# AS Statistics – Chapter 4 – Correlation

Bivariate data is data which has pairs of values for two variables. You can represent bivariate data on a **scatter diagram**. In experiments which generates bivariate data, there is often one **independent** variable for which you control the recorded values, and a **dependent** variable which you measure.

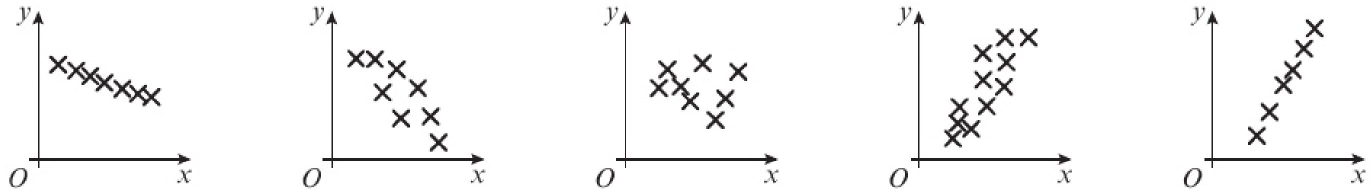
For example, you may choose to measure the temperature of a liquid at 5 minute intervals. In this case, time is the independent variable and temperature is the dependent variable.

On a scatter graph,

- The horizontal axis is the **independent** or **explanatory variable**
- The vertical axis is the **dependent** or **response variable**

The two different variables in a set of bivariate data are often related.

**Correlation** describes the nature of the **linear relationship** between two variables.



Strong negative correlation

Weak negative correlation

No (or zero) linear correlation

Weak positive correlation

Strong positive correlation

For negatively correlated variables, when one variable increases, the other decreases.

For positively correlated variables, when one variable increases, the other increases.

Two variables have a **causal relationship** if a change in one variable causes a change in the other.

**Correlation does not necessarily imply causation!** If two variables are correlated, you will need to consider the context of the question and use your common sense to decide whether the relationship is likely to be a causal one.

## Regression Lines

At GCSE, you drew lines of best fit on scatter diagrams using your own judgement. These are generally inaccurate, and two lines of best fit for the same data may differ significantly.

There are mathematical approaches to drawing lines of best fit, one of which is the **least squares regression line**, which we will refer to simply as the **regression line**. This is a straight line which is designed to minimise the sum of the squares of the vertical distances of each data point from the line – you don't need to know how this is done!

**The regression line of $y$ on $x$ is written as $y = a + bx$**

You do not need to calculate regression lines – the equation will be given in the question if needed.

Note that this is of the form $y = mx + c$, where $b$ is the gradient and $a$ is the vertical intercept.

**The coefficient $b$ tells you the change in $y$ for each unit change in $x$, and its sign tells you the type of correlation.**

Be careful when using regression lines to predict values. They only work one way, to predict a value of the **dependent variable** for a given value of the **independent variable**:

**A regression line of $y$ on $x$ can only be used to estimate values of $y$ given $x$, and not values of $x$ given $y$**

You should also only use regression lines to make estimates using values of the independent variable within the range of the given data (**interpolation**), and not outside the range (**extrapolation**).