Scatter graphs and correlation

S6 · Statistics · revision notes

Scatter graphs show whether two variables are related. You need to plot points, describe the correlation, draw a line of best fit, use it to make estimates, spot outliers and explain the limits of your conclusions. These questions come up on both tiers, usually for 1 to 4 marks.

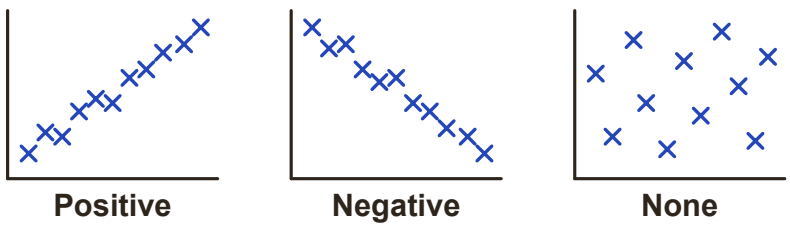
Key facts

- Positive correlation: as one increases, the other increases
- No correlation: no clear pattern
- Interpolation (inside the range of the data): fairly reliable
- Outlier: a point that does not fit the general pattern
- Grade 6+ only Gradient of the line of best fit = change in y for each increase of 1 in x
- Negative correlation: as one increases, the other decreases
- Line of best fit: one straight ruled line following the trend, ignoring outliers
- Extrapolation (outside the range of the data): unreliable
- Correlation does not prove causation

Notes

Scatter graphs and correlation

- A scatter graph shows two sets of data about the same items (bivariate data), e.g. the age and the value of some cars. Each item is plotted as one point.
- **Positive correlation**: as one variable increases, the other tends to increase.
- **Negative correlation**: as one variable increases, the other tends to decrease.
- **No correlation**: there is no clear pattern.



The closer the points are to a straight line, the stronger the correlation.

- The closer the points are to a straight line, the **stronger** the correlation. Widely scattered points show **weak** correlation.

Lines of best fit

- A line of best fit is a single straight line, drawn with a ruler, that follows the trend of the points.
- Aim for about the same number of points on each side, spread along its length. It does not have to go through the origin or through any particular point.
- Ignore any outlier when you draw it, and only draw one when there is correlation.

Making estimates

- To estimate, go from the given value to the line of best fit, then across to the other axis. Use the line, not the nearest point.
- **Interpolation** means estimating inside the range of the data. This is fairly reliable.
- **Extrapolation** means estimating outside the range of the data. This is unreliable, because the pattern may not continue.
- The **gradient** of the line of best fit tells you how much y changes, on average, for each increase of 1 in x, e.g. a gradient of 5 could mean about 5 more marks for each extra hour of revision. Grade 6+ only

Outliers and causation

- An **outlier** is a point that does not fit the pattern of the others. It may be an error or an unusual item.
- Correlation does not prove that one variable causes the other. Both may depend on a third factor: ice cream sales and cases of sunburn both rise in hot, sunny weather.

How to answer each type of question

M = method mark, A = answer mark, B = a mark on its own, C = a communication mark.

Describe the correlation and the relationship 1 to 2 marks Grade 3

1. Look at the overall direction of the points from left to right.
2. Name the correlation: positive, negative or none.
3. For 'describe the relationship', use the names of the variables, e.g. 'the older the car, the lower its value'.

Example. The table shows the ages and the values of 8 cars.

Age (years)	1	2	3	4	5	6	7	8
Value (£1000s)	18	16	15	12	11	9	8	5

A scatter graph is drawn of these data.
(a) What type of correlation does it show?
(b) Describe the relationship between the age and the value of these cars.

Model answer.

(a) Negative correlation (B1)
(b) The older the car, the lower its value (C1)

Draw a line of best fit and make an estimate

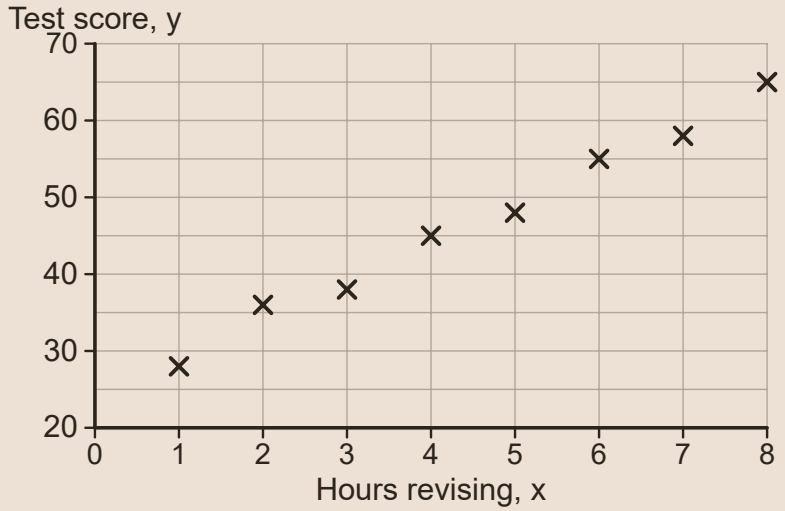
2 marks | **Grade 5**

1. Draw one straight line through the middle of the points, ignoring any outlier.
2. Go up from the given value to your line, then across to the other axis.
3. Read the value carefully and give the units, if there are any.

Example. The table shows the number of hours, x, that 8 students spent revising and their scores, y, in a test marked out of 80.

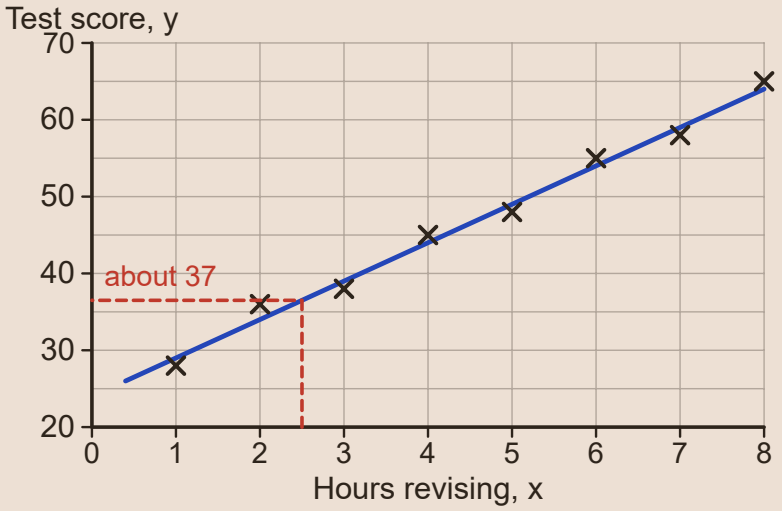
x	1	2	3	4	5	6	7	8
y	28	36	38	45	48	55	58	65

These data are plotted on the scatter graph below.



Draw a line of best fit and use it to estimate the score of a student who revised for 2.5 hours.

Model answer.



A single straight line of best fit drawn through the points (M1)

About 37 marks (A1, accept answers from about 34 to 40 that match your line)

Explain why an estimate may not be reliable

1 mark | Grade 5

1. Check whether the value is inside or outside the range of the data.
2. If it is outside, say that it is outside the range of the data, so the trend may not continue.
3. If you can, add a reason from the context.

Example. Sam uses the line of best fit for the revision data to estimate the score of a student who revised for 15 hours.
Explain why Sam's estimate may not be reliable.

Model answer. 15 hours is outside the range of the data (1 to 8 hours), so the trend may not continue.
For example, the line gives close to 100 marks, which is impossible in a test marked out of 80 (C1)

Explain whether one variable causes the other

1 mark | Grade 5

1. Say that correlation does not prove causation.
2. Suggest a third factor that could affect both variables.

Example. At a beach, there is strong positive correlation between the number of ice creams sold each day and the number of people treated for sunburn each day.
Kai says, “Eating ice cream causes sunburn.”
Is Kai correct? Give a reason for your answer.

Model answer. No. Correlation does not prove causation: both numbers are likely to go up on hot, sunny days (C1)

Use and interpret the equation of a line of best fit

1 to 3 marks | Grade 6

1. The gradient is the number multiplying x: the change in y for each increase of 1 in x.
2. Describe it in context, using both variables.
3. To estimate, substitute the value of x into the equation.

Example. The line of best fit for the revision data has equation $y = 5x + 24$, where x is the number of hours spent revising and y is the test score.
(a) Explain what the gradient of the line tells you.
(b) Use the equation to estimate the score of a student who revised for 5.5 hours.

Model answer.

(a) For each extra hour of revision, the score increases by about 5 marks (C1)
(b) 5 × 5.5 + 24 (M1)
= 51.5 marks (A1)

Don’t lose marks

- Joining the points dot to dot, or drawing a line of best fit that is not straight.

- Forcing the line of best fit through the origin.
- Including the outlier when you draw the line of best fit.
- Answering 'describe the relationship' with only 'positive'. Use the context.
- Saying that correlation proves one variable causes the other.
- Misreading the scale: check what each small square is worth on each axis.

More tips

Memory tricks

- Positive slopes uphill from left to right; negative slopes downhill.
- Inter = inside the data (safe); extra = beyond the data (risky).
- Line check: about half the points should be above your line and half below, all the way along it.
- When you read an estimate, draw the lines: up to the line of best fit, then across. It shows your method and stops misreadings.

Exam technique

- Use a ruler, and draw the line of best fit across the whole spread of the points.
- Answers read from a graph are accepted within a range, so draw carefully and show your reading lines.
- For reliability, the key phrases are 'outside the range of the data' and 'the trend may not continue'.
- For causation, write 'correlation does not mean causation' and suggest a third factor.

What each grade needs

What you need to be able to do, from the first marks up to the top grade.

Grade 2 — **Plot points on a scatter graph**
Plot each pair of values as a cross, reading both scales carefully.

Grade 3 — **Name the type of correlation**
Positive, negative or no correlation, and whether it is strong or weak.

Grade 3 — **Describe the relationship in context**
Say how one variable changes as the other increases, using the names of the variables.

Grade 4 — **Draw a line of best fit**
Draw one straight ruled line that follows the trend, with the points spread evenly on either side.

Grade 4 — **Identify and explain an outlier**
An outlier is a point that does not fit the general pattern; leave it out when you draw the line.

Grade 5 — **Make estimates using a line of best fit**
Read from the line inside the range of the data; estimates outside it (extrapolation) are unreliable.

Grade 5 — **Explain why correlation does not prove causation**
Two variables can be correlated because both depend on something else.

Grade 6 — **Interpret the gradient of a line of best fit**
The gradient is the change in y for each increase of 1 in x, described in context.

Free notes, Full packs, practice papers and mock exams at chhetriacademy.com. Written for this site; not taken from past papers. Stuck? 1-to-1 tutoring with a free 30-minute trial lesson.