Scatter graphs & correlation (8–9)

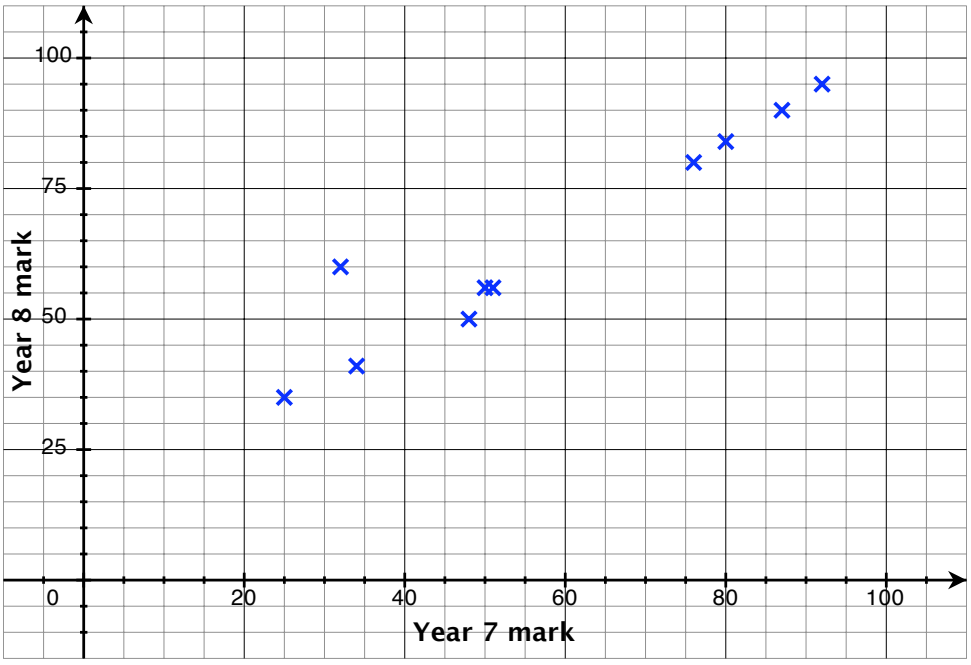
Contents

1 Scatter graphs

If we only have one type of data to show, we can use a bar chart, histogram, pie chart etc. A scatter graph is used when we wish to compare **two** types of data.
E.g. Imagine a teacher wanted to see how her class had done in the year 7 summer exam and in the year 8 summer exam. The following table shows the mark that each student got in year 7 and in year 8.

Year 7	50	80	76	34	51	48	25	92	87	32
Year 8	56	84	80	41	56	50	35	95	90	60

We can plot these on a **scattergraph** as below:



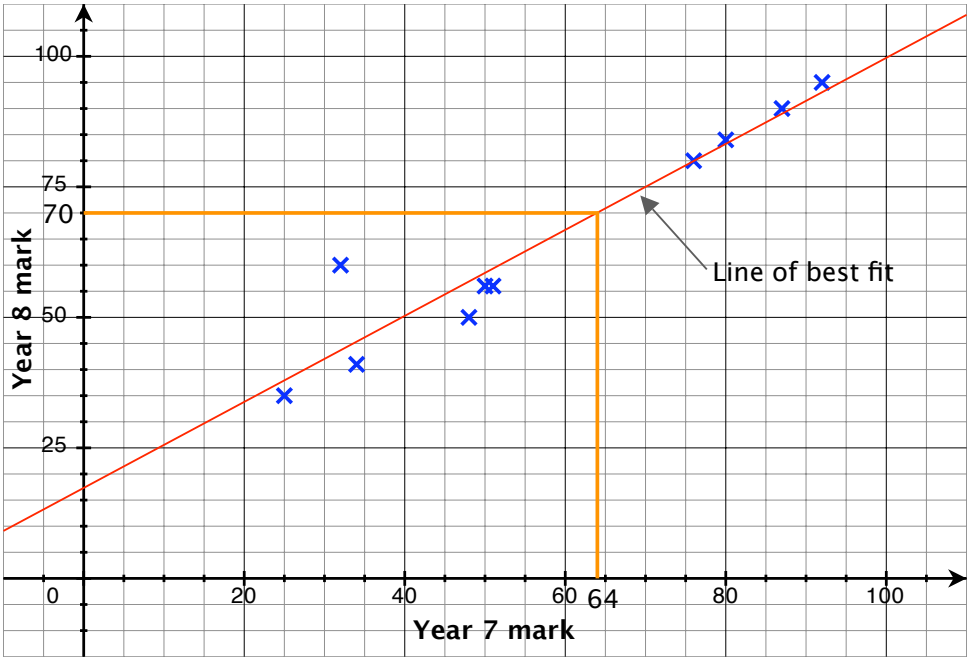
Scatter graph to show year 7 against year 8 summer exam marks

The graph shows us that, in general, the better you did at the end of year 7 the better you went on to do at the end of year 8. i.e. a student scoring low in year 7 would probably score low in

year 8, a student scoring high in year 7 would probably score high in year 8.

We use the words "in general" since the relationship between year 7 and year 8 marks is not perfect. If it was, all the points would lie on a straight line. One student, for example, did not do very well in year 7 (32%) but went on to do quite well in year 8 (60%).

Since the points have a general upwards trend, we can insert a **line of best fit**, which is a straight line that best follows the trend of the points (see the following diagram).



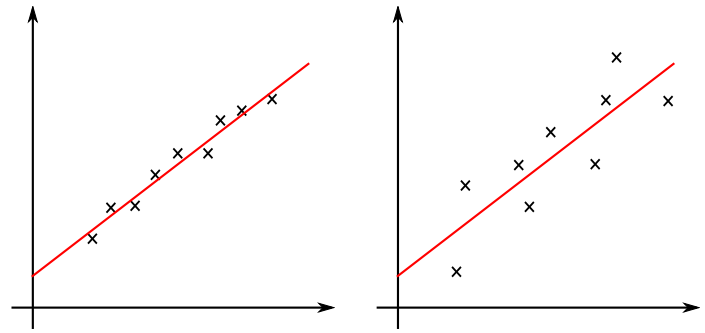
Scatter graph to show year 7 against year 8 summer exam marks

The line can be used to work out missing data e.g. imagine a new student joined in year 8 and scored 70%. What was their likely mark in year 7? Viewing the lines we have inserted into the diagram (orange lines), we see the student probably would have got around 64% in year 7.

2 Correlation

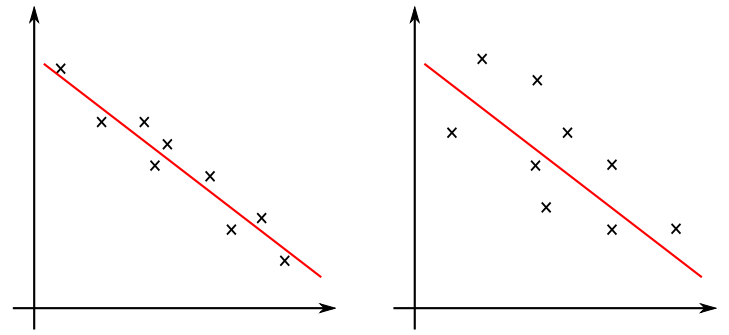
This is the mathematical word for "relationship".

Positive correlation. As one variable increases, so does the other:



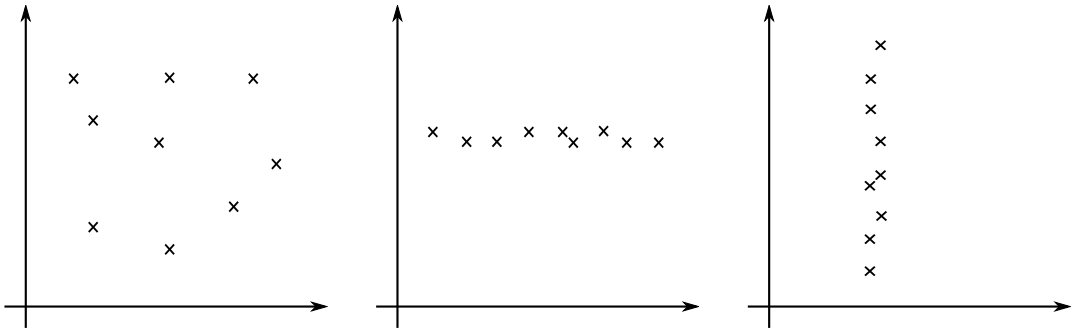
Strong positive correlation *Weak positive correlation*

Negative correlation. As one variable increases the other decreases and vice versa.



Strong negative correlation *Strong negative correlation*

No correlation. There is no relationship between the variables.



No correlation — DO NOT draw a line of best fit!

Notice how we only insert a line of best fit when there is some correlation (how can you put in a line that best follows the points if the points are completely random?).
What type of correlation would you expect between the following?

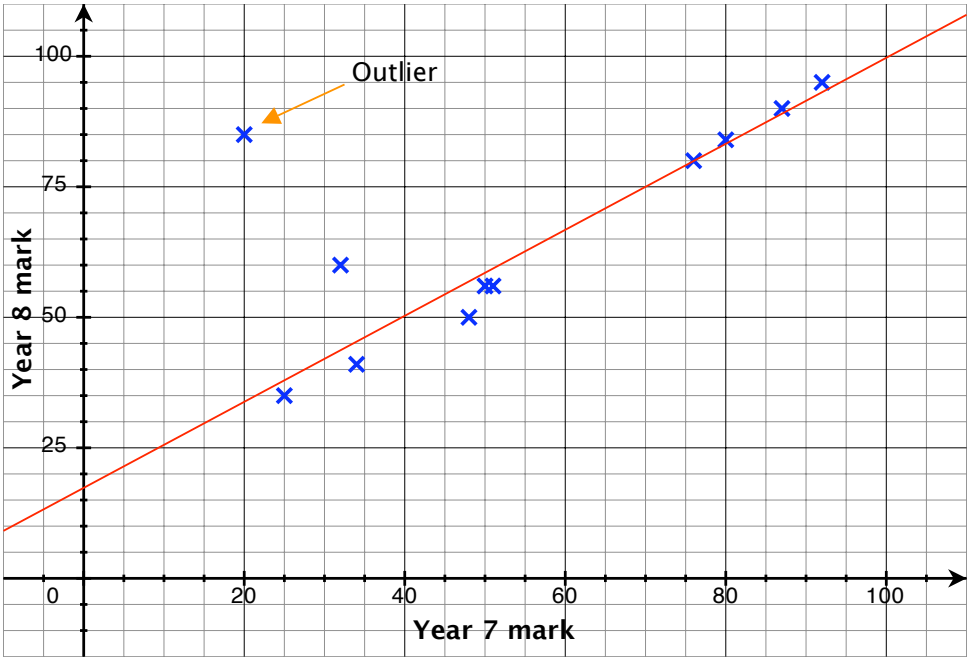
Height & Weight. Taller people are probably heavier than shorter ones, although you can get skinny tall people and rounder short people: *weak positive correlation*

Age & price of a car. Older cars probably cost less and new cars cost more: *strong negative correlation*

Pocket money and height. It is probably not true that taller people get more pocket money (although you could argue that taller people are probably older so do get more): emphno correlation

3 Outliers

An outlier is a correct point but that is very different from the other points. For example, imagine that the scattergraph of test results looked like the graph below. You can see one point that is very far from the general trend of the rest of the points.



Scatter graph to show year 7 against year 8 summer exam marks

This is a student who did very poorly in year 7 (20%) but got an excellent mark in year 8 (85%). There may be a reason why this happened. For example she could have been very ill on the day of the year 7 test, or could have had a long time out of school in year 7.

4 Important points about scatter diagrams

- Never draw a scatter diagram with only a few points.
- Only add a line of best fit if there is an obvious general relationship,, i.e. you can see some correlation when you look at the graph.
- A line of best fit *does not* have to go through the origin
- When you insert a line of best fit by eye, insert it roughly half way through the points, getting an equal number on either side and as many on the line as possible
- Make sure you axes are the right way around e.g. in the example above, the mistakes you make in your driving lesson is dependent on the lessons you had, so mistakes go on the y-axis (it wouldnt make sense to say lessons depend on mistakes)