8 Looking for relationships: batches and scatter graphs

Key words: population, sample, random sample, batch, variability, stem-and-leaf diagram, histogram, box plot, median, quartile, range, interquartile range, outlier, percentile, scatter graph, variable, correlation, line of best fit.

In the biological sciences, it is quite common to have a data set in which there is an underlying variability in the things being measured. This contrasts with the physical sciences, where variability is more often due to measurement uncertainty. This leads to differences in the way that data are collected (e.g. in thinking about sampling) and the way they are analysed (e.g. in dealing with correlations).

8.1 Different kinds of relationship

Both of the following questions are about how one variable is related to another:

- How does the rate of reaction of zinc with hydrochloric acid relate to temperature?
- How does the lifespan of mammals relate to their heart rate?

Although these look like rather similar types of question, there are important ways in which they are very different.

In the first case, it would be necessary to control all of the variables apart from temperature, such as the amount of zinc, the size and shape of the granules, and the volume and concentration of the acid. However, having done that, we would expect that for every value of the temperature, there would be a unique value of the reaction rate. If the experiment were repeated with *exactly* the same conditions, we would expect that the reaction rate would be *exactly* the same. Of course, the actual values obtained might not be the same because of measurement uncertainty. However, this is due to the limitations in the measuring equipment or our ability to measure, rather than a difference in the phenomenon itself.

The second question is rather more complex. There are different species of mammal, so would it be necessary to collect data for *every* species, or would it be sufficient to make a *selection*? In addition, what does it mean to talk about the lifespan or heart rate for a particular type of mammal, such as a tiger? Different individuals have different lifespans and heart rates. Perhaps taking an average for all tigers? But it would be impossible to collect data on all tigers, so perhaps just a selection?

What this illustrates is that, in designing the collection of data to help answer this kind of question, it is important to consider what to select – in other words, how to *sample*.

8.2 Populations and samples

To discuss the nature of sampling, a simple artificial example will be used. Suppose you have a bag containing a very large number of 1p coins. The composition of 1p coins was changed in 1992, so from that year they were magnetic while before then they were not. Without using a magnet, how could you estimate the proportion of the old non-magnetic coins in the bag? One way would be to check the year of every coin in the bag, but this would take a long time, so better would be to check just a small number at random and make a guess – a process of *sampling*. If you drew out just four coins and found that one of them was pre-1992, you might not be very confident in guessing the proportion in the bag. If however, you drew out 40 coins and found that 10 of them were pre-1992, you would be more confident in saying that the bag might contain around 25% of these coins.

There are two important technical words that can be applied to a situation like this. The collection of all the coins in the bag is known as a ***population***. The smaller set of coins selected for checking is known as a ***sample***. Note that, in statistics, the term 'population' has a different meaning to that in everyday language – it means a set of things of a similar nature that is of interest as a whole. In everyday language, a 'population' usually refers to a group of people or animals living in a particular area. In a statistical population, however, the 'individuals' could be any kind of object or event.

For example, if the group of interest were the 'population of people in the UK' (an everyday expression) then this would also be a 'population' in the statistical sense. In a factory making computer chips, however, the quality control department might define a 'population' as the group of chips that are made each day, in order that a sample of these can be selected and tested. Not all 'populations' consist of objects: one could think of a 'population' of earthquakes, in which each earthquake is seen as an 'individual'.

'Populations' can apply at different levels. A research study on bees might be interested in all the beehives in a particular area. This collection of beehives is the 'population', and each beehive is an 'individual'. Another study might just look at the worker bees within a single hive. Here, the population is the collection of all the worker bees in that hive, and each worker bee is an individual.

When sampling from a population, it is important that the sample is *representative of the population*. For a collection of 1p coins in a bag, this is fairly easy: coins can just be taken out of the bag *at random*. This means that every coin in the bag has *an equal chance of being selected*, and thus the sample is representative of the population. Such a sample is called a ***random sample***. Giving the bag a good shake and not always choosing a coin from the same part of the bag would be a good way of ensuring this.

Another important principle in sampling is the effect of *sample size*. The larger the sample size, the more likely it is that the sample will be *representative* of the population. With samples of just a few coins, there will be a lot of random variation. A sample of a large number of coins is more likely to have a composition similar to that of the whole set of coins. Choosing a small sample means that collecting the data is easier, but with larger samples there is more confidence that the data are representative. In practice, sample size is determined by the balance between these two factors.

Sampling a bag of coins is straightforward but collecting data to answer the question 'How long does a tiger live?' is rather more complex. The 'population of tigers' is harder to define than the 'population of coins'. The 'individuals' in the population would be individual tigers, but which ones? Those that die in a particular year? It is also harder to collect appropriate

data and to design sampling procedures that ensure the samples are representative. However, even though the problem is more difficult, the same basic principles of sampling still apply. Having found a 'typical value' for the lifespan of a tiger, this might then be compared to the lifespans of other mammals (polar bears, chimpanzees, grey squirrels, and so on) to see what factors influence the lifespan of different types of mammal. This could involve sampling from a population of 'all types of mammals' in which the individuals would be 'types of mammal'.

8.3 Analysing a batch of data

The table in Figure 8.1 shows the potential lifespan of some selected types (or 'orders') of mammals. The whole data set in fact contains 75 selected mammals listed alphabetically (taken from an article on the internet), but just the first few are given here.

Figure 8.1 Potential lifespans of mammals

Mammal	Lifespan (years)
African elephant	60
African giant rat	5
African porcupine	20
Alpine marmot	13
American beaver	19
American bison	23
Asiatic or Indian elephant	78
Australian sea lion	12
Aye-aye	7
Bactrian camel	26
Baikal seal	56
etc.	

A data set like this is sometimes called a ***batch*** of data – it contains a set of values about the *same kind of thing*. The values therefore relate to just a *single quantity or variable*. Such data sets are discussed in Chapter 6 *Dealing with variability*, in which it is shown how a ***box plot*** is a useful visual display for getting a sense of the size and ***variability*** of the values. Note that the term 'batch' is commonly used in data analysis, though not so much at school level: it is used here, as it is very helpful to have a simple term to describe this kind of data.

This section looks at the techniques for drawing a single box plot for one batch of data. The way in which relationships can be explored by analysing more than one batch of data is discussed later in Section 8.5 *Comparing batches of data* on page 81.

Drawing a box plot requires the values to be put in order of size, so that five summary values can be identified (see Section 6.6 *How much do the values vary?* on page 56). Ordering values is easy to do with a computer spreadsheet but if the values only exist on paper it would take a long time to enter them.

A quick and simple way of organising a large set of data by hand is to construct a ***stem-and-leaf diagram***. In this method, the values are first roughly sorted in order of size along a 'stem', and then in a second pass, the individual values (the 'leaves') are put into exact order.

Figure 8.2a shows how to make a start: a vertical 'stem' is drawn with each digit representing 'tens of years' (i.e. 0, 1, 2, 3 . . . 8 represent 0, 10, 20, 30 . . . 80 years). The final digits of each of the data values are the 'leaves' (with units of 'years'), written in the appropriate positions on the 'stem'.

So, the first value in the list of data is 60 years: to the right of the '6' on the stem is written '0'. The next data value is 5 years, and to the right of the '0' on the stem is written '5'. This is continued for each of the values in the batch, writing each new 'leaf' to the right of the existing ones. Thus, in Figure 8.2a, next to the '1' on the stem, are the digits '3 9' – these represent two data values, 13 years and 19 years.

Figure 8.2b shows the diagram with all 75 values entered. Putting values into groups like this is similar to the construction of frequency tables from discrete data (see Section 3.2 *Using tables to process data* on page 24). The shape of this diagram is similar to the outline of a ***histogram*** 'on its side'. It shows that there are many values in the intervals 10–19 years and 20–29 years, so this gives a sense of the 'typical' lifespan.

Figure 8.2 Making a stem-and-leaf diagram

(a) The first few values

8	
7	
6	0
5	
4	
3	
2	0
1	3 9
0	5

stem: 10 years
leaves: 1 year

(b) All values added

8	
7	8 3
6	0
5	6 0 5 5 0
4	0 6 7 2 7 0 9 5
3	4 1 0 0 0 2
2	0 3 6 8 4 0 0 0 3 0 4 0 0 1 9 7 0 4 0 6
1	3 9 2 8 6 7 6 2 7 5 4 5 6 4 6 5 3 5 6 0 4 2
0	5 7 9 7 6 6 4 3 5 7 3

stem: 10 years
leaves: 1 year

Having ordered the values along the stem, the next step is to put the 'leaves' in order, as shown in Figure 8.3a. Reading from left to right starting at the bottom and going up the stem, the whole batch of values is now completely ordered. (The easiest way to do this is to cross off the digits on the unordered diagram as they are entered on the new ordered diagram.)

Figure 8.3 Ordering and finding summary values from a stem-and-leaf diagram

(a) Ordered

8	
7	3 8
6	0
5	0 0 5 5 6
4	0 0 2 5 6 7 7 9
3	0 0 0 1 2 4
2	0 0 0 0 0 0 0 0 0 1 3 3 4 4 4 6 6 7 8 9
1	0 2 2 2 3 3 4 4 4 5 5 5 5 6 6 6 6 6 7 7 8 9
0	3 3 4 5 5 6 6 7 7 7 9

stem: 10 years
leaves: 1 year

(b) Summary values

8	
7	3[8]
6	0
5	0 0 5 5 6
4	0 0 2 5 6 7 7 9
3	0 0[0 1]2 4
2	0 0 0 0[0]0 0 0 0 1 3 3 4 4 4 6 6 7 8 9
1	0 2 2 2 3 3 4[4 4]5 5 5 5 6 6 6 6 6 7 7 8 9
0	[3]3 4 5 5 6 6 7 7 7 9

stem: 10 years
leaves: 1 year

The shape of the ordered stem-and-leaf diagram is the same as before but it is now possible to find the summary values for the box plot. These are indicated in Figure 8.3b. The highest and lowest values are easy to identify (78 and 3 years). Since there are 75 values in the whole batch, the ***median*** is the 38th value (with 37 values below it and 37 values above it): it is 20 years.

To find the upper and lower ***quartiles***, the upper and lower halves of the data are each taken as including the median, so they each have 38 values. (Note that this convention is not universal – some sources say that the upper and lower halves are taken without including the median.) Since this is an even number, the 'middle' consists of two values (the 19th and 20th), and a mean is taken of these two. This gives an upper quartile of 31 years (from 30

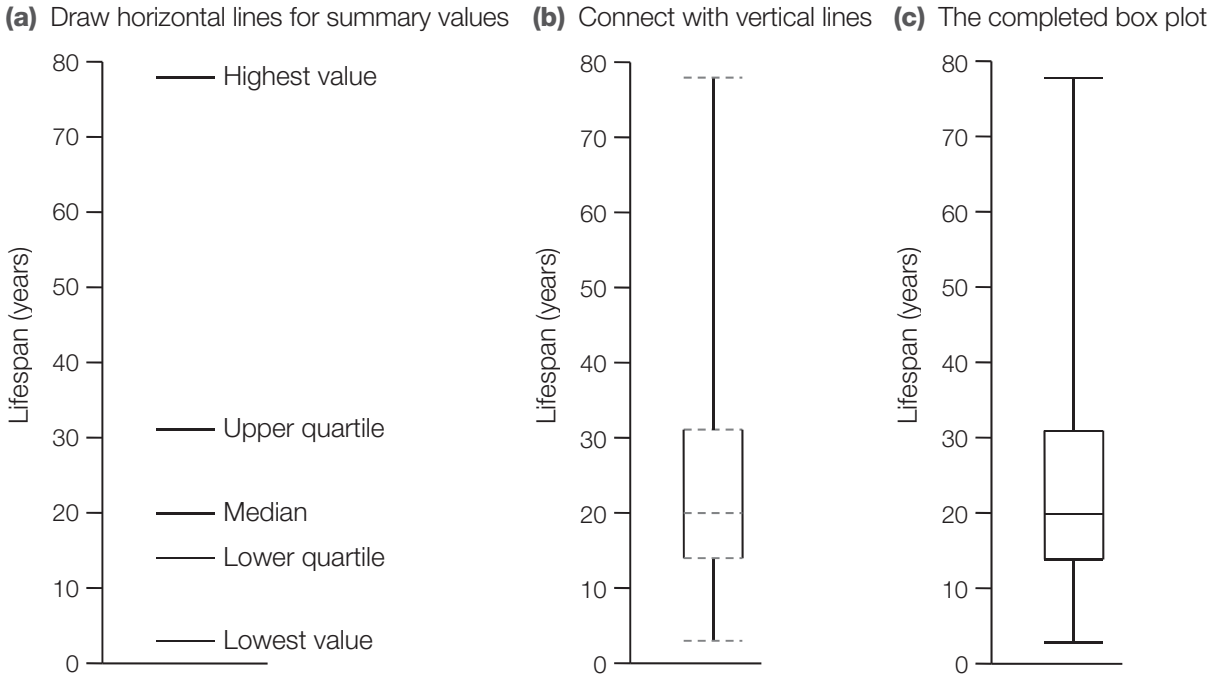
and 31 years, rounded up from a mean of 30.5 years), and a lower quartile of 14 years (from the values 14 and 14 years).

The five summary values are therefore:

- highest value: 78 years
- upper quartile: 31 years
- median: 20 years
- lower quartile: 14 years
- lowest value: 3 years

From these summary values a ***box plot*** can be drawn. Figure 8.4 shows the easiest way of doing this by hand. The first step is to draw an appropriate vertical scale and then five horizontal lines corresponding to the summary values, as shown in Figure 8.4a. The next step is to draw the vertical lines to create the box, and connecting the highest and lowest values, as shown in Figure 8.4b. The completed box plot is shown in Figure 8.4c.

Figure 8.4 Drawing a box plot: potential lifespans of mammals



The box plot gives a very clear sense of the variability of mammalian lifespans. While the median value is 20 years, there is a very large variation, with a ***range*** of nearly 80 years (in fact from 3 to 78 years). As noted in Section 6.6 *How much do the values vary?* on page 56, a better measure of spread is the ***interquartile range***, i.e. the range of values included within the box on the box plot. Its value here is 16 years (the difference between the upper quartile and the lower quartile: 30 − 14 years). One half of the mammals in this batch have potential lifespans within this range.

The box plot also shows that the batch of data is *skewed*, with the upper part being 'stretched out' and the lower part being 'squashed together'. This contrasts with the display in Chapter 6 (Figure 6.3) which skewed in the opposite direction.

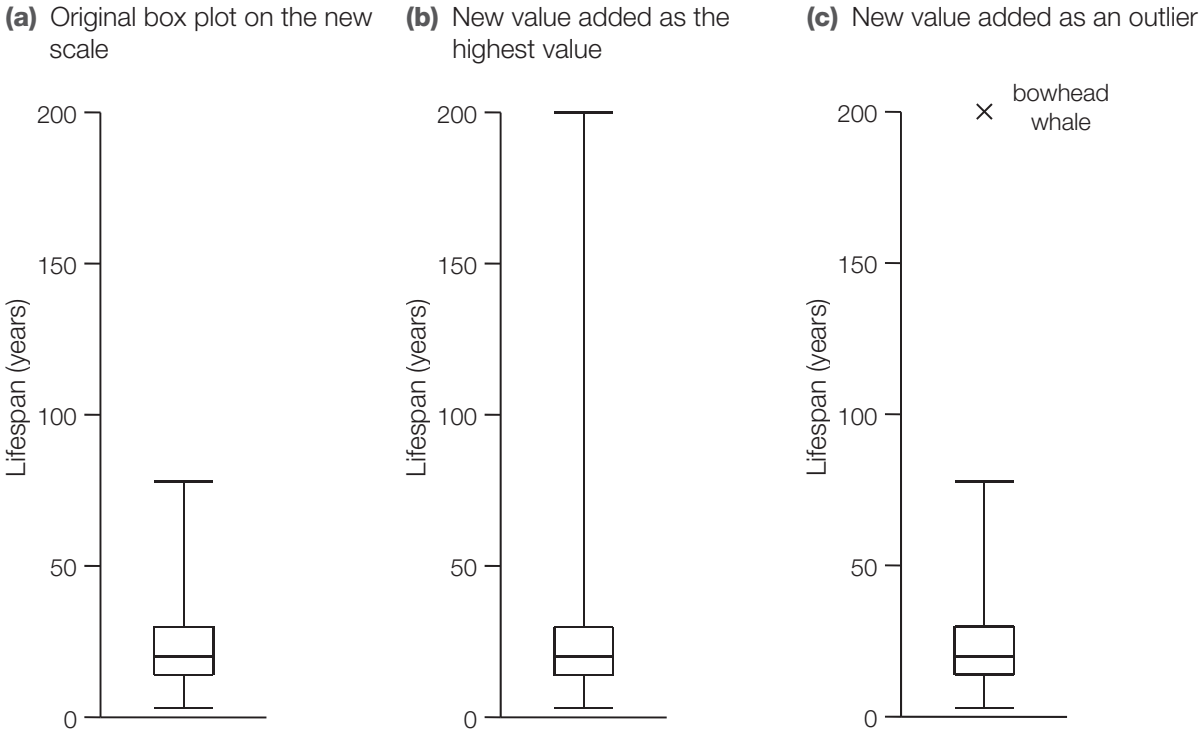
Note that, in mathematics, pupils may have seen stem-and-leaf diagrams drawn so that values of the stem increase *downwards*. Many books on analysing data adopt the convention that values of the stem increase *upwards*. This is the convention that has been used here, since the

stem then matches the scale used for the box plot, and it corresponds to the terms used to describe the summary points (e.g. the highest value is at the top, and so on).

Sometimes, batches of data contain extreme values that are very different from the rest of the values. These are known as ***outliers***. For example, a bowhead whale has a lifespan of about 200 years. This was not included in the original batch of data but, if it had been, how could this value be handled? Figure 8.5a shows the box plot of the original batch drawn on a new scale that extends upwards to 200 years. Figure 8.5b shows the box plot redrawn with the upper line extended to the new highest value. Such a plot could be misleading since it may give the impression that there are quite a few types of mammal with lifespans approaching 200 years.

In Figure 8.5c, the original box plot has been drawn, with the new value indicated as an outlier by showing a separate data point labelled 'bowhead whale'. This gives a much better impression of how different this particular type of mammal is from the others in the batch. Such a plot prompts questions to be asked about the reasons for this difference, perhaps related to the whale's size, habitat, diet, and so on.

Figure 8.5 Dealing with outliers: potential lifespans of mammals



There are no hard-and-fast rules about how to identify what should be considered an outlier. Sometimes there will be values that are very much bigger or very much smaller than the rest of the values and which are clearly outliers; sometimes all of the values in a batch lie fairly close together where there are clearly no outliers. Between these extremes, the decision is a matter of judgement depending on the nature of the data and on what is of interest in the analysis.

Pupils may encounter the term ***percentile*** in science lessons. This is a similar idea to a *quartile*. You can think of *quartiles* as the values that split a batch of data into *four parts*; in the same way, the *percentiles* are the values that split a batch of data into a *hundred parts*. For example, the World Health Organization has data on the weights of babies at different ages. For baby girls aged 12 months, it gives the value of the 90th percentile as 10.5 kg. This means that 90% of these babies are under this weight, while 10% are over. Since a *median*

is the value in the *middle* of a batch, it is the *50th percentile* (50% above and 50% below). The *upper and lower quartiles* correspond to the *75th and 25th percentiles* respectively. Using percentiles can be useful when looking at the way that values are distributed in a batch in a more detailed way than simply using quartiles.

8.4 Dealing with more than one batch of data

The previous section looked at the analysis of a *single* ***batch*** of data. There are two distinct types of situation where you might be dealing with more than one batch of data. These are illustrated in Figure 8.6.

The first set of data, represented in part in Figure 8.6a, consists of two batches of data: the lifespans for two different types of mammals (rodents and primates). Here, the data are about the *same quantity* for *two different samples*.

The second set of data, represented in part in Figure 8.6b, also consists of two batches of data: the mean heart rates and mean lifespans for selected types of mammals. Here, however, the data are about *two different quantities* related to the *same sample*.

Figure 8.6 Different structures of data

(a) Lifespans for rodents and primates

Rodent	Lifespan (years)
African giant rat	5
African porcupine	20
etc.	

Primate	Lifespan (years)
Aye-aye	7
Chimpanzee	55
etc.	

same quantity
two different samples

(b) Mean heart rates and mean lifespans for different types of mammal

Mammal	Heart rate (beats/min)	Lifespan (years)
Badger	138	11
Cat	120	15
etc.		

two different quantities
same sample

So, although both sets of data consist of two batches, they have *different structures*, and this leads to *different types of questions* that can be asked about the data, for example:

(a) What is the relationship between *lifespan* and *type of mammal* (rodent and primate)?
This is a question about the relationship between a ***continuous*** *variable* and a ***categorical*** *variable*.

(b) What is the relationship between *lifespan* and *heart rate* for different types of mammal?
This is a question about the relationship between *two* ***continuous*** *variables*.

Since these are different types of question involving data of different structures, the data are analysed in different ways. These are discussed in the next two sections.

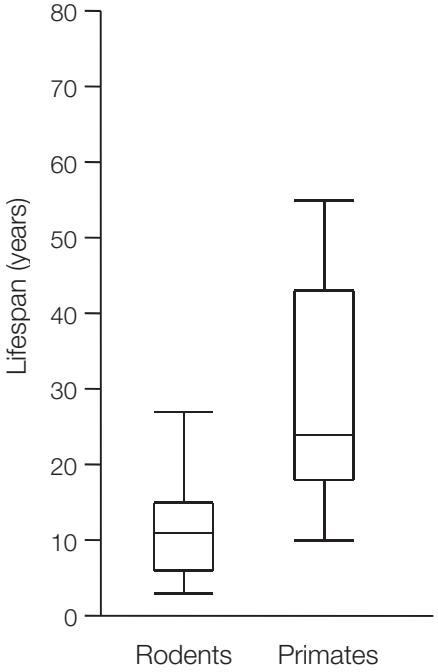
8.5 Comparing batches of data

Box plots are particularly useful when analysing two or more ***batches*** of data, such as the tables shown in Figure 8.6a, rather than just a single batch. The lifespans of rodents and primates can be compared by drawing a box plot for each of these batches side-by-side, as shown in Figure 8.7.

As for a single box plot, the same questions can be asked about each of these box plots individually: What is the typical lifespan for each of these types of mammal and how much do they vary? In addition, the two batches can be compared. It seems that primates typically live about twice as long as rodents (the median for primates is about twice that for rodents), but there is a lot of overlap. For example, the median value for rodents is a bit higher than the lowest value for primates. This means that a bit over a half of the rodents live longer than the shortest lived primate.

Box plots become even more useful when comparing larger numbers of batches. For example, Figure 8.7 could be extended to include other types of mammal, and this would enable a great deal of data about lifespans to be compared quite easily.

Figure 8.7 Lifespans of rodents and primates compared



8.6 Judging whether a difference is significant

Box plots can be useful when analysing the results of experiments involving a 'control' group and a 'treatment' group. For example, when looking at the effects of a fertiliser on the height of plants, one would expect there to be a variation in the heights of individual plants in both groups. However, the question is whether, *overall*, the plants in the two groups seem to be of different heights, i.e. are the medians of the two groups noticeably different?

If there is no overlap of the 'boxes' (i.e. of the *interquartile ranges*) in the two box plots then it might seem that the batches are very different and the fertiliser had an effect. However, if there is a lot of overlap of the boxes and the medians are quite close then the observed difference may simply be due to chance.

In post-16 biology, students learn about more formal statistical tests of significance to judge whether two batches might really be different or whether the difference could have arisen by chance. Instead of using the median and the interquartile range in the comparison of box plots, these tests use the mean and the standard deviation. However, the principles of these formal tests are similar to how judgements are made by eye using box plots, so representing the data visually in this way forms a good basis for further understanding.

Note that the word '*significant*' has both an everyday meaning and a technical statistical meaning. In everyday language, a *significant effect* is often used to mean a *big effect*. However, in statistics, a significant effect means an effect that is *unlikely to have happened by chance*: it does *not* necessarily indicate that it is a big effect. This is a subtle idea, and certainly goes beyond what pupils need to know in 11–16 science. However, because 'significance' is a term that is encountered in media reports and can cause confusion, it is discussed in outline here.

The idea can be illustrated by thinking about coin tosses (see Section 6.9 *Basic ideas in probability* on page 59). A *fair* coin is one that has an equal probability of landing as a head or a tail. If you tossed a coin 10 times, you might get equal proportions of heads and tails (i.e. 5 of each), but you would not be surprised if you got another outcome (e.g. 7 heads and 3 tails). There is random variation in the outcomes. However, if you tossed the coin many thousands of times, you would expect the proportions of heads and tails to be very close to equal.

Now, suppose you were given an *unfair* or '*biased*' coin, which had an 80% probability of landing heads. If you did not know whether it was biased or not, you might start to suspect after only a few throws that it was not a fair coin. The greater the number of tosses, the more you might believe that it was biased. Eventually you might say that you were 'fairly certain'. You could never be *completely* certain, because an outcome with a very large proportion of heads could still be possible, even though it might be highly unlikely.

Suppose you were given another biased coin, but this time much less so, with just a 51% probability of heads. From a small number of coin tosses, you would not notice that it was biased. By counting a larger number, you might suspect something, but you would need a much larger sample of coin tosses with this coin before you might say you were 'fairly certain'.

What you are really trying to judge here when using a biased coin is whether you might have got these results using a fair coin. If the proportion of heads seems rather too large, you might say that that you are 'fairly certain' that the outcome could not have happened by chance using a fair coin. With some complex calculations, it would be possible to turn 'fairly certain' into a value of a probability; for example, 'there is a 95% probability that this outcome could not have happened by chance'. Another way of saying the same thing is that there is only a 5% probability that the outcome could have happened by chance.

This is the essence of statistical significance. If an outcome is reported as being 'significant at the 5% level', it means that the probability of the outcome happening by chance is only 5% or less. If the reported level of significance is lower (e.g. 1%), it means that there is an even smaller probability that the outcome could have happened by chance.

These ideas are important in experiments involving a 'control' group and a 'treatment' group. Suppose that these were the results of two different studies on fertilisers (with a significance level of 5%):

- With fertiliser A, there was a 7% bigger growth in the treatment group than in the control group. The sample sizes were small and the result was not significant.
- With fertiliser B, there was a 0.1% bigger growth in the treatment group than in the control group. The sample sizes were large and the result was significant.

What these results illustrate is that with small sample sizes you might not get a significant result even with a large effect. With large samples sizes you might get a significant result, even though the effect is small.

So, although the result for fertiliser A is not significant, the difference in the growth seems quite big. It might be worth doing another study with larger samples to see whether the effect might be real, or whether it just happened by chance. For fertiliser B, the result was significant, but the effect is so small that it may not justify the cost of using the fertiliser.

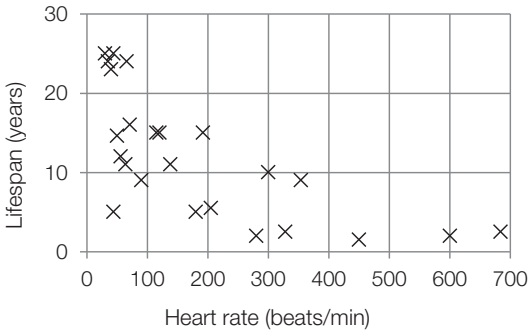
In summary, box plots can be used to compare samples and look at the *sizes of the differences* between them. Statistical tests are used to judge whether the differences are *significant*, in other words, whether it is unlikely they could have occurred by chance.

8.7 Relationships between variables: scatter graphs and correlation

Scatter graphs are useful for looking at the relationship between two ***variables*** of the same sample of individuals. The table shown earlier in Figure 8.6b illustrates this kind of data. The sample consists of selected types of mammals and the quantities are mean heart rate and mean lifespan. Figure 8.8. shows a scatter graph of these data.

This graph suggests that there is *some* relationship between lifespan and heart rate. It seems that, very roughly, as the heart rate of a mammal *increases*, its lifespan tends to *decrease*.

Figure 8.8 A scatter graph of mean lifespan against mean heart rate for some selected types of mammal



However, it is certainly not an *exact* relationship. It is very different, for example, to the relationship between the mass suspended from a spring and the length of the spring. Here, it is straightforward to control all the variables that affect the length of a spring, and just look at the effect of the suspended mass. For every value of mass, there is a *unique* value for the length of the spring.

For the data on mammals, it is much harder to control all of the variables that could affect a dependent variable. In such a case, it may be possible to see the effect of an independent variable on the dependent variable, but it will be masked by the effects of additional variables that are not controlled. We would *not* expect that for every value of one variable there would be a unique value for the value of the other. There is *variability* in the population.

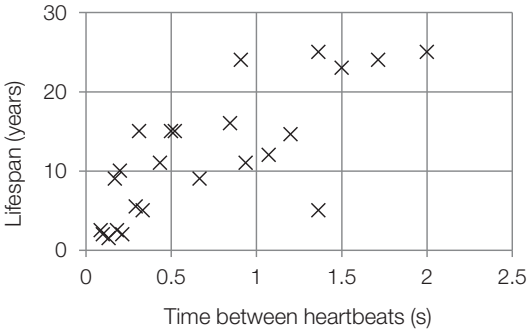
The distinction between these two examples is also discussed in Section 3.6 *Line graphs and scatter graphs: two related quantities* on page 29. (Note: for the purposes of the discussion in that section, the graph of the mammalian data showed fewer data points.)

One way of explaining the apparent relationship between lifespan and heart rate would be that every mammal tends to have the same fixed number of heart beats in its lifetime. So, if the time between heart beats doubles then the lifespan doubles too.

Figure 8.9 has been drawn to test this idea. The quantity plotted along the horizontal axis is the *time between heartbeats.* This has been calculated from: time between heartbeats in seconds = 60 / heart rate in beats per minute.

Re-expressed in this way, the pattern of the data points now shows a general upward slope. The graph shows that, as the time between heart beats increases, the lifespan increases. The pattern also appears less curved: since it is a bit more 'straight', it suggests that, *very roughly*, if the time between heart beats doubles then the lifespan doubles too. It is clearly not possible to fit a straight line that passes through or close to all of these points, though it can still be useful to draw a line of best fit on a scatter graph like this. (This is discussed later in Section 8.8 *Drawing a line of best fit on a scatter graph* on page 85.)

Figure 8.9 A scatter graph of mean lifespan against time between heartbeats for different types of mammal



A ***correlation*** is a way of expressing the strength of the relationship between two variables. If a scatter graph appears to show that there is a relationship between two variables then we can say that they are *correlated*.

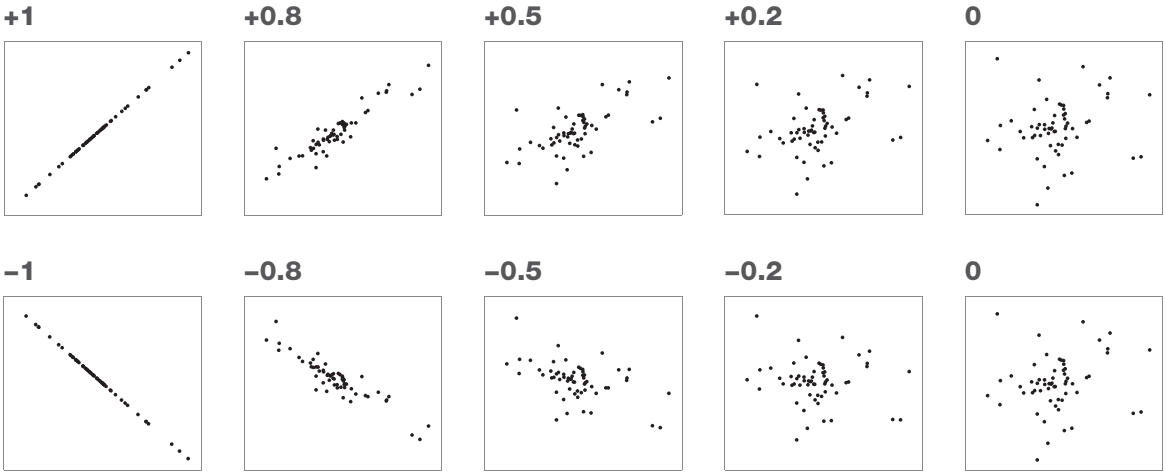
In 11–16 science, pupils only need to be able to talk about correlation qualitatively, but it is worth being aware that it is not just a 'vague' idea. A *correlation coefficient* is a value that can be *calculated* for certain types of data, and students of post-16 biology encounter this

quantitative aspect. Calculating a correlation coefficient gives a value between +1 and −1. A positive value indicates that, as one variable increases, the other also tends to increase; a negative value indicates that, as one variable increases, the other tends to decrease. A value of 0 indicates no correlation between the variables.

Figure 8.10 shows a series of scatter graphs with varying degrees of correlation. These can be described *qualitatively* using the following terms:

- *positive correlation*: as one variable increases, the other tends to increase as well
- *negative correlation*: as one variable increases, the other tends to decrease
- *no correlation*: there does not appear to be any relationship between the variables
- *strong* or *weak* correlation: to describe how closely the variables appear to be related.

Figure 8.10 Correlations (positive and negative) of different strengths



Thus, for these graphs, we could talk about a strong positive correlation between two variables (+0.8), a weak negative correlation between two variables (−0.5), or no apparent correlation (0). (Note that the variables in the graphs labelled '+1' and '−1' are *perfectly correlated* – they show the points you would expect on straight line graphs.)

An important point that is often made is that *correlation does not imply causation*. If A and B are correlated then it is possible that A causes B, but another possibility is that B causes A. It may be that they are not causally related to one another at all, but that they are both causally related to a third variable C. It is also possible that the apparent correlation between them is just coincidental and happened by chance.

Note that a correlation coefficient is a value that indicates the *size* of an apparent relationship. In more advanced statistical work, a further test needs to be done to judge whether the effect is *significant* or whether it could have arisen by chance. The larger the sample size, the more likely an effect is to be significant.

8.8 Drawing a line of best fit on a scatter graph

The data points on a ***scatter graph*** are usually, as the name indicates, *scattered*. This is because of the underlying variability in the population concerned. For this reason, the data points do not lie close to any ***line of best fit***.

It is, however, still possible to draw a line of best fit on a scatter graph, though it is important to be clear about its meaning. It has a different meaning to the lines of fit discussed in

Chapter 7 *Looking for relationships: line graphs*. The difference arises because of the differences in the nature of the data. For 'line graph' type data, the data points may not all lie on the fitted line because of *measurement uncertainty*. For 'scatter graph' type data, the differences from a fitted line are due to *differences between individuals in a population*. The distinction between these two different kinds of variability is discussed in Section 6.1 *Where does variability come from?* on page 50.

Figure 8.11 A line of best fit on a scatter graph

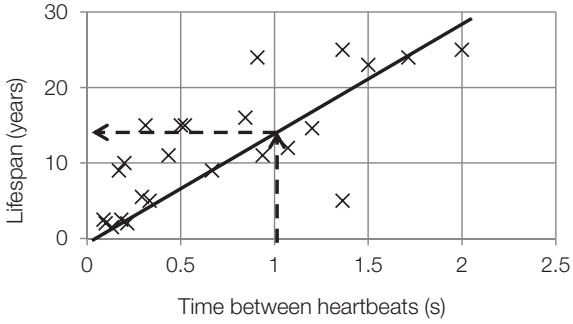


Figure 8.11 shows the same scatter graph as in Figure 8.9, but here a line of best fit has been added. A straight line has been chosen, since the distribution of data points does not look particularly 'curvy' The criteria for fitting the line are that it should have similar numbers of points above and below the line and the gradient of the line should reflect the distribution of the points (as discussed in Section 7.4 *Lines of best fit: linear relationships* on page 69). Deciding on where to draw the line when the points are very scattered involves more judgement than when they are close to a line. It is not so easy to decide on which is the 'best' line.

The straight line on the graph can now be used to make *estimates* or *predictions*. For example, suppose you were asked what lifespan a mammal might have if it had a 'time between heartbeats' of 1 second. The dotted lines in Figure 8.11 show that the best guess would be about 14 years. This is only a rough estimate but it is a better guess than if we had no information at all about the time between heartbeats. It is not a perfect guess since the data points do not lie exactly on the line because of the variability in the population.

In this example, a *straight* line was chosen as the line of best fit because the distribution of the data points suggested it (the formal term for such a line is a *regression line*). In mathematics, pupils are generally given data for which a straight line is a good fit; in science, pupils may need to decide whether a straight line or a curve is the best fit. For example, in the scatter graph shown in Figure 3.10, a curve would be a better fit for the data than a straight line.