to check.

CHECK YOURSELF

Try each one before you look. A wrong answer tells you where to go next.

1. The chart shows the mean jump heights of two classes. What is wrong with it?



A Class B jumps three times as high as Class A.

B The axis is cut, so a small difference looks big.

C The mean is the wrong average for jump heights.

Answer at the back of the book.

2. Waiting times in minutes are 4, 5, 5, 6, 7, 9 and 48. Which average best describes a typical wait?

A The median, 6 minutes

B The mean, 12 minutes

C The range, 44 minutes

Answer at the back of the book.

3. A leaflet says: "Route A has the lower mean time, so it is more reliable." Why is this not enough?

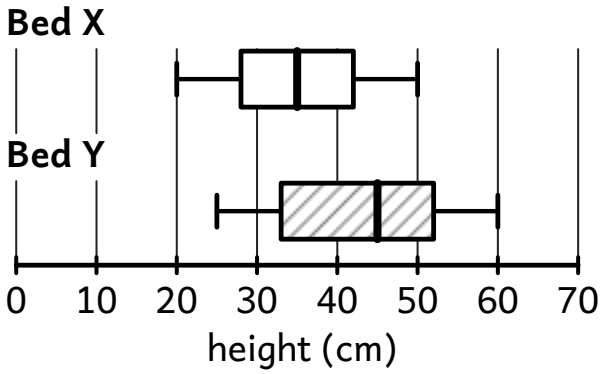
A The mean cannot be used for times.

B A lower mean means every journey on Route A is quicker.

C Reliable means little spread, so the IQRs or ranges must be compared too.

Answer at the back of the book.

4. The box plots show the heights of plants in two beds. Which statement is true?



A Bed X has the higher median.

B The two beds have the same spread.

C Bed Y has the higher median and the larger spread.

Answer at the back of the book.

5. Which of these can make a bar chart misleading?

A Labelling both axes with units.

B Starting the vertical axis above zero.

C Using the same scale for every bar.

Answer at the back of the book.

6. What is an outlier?

A A value that is far from the rest of the data.

B The most common value.

C The middle value of ordered data.

Answer at the back of the book.

7. **From an earlier chapter**

The four quarterly readings of a year are 14, 26, 18, 10. What is the mean for the year?

Answer at the back of the book.

8. Find the median of these 7 journey times, in minutes: 9, 10, 11, 12, 14, 15, 60.

Answer at the back of the book.

KEEP IT

Copy this into your exercise book.

Check the scale: a cut axis misleads. An outlier pulls the mean, so use the median. To compare groups, give the median and the IQR, with the units.

Draw: Two bar charts of the same two numbers, one from 0 and one cut, with the words "same numbers" between.

Key words, A to Z

Every key word in this book, what it means, and the chapter that teaches it.

biased
leaning one way, because some members are more likely to be chosen. M4.7.1

box plot
a picture of the smallest value, Q1, the median, Q3 and the largest value on one scale. M4.7.4

consistent
not spread out: the values stay close together. M4.7.4

correlation
the pattern between two measurements: positive, negative or none. M4.7.6

cumulative frequency
a running total of the frequencies. M4.7.3

estimate
an answer that is close, not exact, because some information is missing. M4.7.2

extrapolation
using a line beyond the range of the data, which is unreliable. M4.7.6

frequency density
frequency divided by class width. It is the height of a bar in a histogram. M4.7.5

frequency table
a table that lists each value and how many times it happens. M4.7.2

histogram
a chart for measured data in which the area of each bar is the frequency. M4.7.5

interquartile range
the upper quartile minus the lower quartile, the length of the middle half of the data. M4.7.3

line of best fit
a straight line drawn through the middle of the dots on a scatter graph. M4.7.6

midpoint
the number halfway between the two ends of a class. M4.7.2

misleading
giving a wrong idea, even if the numbers are true. M4.7.8

modal class
the class with the greatest frequency. M4.7.2

outlier
a value that is far from the rest of the data. M4.7.8

population
everything we want to know about. M4.7.1

quartile
a value that splits ordered data into quarters. Q1 has a quarter below it, Q3 a quarter above. M4.7.3

random sample
a sample where every member has the same chance of being chosen. M4.7.1

sample
a part of the population that we look at. M4.7.1

scatter graph
a graph with one dot for each pair of measurements. M4.7.6

seasonal pattern
a rise and fall that repeats after the same number of readings. M4.7.7

stratified sample
a sample that takes from each group in proportion to its size. M4.7.1

time series
readings taken at regular times, plotted in order. M4.7.7

trend
the general direction of a time series over a long time. M4.7.7

upper boundary
the top end of a class, where its cumulative frequency is plotted. M4.7.3

whisker
the line from the box to the smallest or the largest value. M4.7.4

Answers

M4.7.1. Samples and bias

1. **Choose names at random from a list of everyone in the town.**

   Every person on the list has the same chance of being chosen, so the sample is not biased. The other two ways make some people more likely to be asked.

   - **If you chose A:** Friends are often alike, in how they live and when they sleep. Some people in the town have no chance of being asked. Go to M4.7.1.
   - **If you chose B:** People up before dawn are more likely to sleep fewer hours. This sample leans one way. Go to M4.7.1.

2. **A big sample can still be biased if the way of choosing favours some people.**

   Bias comes from how members are chosen. Ravi chose from a place where active people are more likely to be, so more of them only repeats the lean.

   - **If you chose B:** Size does not remove bias. All 500 people came from a place where people who exercise are more likely to be. Go to M4.7.1.
   - **If you chose C:** A sample does not need to be half the population. A random sample of a few hundred can be fair for a very large population. What counts is how it is chosen. Go to M4.7.1.

3. **Every member of the population has the same chance of being chosen.**

   That is what random means in statistics. It does not depend on size or on looks.

   - **If you chose A:** Choosing by eye lets the picker's own ideas decide. Some members are then more likely to be chosen than others. Go to M4.7.1.
   - **If you chose C:** In everyday talk "random" can mean careless. Picking fast tends to favour members that are quick to reach. A random sample gives every member the same chance. Go to M4.7.1.
   - **If you chose D:** A large sample can be biased. What makes a sample random is the equal chance for every member. Go to M4.7.1.

4. **Plants 12, 38, 5 and 26**

   47 is above 40, so skip it. Then 12 and 38 are chosen. The second 12 is a repeat. The pair 05 is plant 5, and 26 is plant 26, which fills the sample.

   - **If you chose A:** There is no plant 47, because the bed has 40 plants. A number that is too big is skipped. A plant also cannot be chosen twice. Go to M4.7.1.
   - **If you chose B:** A plant cannot be in the sample twice. The second 12 is a repeat, so skip it. Go to M4.7.1.
   - **If you chose C:** The plants have two digit numbers, so read the digits two at a time, not one at a time. Go to M4.7.1.

5. **6**

   Group A is 48 ÷ 80 of the population, so it gives 48 ÷ 80 × 10 = 6 of the sample.

   - **If you chose A:** You took the same number from each group. Group A is bigger than group B, so it must give more. Go to M4.7.1.
   - **If you chose C:** That is the share for group B, 32 ÷ 80 × 10. Group A is the bigger group. Go to M4.7.1.

6. **Only people who listen and care enough to phone take part.**

   People choose themselves. Those who feel strongly are more likely to phone, and people who do not listen cannot. So some members are more likely than others to be in the sample.

   - **If you chose B:** A miscount is a mistake in counting. Bias is different. Even a perfect count would lean one way, because only some people choose to vote. Go to M4.7.1.
   - **If you chose C:** Listeners decide for themselves whether to vote. That is not random: people who care a lot are more likely to take part. Go to M4.7.1.

7. **10**

   Bean plants are 60 ÷ 90 of the garden. 60 ÷ 90 × 15 = 10. Check: peas give 30 ÷ 90 × 15 = 5, and 10 + 5 = 15.

8. **5**

   Maize is 25 ÷ 100 of the field. 25 ÷ 100 × 20 = 5. Check: beans 8 and peas 7, and 8 + 7 + 5 = 20.

M4.7.2. Averages from frequency tables

1. **2.25**

   1 × 5 + 2 × 8 + 3 × 4 + 4 × 3 = 5 + 16 + 12 + 12 = 45 seeds. There are 5 + 8 + 4 + 3 = 20 pods. 45 ÷ 20 = 2.25.

   - **If you chose A:** You found the mean of the numbers 1, 2, 3 and 4, and ignored how many pods have each. Multiply each x by its f first. Go to M4.7.2.
   - **If you chose B:** 45 is the total number of seeds, Σfx. The mean is the total divided by the number of pods. Go to M4.7.2.
   - **If you chose C:** You found the mean of the frequencies 5, 8, 4 and 3. The mean wanted is of the seeds in a pod, not of the numbers of pods. Go to M4.7.2.

2. **22 cm**

   The midpoints are 5, 15, 25 and 35. 5 × 3 + 15 × 9 + 25 × 12 + 35 × 6 = 15 + 135 + 300 + 210 = 660. Then 660 ÷ 30 = 22.

   - **If you chose A:** You found the mean of the midpoints 5, 15, 25 and 35, and ignored the frequencies. More plants are in the middle classes, so the mean is higher. Go to M4.7.2.
   - **If you chose C:** 660 is the total Σfx. The mean is the total divided by the number of plants. Go to M4.7.2.
   - **If you chose D:** You used the lowest value of each class, not the midpoint. The midpoint stands for the whole class. Go to M4.7.2.

3. **10 ≤ t < 20**

   There are 45 values, so the median is the (45 + 1) ÷ 2 = 23rd. The first class holds 12 values. Adding the second gives 12 + 20 = 32, so the 23rd is in the second class.

   - **If you chose B:** That is the middle class by position, the third of five. The median is the middle value, and 12 + 20 = 32 values already lie below 20. Go to M4.7.2.
   - **If you chose C:** This class holds only the first 12 values. The median is the 23rd, so we must go further up. Go to M4.7.2.

4. **50**

   The midpoint is (40 + 60) ÷ 2 = 50.

   - **If you chose A:** That is the lower end of the class. The midpoint is halfway between the two ends. Go to M4.7.2.
   - **If you chose B:** That is the width of the class, 60 − 40. The midpoint is a place on the scale, halfway along it. Go to M4.7.2.
   - **If you chose D:** That is the upper end of the class. The midpoint is halfway between 40 and 60. Go to M4.7.2.

5. **We know only which class each value is in, so we use the midpoint as its value.**

   The exact values are hidden by the classes. The midpoint is our best single guess for each class.

   - **If you chose A:** Grouping still counts every value. It hides where each value lies inside its class. Go to M4.7.2.
   - **If you chose C:** Some values are below the midpoint and some above. The errors can go either way and often cancel. Go to M4.7.2.

6. **The total of all the values in that row.**

   If f values are each equal to x, they add up to f × x. The mean is then the sum of these totals divided by Σf.

   - **If you chose A:** The number of values in the row is f, the frequency. Multiplying by x gives their total. Go to M4.7.2.
   - **If you chose B:** Every value in the row is x, so the mean of the row is just x. f × x is their total. Go to M4.7.2.

7. **6**

   Herbs are 30 ÷ 80 of the garden. 30 ÷ 80 × 16 = 6.

8. **2.85**

   1 × 3 + 2 × 5 + 3 × 6 + 4 × 4 + 5 × 2 = 3 + 10 + 18 + 16 + 10 = 57 peas. There are 3 + 5 + 6 + 4 + 2 = 20 pods. 57 ÷ 20 = 2.85.

9. **22 cm**

   The classes are 0 ≤ h < 10 and so on, so the midpoints are 5, 15, 25 and 35. 5 × 2 + 15 × 6 + 25 × 8 + 35 × 4 = 10 + 90 + 200 + 140 = 440. Then 440 ÷ 20 = 22.

M4.7.3. Cumulative frequency

1. **(30, 24)**

   The running total up to this class is 4 + 5 + 15 = 24 plants. They are all under 30 cm, so the point is at the upper boundary: (30, 24).

   - **If you chose A:** The running total belongs at the upper boundary of the class, not at its middle. All 24 plants are under 30 cm. Go to M4.7.3.
   - **If you chose B:** 15 is the frequency of this class. A cumulative frequency adds the classes before it: 4 + 5 + 15. Go to M4.7.3.
   - **If you chose C:** The running total of 24 counts plants under 30 cm, so it is plotted at 30, not at the lower boundary 20. Go to M4.7.3.

2. **34 cm**

   Half of 60 is 30. Read across from 30 to the graph, then down: the median is 34 cm.

   - **If you chose B:** 30 is the place to read at, half of 60. You must then go across to the graph and down to the height axis. Go to M4.7.3.
   - **If you chose C:** 24 is the cumulative frequency at a height of 30 cm. It is a number of plants, not a height. Read across from 30, then down. Go to M4.7.3.
   - **If you chose D:** 45 is the place where the upper quartile is read, 3 × 60 ÷ 4. The median is read at half of 60. Go to M4.7.3.

3. **19**

   IQR = Q3 − Q1 = 41 − 22 = 19.

   - **If you chose A:** You added the two quartiles. The interquartile range is the difference, Q3 − Q1. Go to M4.7.3.
   - **If you chose B:** 41 is the upper quartile itself. The interquartile range is how far Q3 is above Q1. Go to M4.7.3.

4. **It leaves out the lowest and highest quarters, so one odd value cannot change it much.**

   The range depends on the two most extreme values. The IQR looks at the middle half only.

   - **If you chose A:** The IQR uses only the middle half of the data. That is the reason one odd value cannot change it much. Go to M4.7.3.
   - **If you chose C:** The IQR is part of the range, so it is never bigger. The middle half cannot be wider than all of the data. Go to M4.7.3.

5. **39**

   Read up from 40 on the height axis. The graph is at 39 plants.

   - **If you chose A:** 15 is the frequency of the class $30 \leq h < 40$. Plants under 40 cm include the lower classes too. Go to M4.7.3.
   - **If you chose B:** 21 plants are 40 cm or more, since 60 − 39 = 21. The question asks how many are shorter. Go to M4.7.3.
   - **If you chose D:** 24 is the reading at 30 cm. The question asks about 40 cm. Go to M4.7.3.

6. **60**

   The upper quartile is three quarters of the way up: 3 × 80 ÷ 4 = 60.

   - **If you chose A:** You read the lower quartile, a quarter of 80. The upper quartile is three quarters of the way up. Go to M4.7.3.
   - **If you chose B:** You read the median, half of 80. The upper quartile is higher up. Go to M4.7.3.
   - **If you chose C:** You took 75% of 100. The quartile place depends on the number of values, so it is three quarters of 80. Go to M4.7.3.

7. **34.5 minutes**

   The midpoints are 5, 15, 25, 35, 45 and 55. 5 × 4 + 15 × 7 + 25 × 18 + 35 × 22 + 45 × 18 + 55 × 11 = 20 + 105 + 450 + 770 + 810 + 605 = 2 760. Then 2 760 ÷ 80 = 34.5.

8. **21**

   IQR = Q3 − Q1 = 52 − 31 = 21.

9. **18**

   42 values are under 40 cm. The rest are 40 cm or more: 60 − 42 = 18.

M4.7.4. Box plots

1. **5**

   The median is the 4th value, 9. Leave it out. The lower half is 3, 5, 8. Its middle value is 5.

   - **If you chose A:** You kept the middle value, 9, in the lower half, so the lower half was 3, 5, 8, 9. With 7 values the middle value is left out of both halves. Go to M4.7.4.
   - **If you chose B:** 3 is the smallest value. The lower quartile is the median of the lower half. Go to M4.7.4.
   - **If you chose D:** You took the 8, the value just below the median. The lower quartile is the middle of the lower half: the lower half is 3, 5, 8, so it is 5. Go to M4.7.4.

2. **35 minutes**

   The thick line marks the median. It is at 35.

   - **If you chose A:** 30 is where the box starts, the lower quartile. The median is the thick line. Go to M4.7.4.
   - **If you chose C:** 40 is the middle of the box. The median is the thick line, and it need not be in the middle of the box. Go to M4.7.4.
   - **If you chose D:** 50 is where the box ends, the upper quartile. The median is the thick line. Go to M4.7.4.

3. **20 minutes**

   The box runs from 30 to 50. IQR = Q3 − Q1 = 50 − 30 = 20.

   - **If you chose A:** 45 is the range, from the end of one whisker to the other. The IQR is the length of the box. Go to M4.7.4.
   - **If you chose B:** 35 is the median, *a* place on the scale. The IQR is *a* length: Q3 − Q1. Go to M4.7.4.
   - **If you chose C:** 50 is Q3, *a* place on the scale. The IQR is how far Q3 is above Q1. Go to M4.7.4.

4. **Route B is more consistent, because its box is shorter.**

   The medians are both 25 minutes. Route B has the smaller IQR, so its journey times are closer together.

   - **If you chose B:** The medians are equal, but the boxes are different lengths. The two routes have different spreads. Go to M4.7.4.
   - **If you chose C:** A longer box means more spread, not quicker. Both routes have the same median time. Go to M4.7.4.

5. **About half**

   A quarter of the values lie below Q1 and a quarter above Q3. That leaves about half between them, in the box.

   - **If you chose A:** A quarter of the values lie in each whisker, and in each half of the box. The two halves of the box add up to more. Go to M4.7.4.
   - **If you chose C:** The box is only from Q1 to Q3. A quarter of the values lie below Q1 and a quarter above Q3. Go to M4.7.4.

6. **The top quarter of the values is more spread out than the bottom quarter.**

   The same quarter of the values lies along a longer stretch of the scale, so those values are further apart.

   - **If you chose A:** Each whisker holds about a quarter of the values, whatever its length. Go to M4.7.4.
   - **If you chose C:** Half the values lie above the median and half below, whatever the whiskers look like. A long whisker shows spread, not more values. Go to M4.7.4.

7. **30**

   The lower quartile is a quarter of the way up: 120 ÷ 4 = 30.

8. **12**

   The median is the 6th value, 13. The lower half is 4, 6, 7, 10, 12, so Q1 = 7. The upper half is 15, 18, 19, 22, 25, so Q3 = 19. IQR = 19 − 7 = 12.

M4.7.5. Histograms with unequal widths

1. **1.8 plants per cm**

   The width is 40 − 20 = 20 cm. Frequency density = 36 ÷ 20 = 1.8.

   - **If you chose A:** You divided the width by the frequency, 20 ÷ 36. Frequency density is the frequency divided by the width. Go to M4.7.5.
   - **If you chose B:** You multiplied 36 by 20. The area is width × height, so to get the height we divide the frequency by the width. Go to M4.7.5.
   - **If you chose D:** 36 is the frequency itself. The density is the frequency shared out over the 20 cm of the class. Go to M4.7.5.

2. **15**

   The area of the bar is its frequency: 6 × 2.5 = 15.

   - **If you chose A:** You added the width and the height. The frequency is the area of the bar, which is width × height. Go to M4.7.5.
   - **If you chose B:** You divided the width by the height. The frequency is the area, so multiply. Go to M4.7.5.
   - **If you chose C:** You divided the height by the width, 2.5 ÷ 6. The frequency is width × height. Go to M4.7.5.

3. **10 ≤ h < 30**

   The areas are 10 × 1.5 = 15, 20 × 1.2 = 24 and 10 × 0.6 = 6. The middle bar has the biggest area, 24 plants.

   - **If you chose A:** Its bar is the tallest but only 10 cm wide. Area counts: 10 × 1.5 = 15 plants. Another bar has a bigger area. Go to M4.7.5.
   - **If you chose C:** The area of each bar is its frequency, so the chart does tell us: 15, 24 and 6. Go to M4.7.5.

4. **The frequency density: the frequency divided by the class width.**

   The height of a bar is frequency ÷ width, so the area of the bar is the frequency.

   - **If you chose A:** That is a bar chart. In a histogram with unequal widths, the frequency is the area of the bar, not its height. Go to M4.7.5.
   - **If you chose B:** The width of a class is shown along the horizontal axis, by the width of its bar. Go to M4.7.5.

5. **5**

   The piece from 50 to 60 is 10 cm wide. The area is 10 × 0.5 = 5 plants.

   - **If you chose A:** 15 plants are in the whole class from 50 to 80. The question asks for the part from 50 to 60. Go to M4.7.5.
   - **If you chose B:** 0.5 is the height of the bar, the density. The number of plants is the area of the part. Go to M4.7.5.
   - **If you chose C:** You took half of the class. From 50 to 60 is 10 cm out of 30 cm, which is a third of it. Go to M4.7.5.

6. **When the classes have different widths.**

   With unequal widths, bars as tall as the frequency would make wide classes look too big. Using the density makes the area show the frequency.

   - **If you chose B:** The number of values does not matter. What matters is whether the widths of the classes are all equal. Go to M4.7.5.
   - **If you chose C:** With equal widths the height can be the frequency, and the picture is still fair, because every bar has the same width. Go to M4.7.5.

7. **25**

   There are 7 values, so the median is 20, the 4th. Leave it out. The upper half is 23, 25, 31. Its middle value is 25.

8. **1.4**

   Frequency density = frequency ÷ width = 14 ÷ 10 = 1.4.

9. **30**

   The frequency is the area: 20 × 1.5 = 30.

M4.7.6. Scatter graphs and correlation

1. **Negative correlation**

   The dots fall from left to right. As the climb gets bigger the temperature gets lower.

   - **If you chose A:** In positive correlation the dots rise from left to right. Here they fall: as the climb rises, the temperature falls. Go to M4.7.6.
   - **If you chose B:** The dots follow a clear pattern, falling from left to right. No correlation would be a scatter with no direction. Go to M4.7.6.

2. **No correlation**

   The dots show no pattern. Knowing the day of the month a person was born tells us nothing about the jump height.

   - **If you chose B:** The dots do not rise from left to right. They are scattered with no direction. Go to M4.7.6.
   - **If you chose C:** The dots do not fall from left to right either. They are scattered with no direction. Go to M4.7.6.

3. **166 cm**

   y = x + 1 with x = 165 gives y = 165 + 1 = 166.

   - **If you chose A:** You used the height as the arm span. The line is y = x + 1, so add 1. Go to M4.7.6.
   - **If you chose B:** You took 1 away. The rule is y = x + 1, so the arm span is 1 more than the height. Go to M4.7.6.
   - **If you chose C:** 163 cm is the arm span of the friend in the table who is 165 cm tall. This is a new friend, so use the line. Go to M4.7.6.

4. **Rainy weather probably makes both rise.**

   On rainy days more people buy umbrellas, and wet roads and poor sight bring more accidents. Rain is a hidden third factor. Correlation does not prove that one causes the other.

   - **If you chose A:** Umbrellas do not make crashes happen. Two things can rise together without one causing the other. Go to M4.7.6.
   - **If you chose B:** There is no evidence for that either. A hidden third factor is a better explanation. Go to M4.7.6.

5. **190 cm**

   190 cm is beyond the tallest friend in the data, 176 cm. Extending the line there is extrapolation, and the pattern may not carry on.

   - **If you chose A:** 150 cm is inside the range of the data, so an estimate there uses the pattern the data show. Go to M4.7.6.
   - **If you chose C:** 160 cm is in the middle of the data, where the line is most trustworthy. Go to M4.7.6.

6. **The dots lie close to a straight line.**

   The closer the dots are to a line, the stronger the correlation.

   - **If you chose A:** Steepness depends on the scales of the axes. Strength is about how close the dots are to the line. Go to M4.7.6.
   - **If you chose C:** Many dots can be widely spread. Strength depends on how close the dots are to the line, not on how many there are. Go to M4.7.6.

7. **12**

   The frequency is the area of the bar: 5 × 2.4 = 12.

8. **17**

   y = 2 × 7 + 3 = 14 + 3 = 17.

M4.7.7. Data over time

1. **The trend is a fall.**

   The visitors drop from 90 in week 1 to 55 in week 8. The general direction is down.

   - **If you chose A:** The dots start high and end lower. A rising trend would climb from left to right. Go to M4.7.7.
   - **If you chose B:** Numbers always change a little. The trend is the general direction, and here the dots fall steadily. Go to M4.7.7.

2. **7 days**

   The peaks are on days 6 and 13, which are 7 days apart. The low points on days 7 and 14 are also 7 days apart.

   - **If you chose A:** 6 is the day of the first peak. The question asks how many days pass before the peak comes again. Go to M4.7.7.
   - **If you chose C:** 14 is the number of days in the whole series. The pattern shows twice in it, so it repeats after half of that. Go to M4.7.7.
   - **If you chose D:** The readings do change from day to day, but the repeating shape is the whole rise and fall, which takes longer. Go to M4.7.7.

3. **The number of visitors to an outdoor swimming pool each month for several years.**

   Pool visits follow the seasons. The same months are busy each year, so the pattern repeats every 12 months.

   - **If you chose B:** A growing tree has an upward trend, but its height does not go up and down in a repeating way. Go to M4.7.7.
   - **If you chose C:** The ball keeps rolling further, then slows. It does not rise and fall in a repeating way. Go to M4.7.7.

4. **60 mm**

   88 + 52 + 28 + 72 = 240. Then 240 ÷ 4 = 60.

   - **If you chose A:** 240 is the total of the four readings. The mean is the total shared out over the 4 quarters. Go to M4.7.7.
   - **If you chose B:** 62 is the median, halfway between the two middle readings. The mean uses all four readings added up. Go to M4.7.7.

5. **52**

   The means rise by 4 each year. The next is 48 + 4 = 52.

   - **If you chose A:** 4 is how much the mean rises each year. The next mean is that rise added to the last mean. Go to M4.7.7.
   - **If you chose B:** You added the rise twice. The next year is only one step on: 48 + 4. Go to M4.7.7.
   - **If you chose C:** That is the last mean. The trend says the mean keeps rising, so the next one is higher. Go to M4.7.7.

6. **The general direction of the readings over a long time.**

   The trend says whether the readings are mostly rising, falling or staying level.

   - **If you chose A:** The highest reading is one value. The trend is about the whole series. Go to M4.7.7.
   - **If you chose B:** That is the seasonal pattern. The trend is the general direction under those ups and downs. Go to M4.7.7.

7. **29**

   y = 3 × 9 + 2 = 27 + 2 = 29.

8. **18**

   12 + 30 + 22 + 8 = 72. Then 72 ÷ 4 = 18.

M4.7.8. Telling the truth with data

1. **The axis is cut, so a small difference looks big.**

   The bars start at 40 and not at 0. Class B is only a little higher than Class A, but its bar looks three times as tall.

   - **If you chose A:** That is what the chart seems to say. Read the scale: 43 cm is only 2 cm more than 41 cm. Go to M4.7.8.
   - **If you chose C:** The mean is a fine average for jump heights. What misleads here is the scale on the vertical axis. Go to M4.7.8.

2. **The median, 6 minutes**

   The median is the 4th of the 7 values, 6. It is not pulled up by the outlier, 48.

   - **If you chose B:** The one wait of 48 minutes pulls the mean up. Six of the seven waits are under 12 minutes. Go to M4.7.8.
   - **If you chose C:** The range measures spread, not the middle. It is the largest value minus the smallest, 48 − 4. Go to M4.7.8.

3. **Reliable means little spread, so the IQRs or ranges must be compared too.**

   A mean says nothing about spread. Two routes with the same mean can have very different spreads.

   - **If you chose A:** The mean can be used for times. The problem is that one number cannot show how much the times vary. Go to M4.7.8.
   - **If you chose B:** A mean is one number. Some journeys on Route A can still be slow if its times are spread out. Go to M4.7.8.

4. **Bed Y has the higher median and the larger spread.**

   Bed Y has the median further right, and its box is longer. Its IQR is 52 − 33 = 19 against 42 − 28 = 14.

   - **If you chose A:** The thick line of Bed X is further left than that of Bed Y. A higher median is further right on the scale. Go to M4.7.8.
   - **If you chose B:** The box of Bed Y is longer than the box of Bed X, so its IQR is bigger. The spreads are not the same. Go to M4.7.8.

5. **Starting the vertical axis above zero.**

   A cut axis makes small differences look large, because bar length no longer shows the size.

   - **If you chose A:** Labels and units help the reader. They do not mislead. Go to M4.7.8.
   - **If you chose C:** A common scale is what makes bars fair to compare. Go to M4.7.8.

6. **A value that is far from the rest of the data.**

   An outlier stands apart from the other values. It can pull a mean away from the typical value.

   - **If you chose B:** That is the mode. An outlier is not common. It stands apart from the others. Go to M4.7.8.
   - **If you chose C:** That is the median. An outlier is far from the middle. Go to M4.7.8.

7. **17**

   14 + 26 + 18 + 10 = 68. Then 68 ÷ 4 = 17.

8. **12**

   The values are in order. The 4th of the 7 values is 12. The outlier, 60, does not change the median.

About this book

For learners who can find the mean, median and range of a short list and read a bar chart. In many schools this fits teenagers of about 14 to 16, and adults who want to read statistics in the news with more care can start here too. Each chapter takes one or two sessions. A learner can work alone: every exercise has its answer and the reason for it. Each chapter also has a short page for the adult, with three questions to ask.

Written and drawn by attnom. Edition 1.0, 2026-11-12.

This book is free. You may read it, print it, copy it and share it with anyone.

The newest copy is at https://attnom.com/academy/averages-spread-and-correlation.