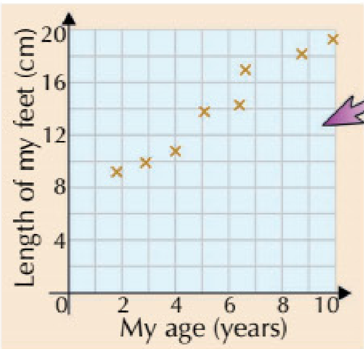
Correlation

There's a fair bit of fancy stats-speak in this section. Correlation is all about how closely two quantities are linked and linear regression is just a way to find the line of best fit. Not so scary now, eh...

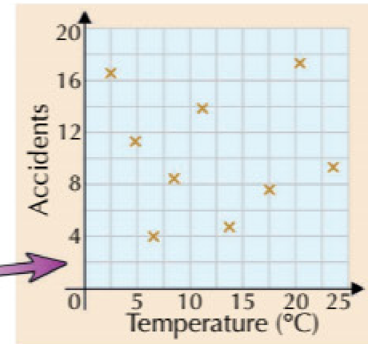
*Draw a **Scatter Diagram** to see **Patterns** in **Data***

1) Sometimes variables are measured in **pairs** — maybe because you want to find out **how closely** they're **linked**. Data made up of pairs of values (x, y) is known as **bivariate data** and can be plotted on a **scatter diagram**.



2) The variables 'my age' and 'length of my feet' seem linked — all the points lie **close** to a **line**. As I got older, my feet got bigger and bigger (though I stopped measuring when I was 10).

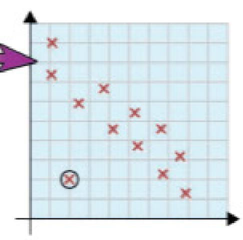
3) It's a lot harder to see any connection between the variables 'temperature' and 'number of accidents' — the data seems **scattered** pretty much everywhere.



Correlation** is a measure of **How Closely** variables are **Linked

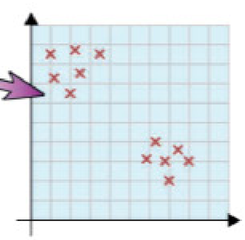
1) If, as one variable gets **bigger**, the other one also gets **bigger**, the scatter diagram might look like the age/length of feet graph above. The line of best fit would have a **positive gradient**. The two variables are **positively correlated** (or there's a positive correlation **between** them).

2) If one variable gets **smaller** as the other one gets **bigger**, then the scatter diagram might look like this one and the line of best fit would have a **negative gradient**. The two variables are **negatively correlated** (or there's a negative correlation **between** them). The circled point is an **outlier** — a point that doesn't fit the pattern of the rest of the data. Outliers can usually be **ignored** when drawing the line of best fit or describing the correlation — they can be **measurement errors** or just '**freak**' observations.



3) If the two variables **aren't** linked at all, you'd expect a **random** scattering of points (like in the temperature/accidents graph above). The variables **aren't correlated** (or there's **no correlation**).

4) Watch out for graphs that show distinct sections of the population like this one — the data will be in **separate clusters**. Here, you can describe both the **overall correlation** and the correlation in **each cluster** — so on this graph, there appears to be **negative correlation** overall, but **no correlation** within each cluster. Different clusters can also be shown on **separate graphs**, with each graph representing a different section of the population.



5) Correlation can also be described as '**strong**' or '**weak**'. The **stronger** the correlation is, the closer the points on the scatter diagram are to being in a **straight line**.

BUT you have to be **careful** when writing about two variables that are correlated — changes in one variable might **not cause** changes in the other. They could be linked by a **third factor**, or it could just be **coincidence**. The formal way of saying this is '**correlation** does not imply **causation**'.

*Decide which is the **Explanatory Variable** and which is the **Response***

The variable along the ***x*-axis** is the **explanatory** (or **independent**) variable
— it's the variable you can **control**, or the one that is **affecting** the other.

The variable up the ***y*-axis** is the **response** (or **dependent**) variable — it's the variable you think is **being affected**.

Example: Tasha wants to plot a scatter diagram to show the variables 'load on a lorry' (in tonnes) and 'fuel efficiency' (in km per litre). Identify the response variable and the explanatory variable.

Changing the load on a lorry would lead to a change in the fuel efficiency (e.g. heavier loads would use more fuel). So **fuel efficiency** is the **response** variable and **load on the lorry** is the **explanatory** variable.

So Tasha should plot load on the x-axis and fuel efficiency on the y-axis.

Correlation

*The **Regression Line** (line of best fit) is in the form **y = a + bx***

The **regression line of y on x** (x is the explanatory variable and y is the response variable) is a **straight line** of the form:

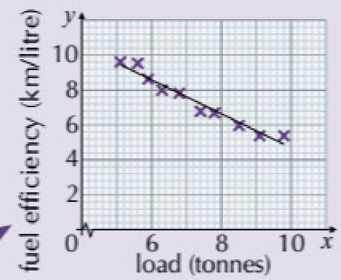
$y = a + bx$, where a = y-intercept and b = gradient

You need to be able to **interpret** the values of a and b.

Example: Tasha's data below shows the load on a lorry, x (in tonnes), and the fuel efficiency, y (in km per litre).

x	5.1	5.6	5.9	6.3	6.8	7.4	7.8	8.5	9.1	9.8
y	9.6	9.5	8.6	8.0	7.8	6.8	6.7	6.0	5.4	5.4

The regression line of y on x is calculated to be $y = 14.5 - 0.978x$.
Plot this data on a scatter graph and interpret the values of a and b.



Plot the scatter graph, with load on the x-axis and efficiency on the y-axis.

a = 14.5: with **no load** ($x = 0$) you'd expect the lorry to do 14.5 km per litre of fuel.

b = –0.978: for every **extra** tonne carried, you'd expect the lorry's fuel efficiency to **fall** by 0.978 km per litre.

*Use regression lines **With Care***

You can use your regression line to **predict** values of the **response variable**. There are two types of this.

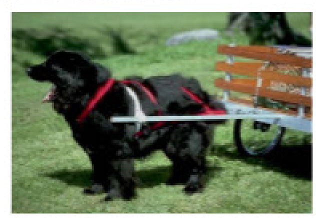
Interpolation — use values of x **within** the data range (e.g. between 5.1 and 9.8 for the lorry example). It's okay to do this — the predicted value should be **reliable**.

Extrapolation — use values of x **outside** the data range (e.g. outside 5.1 and 9.8 for the lorry example). These predictions can be **unreliable**, so you need to be very cautious about them.

Example (continued): Estimate the fuel efficiency when the load is 12 tonnes. Give a reason why your estimate might be unreliable.

Use $x = 12$ in the regression line: $y = 14.5 - 0.978 \times 12 = \mathbf{2.764}$

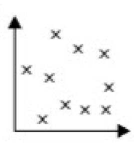
$x = 12$ is outside the data range (5.1 to 9.8) —
this is an **extrapolation** so the estimate may be unreliable.

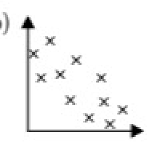


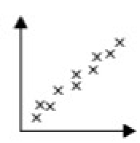
Professor Snuffles had a fuel efficiency of 1.5 km per doggie biscuit.

Practice Questions

Q1 Describe the correlation shown on the graphs to the right:

a) 

b) 

c) 

Q2 Khalid wants to plot a scatter graph for the variables 'barbecue sales' (thousands) and 'amount of sunshine' (hours). Identify the response variable and the explanatory variable.

Exam Question

Q1 The following times (in seconds) were taken by eight different runners to complete distances of 20 m and 60 m.

Runner	A	B	C	D	E	F	G	H
20-metre time (x)	3.39	3.20	3.09	3.32	3.33	3.27	3.44	3.08
60-metre time (y)	8.78	7.73	8.28	8.25	8.91	8.59	8.90	8.05

a) Plot a scatter diagram to represent the data. [2 marks]

b) Describe the correlation shown on your graph. [1 mark]

c) The equation of the regression line is calculated to be $y = 2.4x + 0.7$. Plot it on your scatter diagram. [1 mark]

d) Use the equation of the regression line to estimate the time it takes to run a distance of 60 m, when the time taken to run 20 m is: (i) 3.15 s, (ii) 3.88 s. Comment on the reliability of your estimates. [2 marks]

What's a statistician's favourite soap — Correlation Street...

Watch out for those outliers — you might need to think about why there's an outlier. Sometimes it can only be down to some sort of error, but in others there could be a realistic reason why a data point might not fit the general trend.