

4

Between-individual variation, replication, and sampling

If we are interested in how one characteristic of our experimental subjects (variable *A*) is affected by two other characteristics (variables *B* and *C*), then *A* is often called the *response variable* or *dependent variable* and *B* and *C* are called *independent variables*, *independent factors*, or sometimes just **factors**. It is likely that *A* will be affected by more than just the two factors *B* and *C* that we are interested in.

Any variation in the response variable between individuals in our sample (between-individual variation) that cannot be attributed to the independent factors is called **random variation**, *inherent variation*, *background variation*, *extraneous variation*, *within-treatment variation*, or **noise**.

- In any experiment, your experimental subjects will differ from one another.
- Experimental design is about removing, or controlling for, variation due to factors that we are not interested in (random variation or noise), so we can see the effects of those factors that do interest us (Section 4.1).
- One key aspect to coping with variation is to measure a number of different experimental subjects rather than just a single individual, in other words to replicate (Section 4.2).
- Having identified a population of independent experimental subjects, we then need to sample from that population—and this can be done in a number of different ways (Section 4.3).

4.1 Between-individual variation

Wherever we look in the natural world, we see variation: salmon in a stream vary in their body size; bacteria in test tubes vary in their growth rates. In the life sciences, more than physics and chemistry, variation is the rule; and the causes of variation are many and diverse. Some causes will be of interest to us: bacteria may vary in their growth rates because they are from different species or because a researcher has inserted a gene into some that allows them to utilize an additional type of sugar from their growth media. Other causes of variation are not of interest: maybe the differences in growth rates are due to small unintended differences in temperature between different growth chambers, or to differences in the quality of the media that cannot be perfectly controlled by the experimenter. Some of the measured variation may not even be real, but due to the equipment that we use to measure growth rates not being 100% precise (see Section 11.2). Hence we can divide variation into variation due to **factors** of interest, and **random variation** or **noise**. Of course, whether variation is regarded as due to a factor of interest or to random variation depends on the particular question being

asked. If we are interested in whether different strains of a parasite cause different amounts of harm to hosts, the ages of individual hosts is something that we are not really interested in, but could nevertheless be responsible for some of the variation we observe in the experiment. However, if we are interested in whether parasites cause more harm in older hosts than in younger hosts, the variation due to age becomes the focus of our investigation.

In very general terms, we can think of life scientists as trying to understand the variation that they see around them: how much of the variation between individuals is due to things that we can explain and are interested in, and how much is due to other sources. Whenever we carry out an experiment or observational study, we are either interested in measuring random variation, or (more often) trying to find ways to remove or reduce the effects of random variation, so that the effects that we care about can be seen more clearly. Life scientists have amassed a large array of tools to allow us to do this, and whilst the multitude of these and their specialist names can appear daunting, the ideas underlying them are very simple. In this chapter, we outline the general principles of experimental design, replication, and sampling, and explain how these can be used to remove the masking effects of random variation.



Generally, we want to remove or control variation between subjects due to factors that we are not interested in, to help us see the effects of those factors that do interest us.

4.2 Replication

Random variation makes it difficult to draw general conclusions from single observations. Suppose that we are interested in whether male and female humans differ in height. That is, we want to test the hypothesis:

Sex has an effect on human height.

An admittedly left-field way to test our hypothesis might be to find the heights of Pierre and Marie Curie from historical records, and find that Marie was shorter than Pierre. Does this allow us to conclude that human females are shorter than males? Of course not! It is true that some of the difference in height between Pierre and Marie will be due to any general effects of sex on human height. However, there will be a number of other factors that affect their height that are nothing to do with sex. Maybe Pierre had a better diet as a child, or came from a family of very tall people. There are numerous things that could have affected the heights of these two people that are nothing to do with the fact that Pierre was male and Marie female. The difference in height that we observe between Pierre and Marie might be due to the factor that we are interested in (sex), but could equally just be due to other factors.

The solution to this problem is to sample more individuals so that we have replicate males and females. Suppose we now go and measure ten married couples and find that in each case the man is taller than the woman; our confidence that men really are taller

*Replication involves sampling and taking the same measurements on a number of different subjects. Indeed, subjects are often called **replicates** and their number the *number of replicates*.*

? **Q 4.1** Are we safe in restricting our sample to married couples?

than women on average would increase. The reason for our increased confidence is not very profound. Common sense tells us that, while it is easily possible that differences in height between a single male and a single female could be due to factors other than sex, it is very unlikely that chance will lead us to select ten out of ten couples where the man was taller, when actually couples where the woman is taller are just as common in the population. This is as likely to occur by chance as tossing a coin ten times and getting ten heads. It is unlikely that in all the chosen couples the males had better diets as children than the females, or all the females came from characteristically short families (unlikely but not impossible, see Section 1.4.2 and Chapter 7 for further consideration of such *confounding factors*). If we found the same pattern in 98 out of 100 couples we would be even more confident. What we have done is **replicate** our observation. If differences were due only to chance, we would not expect the same trend to occur over a large sample, but if they are due to real differences between males and females we would expect to see consistent effects over replicate measurements. The more times we make a consistent observation, the more likely it is that we are observing a real pattern. All statistics are based on replication, and are really just a way of formalizing the idea that the more times we observe a phenomenon the less likely it is to be occurring simply by chance. See Box 4.1 for a more detailed treatment of the relationship between random variation and replication.

BOX 4.1 A detailed example of variation and replication: does genetic modification to chicken feed make egg shells thinner?

Given the importance of variation and replication it is worth spending some time thinking about these issues with the use of an example.

Imagine someone working for the research arm of an animal feed company that is considering whether to add grain that has been genetically modified (GM) to their standard chicken feed. However, it has been suggested that the particular type of genetic modification being used may have a negative effect on the shell quality of eggs laid by chickens, making them thinner and more fragile. How should they decide whether this is the case or not? Hopefully, given that you are reading this book, the answer is obvious to you: they should do an experiment.

The researcher will immediately be faced with the problem of biological variation and how to deal with it. Egg shells, like almost everything we might measure in biology, vary in thickness from egg to egg. The thickness of a shell will vary for many reasons. One of these reasons might be whether or not the hen had been fed GM feed, but the thickness of a particular egg shell may also depend on the time that it is laid, the age of the hen, the breed of the hen, how much water she has drunk, the temperature of the hen house ... the list goes on and on. The researcher's job is to try to understand this variation and determine how much, if any, of it is due to the feed the hen has been given. Or to put it another way, the research needs to test the null hypothesis:

The thickness of an egg shell is not influenced by whether the hen that laid it has been fed standard or GM feed.

Let's start by thinking about hens that have been fed the standard feed. Imagine we have five such hens. We take one egg from each hen (why only one? ... this is discussed in detail in Chapter 5) and measure the thickness of the shell. We can be extremely confident of one thing, the thicknesses of these egg shells are very unlikely to be the same—egg shell thickness, like almost everything in biology, varies. We can do our best to reduce the variation by standardizing as many other factors as possible. So we might choose to use hens of the same breed, of the same age, and kept in identical conditions. We might choose eggs laid on the same day, and we would certainly try to make sure that we measured all eggs using the same procedure. Reducing variation in this way is one of the reasons that we carry out experiments in carefully controlled conditions. However, no matter how hard we try to standardize everything, it is still likely that the shells will vary to some degree. Good experimental technique can reduce variation, but it will rarely completely remove it.

Now let's get a picture of what this variation looks like. Imagine that we had the shell thickness measures for literally hundreds of thousands of eggs from hens fed standard feed. If we looked at the distribution of this data the chances are that it would look a lot like one of the distributions in Figure 4.1.

We would likely find that the measures clustered around a mean value, in this case 3 mm, but many are slightly smaller or larger than 3 mm and a few are much smaller or much larger. Statisticians call this kind of bell-shaped curve a *normal* or *Gaussian* distribution, and lots and lots of biological variation looks like this. As well as a mean, every Gaussian distribution has a **standard deviation** (SD), which is a measure of how spread out the data is around the mean. In Figure 4.1 you can see that both distributions cluster around the same mean, but the top distribution is more spread out around the mean value (it has a higher SD, or is more variable). The smaller the SD, the more tightly the data are clustered around the mean and the less likely we are to see extremely high or low values. If you need convincing of this, look at the two distributions again. We can ask in each case, if we pick an egg at random, how likely is it that it will be thicker than 4 mm or thinner than 2 mm (that is, more than 1 mm different from the mean)? To answer this question we need to know what percentages of egg shells in the two distributions are greater than 4 mm or less than 2 mm. We have shaded the sections of the two distributions corresponding to these egg shells. You can see that the shaded area on the top distribution is larger (and so corresponds to a larger fraction of egg shells) than those on the bottom distribution, so the more variation among our egg shells, the more likely we are to pick an unusually thick or unusually thin egg shell in our sample by chance.

We could then do the same for eggs from hens on the GM diet. The pattern of variation will probably look much the same, indeed if feed type actually has no effect on shell thickness it will look exactly the same. However, if feed type really does have an effect on shell thickness, our distribution would be clustered around a different mean. Figure 4.2 illustrates these two scenarios diagrammatically.

In the top panel the null hypothesis is true and GM grain has no effect on shell thickness. This means that the underlying distribution of shell thickness from hens

The **standard deviation** (SD) is a measure of the spread of values around the mean (or average) value. If the distribution of values is a Gaussian shape, then approximately 95% of the values are within two SDs of the mean. So if the mean thickness of shells is 3 mm and the SD is 0.5 mm, then roughly 95% of shells will have a thickness between 2 mm and 4 mm.

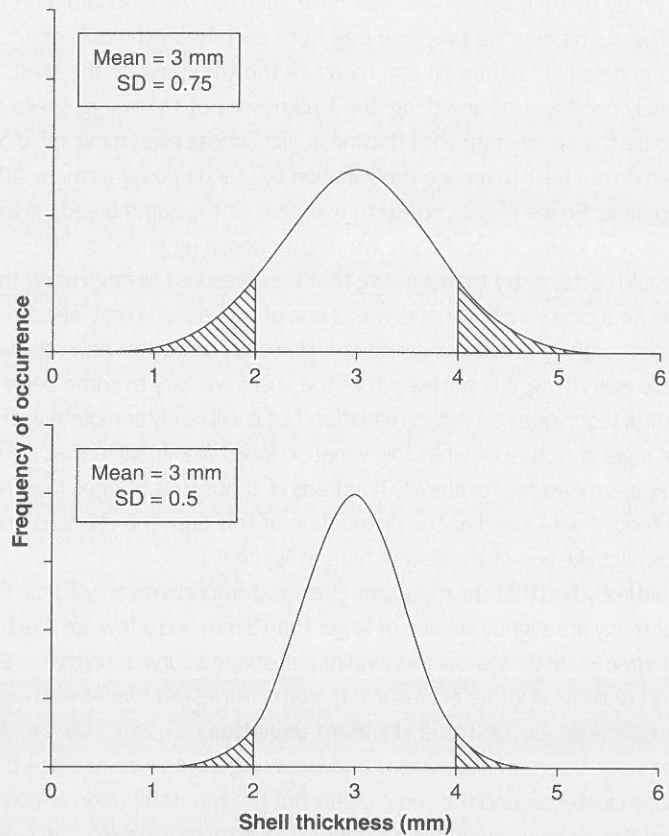


Figure 4.1 Two Gaussian or normal distributions to illustrate the kind of variation we are likely to see if we were to measure the shell thickness of many hundreds of thousands of eggs from hens fed a standard diet. Both distributions cluster around the same mean value (3 mm), but the distribution in the upper panel has a larger standard deviation, because the variation of the values around the mean value is greater. In the upper distribution, approximately 18% of the eggs have shells that are thicker than 4 mm or thinner than 2 mm, whilst in the lower distribution this is only 5%.

fed the two different diets is exactly the same. In the lower panel the null hypothesis is false and the GM grain does reduce the average thickness of egg shells by 0.2 mm. The distribution of egg shells for the hens on the GM and standard diets are still both Gaussian. The distributions also have the same SDs, as the type of food does not affect the underlying variation of shell thickness within experimental groups. However, the two distributions now have different means. We can think of these distributions as representing populations. One population we might call 'All the egg shells of hens fed a standard diet', or 'standard eggs' for short, whilst the other we might call 'All the egg shells of hens fed a GM diet', or 'GM eggs'. If the researcher

could measure every shell in these populations then they could simply say whether their means differed without resorting to any statistics. However, in general, and for obvious reasons, it is not usually possible to carry out that kind of study. Instead we carry out experiments where we examine relatively small samples drawn from the populations that we really want to know about and use the information from these samples to try to say something about the populations that they come from. In this case we would measure samples of eggs from hens fed either standard feed or GM feed and use these samples to estimate the means of the populations they are drawn from. We can then use an appropriate statistical test to examine whether the difference between the means of our samples suggests that the populations that they represent have the same mean as each other or different means.

To illustrate the importance of replication in this process, we are now going to carry out some imaginary experiments using a computer and a random number generator. We will start by assuming that we live in a world like the lower panel in Figure 4.2 and the null hypothesis is actually false, because GM feed does affect

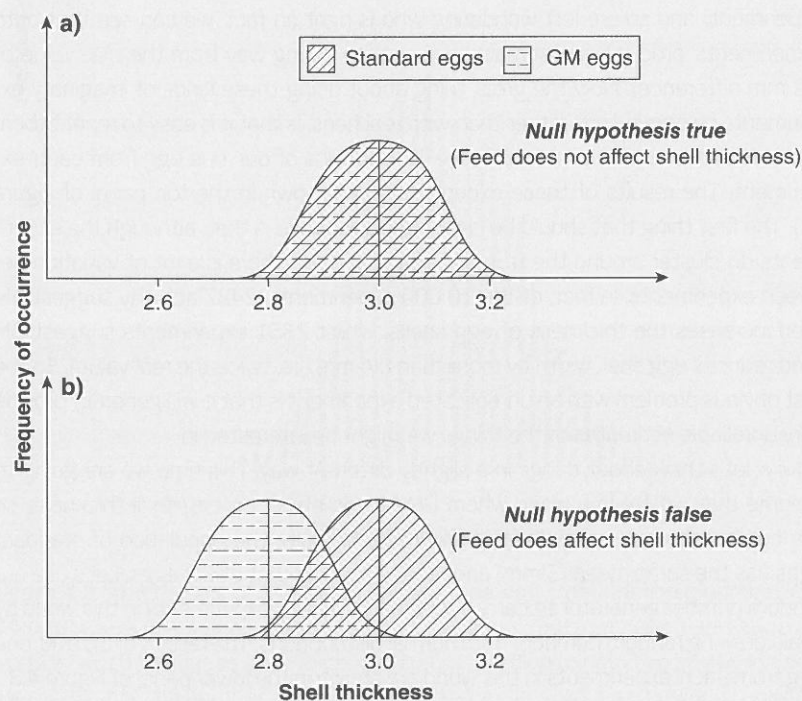


Figure 4.2 The distributions of egg shell thicknesses under two different scenarios. Under both scenarios shell thicknesses follow Gaussian distributions but in the top panel, the two populations have the same mean because the treatment is having no effect on shell thickness (i.e. the null hypothesis is true), whilst in the bottom panel the treatment is affecting shell thickness (i.e. the null hypothesis is false) and the means of the distributions differ.

the thickness of shells, reducing them by 0.2 mm on average. Now let's assume you have ignored all our advice on replication and compare only a single standard egg to a single GM egg. To recreate this experiment on our computer we will draw a single random number from a normal distribution with a mean of 3 mm and a SD of 0.2 mm. This will be the measure for our standard egg. We will then repeat the process to obtain a measure for our GM egg, but this time drawing from a normal distribution with the same SD but a mean of 2.8 mm. The data from our experiment is standard egg = 3.17 mm and GM egg = 2.85 mm, suggesting that the GM feed is reducing egg shell thickness by 0.32 mm. However, now let's imagine our colleague does the same experiment in the same way (computer, two more numbers please!), and their data is standard egg = 2.80 mm, and GM egg = 2.77 mm. These values suggest an effect of GM of -0.03 mm, much smaller than the value that we got in our experiment. Because we know what the populations really look like, we can see that this is because our colleague has measured a relatively small standard egg. However, when we are doing experiments for real we do not have that luxury of knowing the underlying population distribution (indeed, that is what we want to deduce from our experiment) and so are left wondering who is right (in fact, we can see that both 'experiments' produce an estimate that is quite a long way from the true value of 0.2 mm difference). Now the great thing about doing these kinds of imaginary experiments on computers, rather than with real hens, is that it is easy to repeat them many, many times. So let's now create 10 000 runs of our 'one egg from each' experiment. The results of these experiments are shown in the top panel of Figure 4.3. The first thing that should be clear from this figure is that, although the experiments do cluster around the true difference of -0.2, there is a lot of variation between experiments. In fact, of the 10 000 experiments, 2427 actually suggest GM feed increases the thickness of egg shells, whilst 2351 experiments suggest GM feed reduces egg shell width by more than 0.4 mm (i.e. twice the real value). So the first obvious problem with an unreplicated experiment is that it will generally provide very unreliable estimates of the things we might be interested in.

Now let's think about things in a slightly different way. This time we are going to assume that we are in a world where GM has no effect on egg shell thickness, so the null hypothesis is true. Putting this more formally, the population of standard eggs has the same mean (3 mm) and SD (0.2 mm) as that of GM eggs. Let's use our random number generator to carry out some imaginary experiments in this world by again drawing random numbers from normal distributions. The results of 10 000 'one egg from each' experiments in this world are shown in the lower panel of Figure 4.3.

Again, this histogram is centred on the true value (in this case zero, because the populations have the same mean), but is highly variable, with some of our experiments suggesting very large effects of GM (both positive and negative), when we know there is really none (remember, our samples have come from distributions with the same mean). To put this in context, the result that we got in our first experiment when the null hypothesis was false suggested that the effect of GM might be as big as -0.32. However, looking at Figure 4.3 it is very clear that many of our imaginary

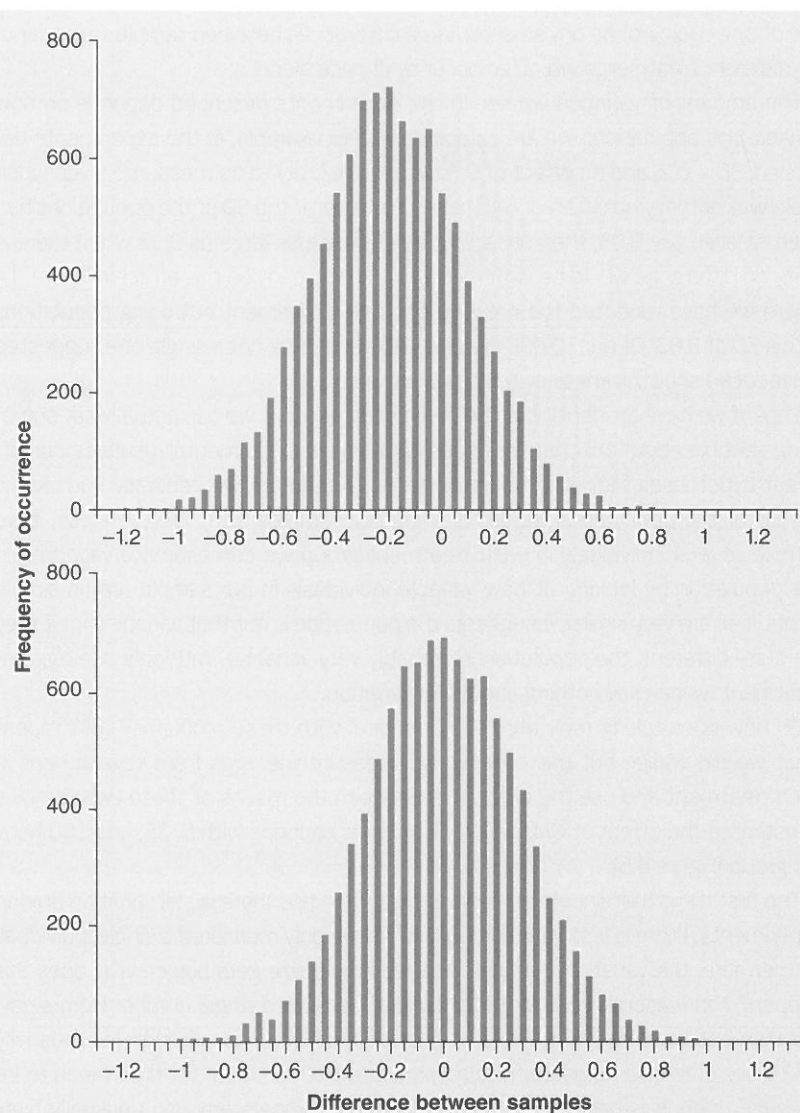


Figure 4.3 The results of 10 000 imaginary 'one egg from each' experiments in a world where GM has no effect on shell thickness (bottom panel) and another world where GM reduces shell thickness by 0.2 mm (top panel). The height of each bar represents the number of experiments that estimated a particular difference (rounded to the nearest 0.05 mm). In both worlds, the estimates from the experiments cluster around the true difference between the population means in that world, but in both cases there is considerable variation among experiments.

experiments generated differences as big as -0.32 , even when there really was no difference between the two populations. In fact, 1274 of our 10 000 imaginary experiments produced a difference as big as this. With small samples (and our sample

size of one could not be any smaller), large differences between samples experiencing different treatments will often occur by chance alone.

The amount of variation we see in the experiments described depends on how variable the populations we are sampling are. For example, in the experiments described, $SD = 0.2$, and an effect of 0.32 was quite likely to be measured, even when there was actually no difference in means. However, if the SD of the populations had been smaller, say 0.05 , then observing a difference as large as 0.32 when there is really no difference would be extremely unlikely. You can see this effect in Figure 4.4 where we have repeated the previous thought experiment but using populations with a SD of 0.05 . Of our 10 000 imaginary experiments, not a single one suggested GM reduced shell thickness by 0.32 mm or more.

Thus, if we have an idea about the underlying variation, we can actually say something sensible about our chances of seeing different experimental results under different hypotheses. However, when we do an experiment we generally don't know the SD of the populations we are drawing our samples from. Nevertheless, if we sample several individuals in each treatment group we can estimate variability in the population by looking at how variable individuals in our sample are (in simple terms: if all are very similar, it suggests the population is not that variable; but if they are quite different, the population is probably very variable). With only one egg per treatment we can say nothing about this variation.

So how does taking replicate measures deal with these problems? Let's repeat what we did earlier, but this time we will measure the eggs from several hens in each treatment and use the difference between the means of these two samples to estimate the effect of GM. Let's look at what happens with 5, 25, and 100 hens per group (Figure 4.5).

The first thing that should be apparent is that, while there is still variation among experiments, there is less variation than when we only measured a single individual. Furthermore, the variation gets less as the sample size gets bigger. Why does this happen? The reason is easy to understand. If we draw a single number from a normal distribution there is a reasonable chance it will be unusually high (or unusually low). If we draw five numbers from a normal distribution, then for their mean to be unusually high, it is necessary that most of the numbers are also unusually high, which is less likely to occur. If we draw 25 numbers, then the chance of the mean being unusually high becomes even less. Another way to think of this is that when we draw several numbers and calculate a mean, whilst some of the numbers will be likely to be unusually high, these will be cancelled out by other numbers being unusually low. In practical terms this implies that our estimates of the population means become more reliable as our samples get bigger, and the difference between sample means becomes a more reliable indication of any difference between population means.

Similarly, in a world where GM has no effect, we have already seen that experiments based on single samples will often show large differences, simply due to chance. As the sample size increases the variation decreases, and the chance that

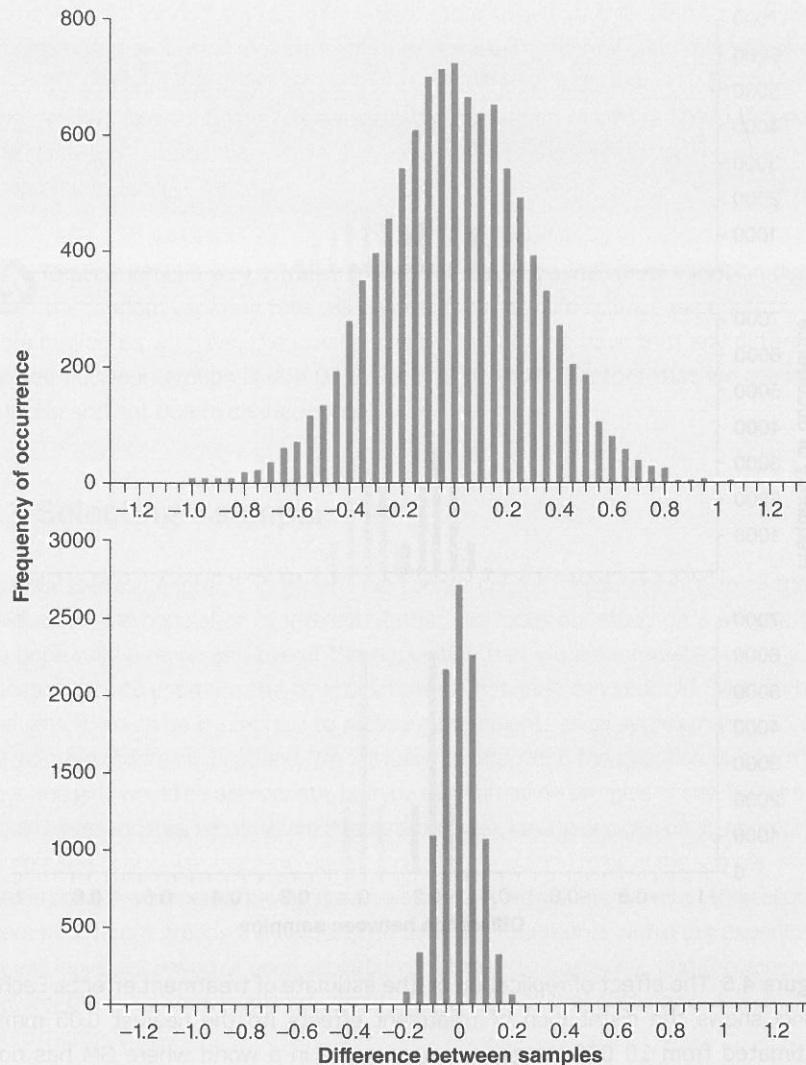


Figure 4.4 The effect of variation in egg shell thickness on experimental estimates. Estimates of the effect of feed type on shell thickness from 10 000 single individual experiments. In each experiment, the effect of GM is estimated by subtracting the thickness of the standard egg shell from the thickness of the GM egg shell. The height of each bar shows the number of experiments that gave a particular effect size (rounded to the nearest 0.05 mm). In both cases there is no effect of GM, and the distributions are centred on zero, but in the top panel the underlying variation in egg shell thickness is greater than in the bottom panel.

we will see a large difference between the means of our treatments if GM really has no effect gets smaller and smaller.

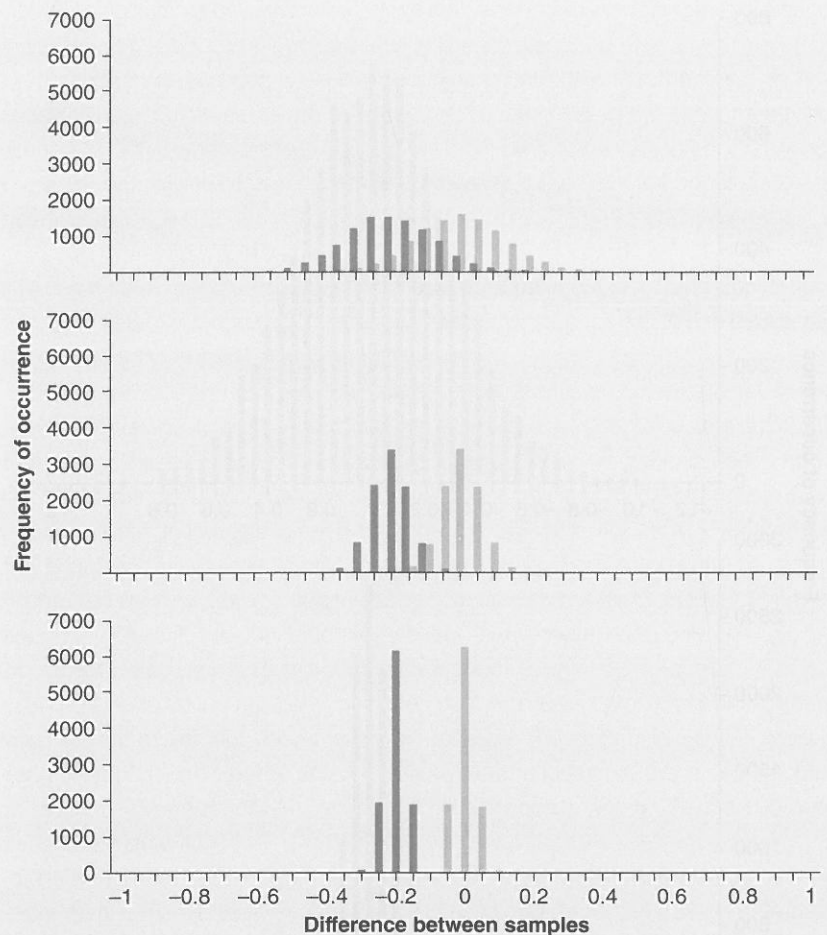


Figure 4.5 The effect of replication on the estimate of treatment effects. Each panel shows the distribution of treatment effects (to the nearest 0.05 mm) estimated from 10 000 imaginary experiments in a world where GM has no effect on eggs shell thickness (light bars), and in an alternative world where GM reduces shell thickness by an average of 0.2 mm (dark bars). In each experiment the effect of GM is estimated by subtracting the mean thickness of the standard egg shell in the sample from the mean thickness of the GM egg shell in the other sample. The sample size of the experiments increases from top to bottom, with five hens per group in the top panel, 25 in the middle, and 100 in the bottom. As sample size increases, the estimate of the treatment effects becomes more precise, and we see fewer experiments yielding extreme estimates. In practice, this means that it becomes easier for us to discriminate between these two worlds from the result of a single experiment.

So replication has two important effects on our experiment; by taking several measures and then taking a mean, some of the random variation is effectively

cancelled out making any real differences easier to see. At the same time, experiments based on several replicates make extreme experimental outcomes less likely to occur by chance, and so if we do see a large difference between our treatments, we can be more confident it is because the population means really do differ. Finally, having replicate measures allows us to estimate the variation in the underlying populations too.



Replication is a way of dealing with the between-individual variation due to the random variation that will be present in any life science experiment. The more replicates we have, the greater the confidence we have that any difference we see between groups is due (at least in part) to the factors that we are interested in and not due to chance alone.

4.3 Selecting a sample

As discussed in Chapter 1, in general we cannot collect measurements on all the individuals in the population of interest. Rather, we focus our study on a sample that we hope will be representative of the population that we are interested in. So if our question was to estimate the height difference between ten-year old Scottish boys and girls, it would be impractical to measure the heights of all approximately 75 000 10-year-old children in Scotland. We will leave to Chapter 6 the question of how many boys and girls would be appropriate, but you might imagine samples of say 100 of each would be reasonable. What we are interested in here, having decided on an appropriate sample size, is how we select individuals from the population to be in the sample. Notice that in some experiments, having drawn a sample of individuals, we must then allocate them to different groups that experience different treatments within our experiment; we will leave discussion of such allocation strategies to Chapter 7, and focus here on drawing the initial sample. We will describe a number of the most commonly encountered methods.



Some further methods (specifically snowball sampling, systematic random sampling, and proportional quota sampling) can be found in Section 4.3 of the supplementary information. Go to: www.oxfordtextbooks.co.uk/orc/ruxton4e/.

4.3.1 Simple random sampling

The simplest method of selecting a sample is called *simple random sampling*. Imagine that from some government census data we obtain a list of all the ten-year-old girls in Scotland: 37 467 of them. We simply put them in a spreadsheet and assign a number from 1 to 37 467 to each. To obtain a simple random sample of 100 of these we then ask the computer to select a random number from 1 to 37 467 where each number has an equal probability of being selected. Let's say that number is 21 456; the girl with that number then becomes the first person in our sample. We repeat this process until we have 100 unique numbers (i.e. we have no duplicates), and so we have identified 100 girls chosen entirely at random and independently of each other. That is our simple random sample.

If your population is smaller then you can get a random sample without using a computer. Imagine we have a herd of 57 cattle and we want to sample 15 of them. Each cow has a unique identification number embossed on a tag on its ear. We simply write each of these numbers down on 57 identical pieces of card, drop the pieces of card into a bag, stir the cards round with our hand for a bit, then pull out 15. This should give a simple random sample. The key thing for the two techniques described above is there should be no way the Devil's Advocate can be concerned about unconscious bias in which individuals are selected for our sample. You could simply make a list of the 57 cows on a sheet of paper, close your eyes then stick your finger onto the piece of paper and read-off the ID number nearest to your finger. You could repeat this until you had 15 unique numbers. This is marginally less work than the 'card in a bag' approach but it introduces risk of bias. Subconsciously if you stuck your finger high on the list of numbers one time, you are likely to go lower on the list next time. This is not in the spirit of random sampling. The key demand of random sampling is that at any point during the selection of your sample every individual who might be added to your sample next (that is who hasn't been picked already) has an equal probability of being the next one picked. That is true of the first two methods we describe, but in this last method it is not because we think there is a risk that after a number from low on the list has been selected the next number selected is more likely to be a high number through your unconscious effort to spread your pointing around the list. So the secret to good random sampling is to remove human factors from the selection process so there can be no risk that someone accuses you of bias in your sampling.

Notice you can always do simple random sampling as long as you can list all members of the population that you want to sample from. In some cases, such as herd of 57 cows, this requirement is pretty easy to meet. But the example of all ten-year-olds in Scotland is an example of a case where this will be less straightforward. Census data of the ages of all residents of Scotland probably exists in some government department; but obtaining access to this data may be difficult or even impossible. If it is possible then this census data will have been collected some time ago and so will not be completely accurate; you have to weigh up how important this inaccuracy is. For example, imagine the census occurred 12 months previously, obviously all the children who were nine then will be ten now; but there will be some children who have died or left the country and some other children who have come to the country only in the last 12 months. However, these unusual cases will likely only lead to a very small inaccuracy in using the 12-month-old census and you can probably defend this approach to a Devil's Advocate.



In a *simple random sample* we list all members of the population of interest, then independently and randomly select a fixed number of them to be in the sample; all have equal probability of being included.

4.3.2 Remember, you want a representative sample

You want a representative sample, and most but not absolutely all random samples are representative. Imagine in the herd of 57 cows there are 20 three-year olds,

19 four-year-olds, and 18 five-year-olds. Imagine now that by chance we notice that our random sample of 15 individuals contains 14 three-year-olds and 1 five-year-old. This seems a very uneven distribution across the age range—is this a problem? Possibly. If we think that the traits that we want to measure on our sample of cows are likely to be influenced by age then the fact that our sample age structure is very different from the age structure of our population of interest is a concern. In this case we should simply reject this sample and draw a fresh sample of 15 that is very unlikely to have such an unusual age distribution. However, such situations will very, very infrequently occur; almost all random samples will be representative. The larger the sample the less likely it is that a random sample will be unrepresentative. You might go through your whole scientific career and never need to reject a random sample on the grounds of it being unrepresentative.

However, eyeball your sample to check that it is representative before you actually carry out measurements on the sampled individuals. One reason to do this early in the process is to save the time, effort, expense, and animal suffering of taking measurements that you will throw away. The other reason is to avoid accusations that you are resampling because the first sample was inconvenient rather than unrepresentative. Imagine that you measured the traits of interest on the 15 cows in your sample, then carried out statistics on your data and got an answer different from the one you were expecting; you then noticed this issue of unusual age distribution and decided to repeat your experiment with another sample. The Devil's Advocate will surely say that you would not have resampled if you had gotten the result you wanted the first time around, so you are repeatedly sampling until you get the result you want. You can see that this is very poor scientific practice—in fact some people would call it scientific misconduct. One way to avoid an unrepresentative sample would be to use stratified sampling, which we'll discuss next.



You aim to have a sample that is representative of the population. Almost all random samples are, especially for larger samples, but very, very occasionally you can get an unrepresentative random sample. Throw it away and pick a new one.

4.3.3 Stratified sampling

Let us return to the example of the herd of 57 cows. Imagine that we had considered before sampling that the traits we are interested in are likely to be strongly age-dependent, and so we very much want the age structure of our sample to reflect that of the statistical population (the herd). We can do that by stratified sampling. We note that the cows fall into three age classes and that each of these has approximately one third of the cows. We then divide our population up into these three strata (three age classes) and use simple random sampling on each of these three strata independently. That is, to get our sample of 15 we draw a sample of five by simple random sampling from the group of 20 three-year-old cows, five from the group of 19 four-year-olds, and five from the group of 18 five-year-olds. This is a little more work to organize but

ensures that, for the trait that we wanted to stratify on (age), the distribution across the sample will be similar to the distribution in the underlying population.

You can stratify on any trait you can measure on subjects, not just age. Consider the case of sex differences in ten-year-olds in Scotland. Here we might have considered that ethnicity might be important to height. The same census data discussed above might also give us the stated ethnicity of each ten-year-old. We could then decide to stratify on that ethnicity. Imagine the distribution of different ethnicities in ten-year-old children was 96.0% White, 2.7% Asian, 0.7% Black, 0.3% Mixed, 0.2% Arab and 0.1% Other. If we were sampling 1000 girls we could split the girls into groups (strata) by ethnicity and for a sample of 1000 select 960 from the 'White' group, 27 from the 'Asian' group etc. This would ensure that the distribution of ethnicities in our sample is similar to that of the underlying population.

In the case above where ethnicity will not have a huge effect on height (Black ten-year-olds will not be twice as tall on average as Asian ones) and where there is one dominant stratum (96% are White) then there is not a strong need to stratify. However, if you stratify correctly then there is no drawback to stratified sampling except for the extra effort involved. Where stratification might be very valuable is if a small fraction of the population might be very influential. For example, imagine that we were sampling positions in a large forest where we will place pitfall traps in order to measure terrestrial invertebrate biodiversity (see Figure 4.6). The margins of the forest might be a relatively small fraction of the overall land cover of the forest but might be very high in biodiversity compared to the interior. By stratifying we can make sure that our random sample includes an appropriate (small) fraction of these sites that we think might be uncommon but influential. If we selected sites by simple random sampling we might not select any sites near the margins and so might not have the ability to evaluate this part of the study population that we think might be important.

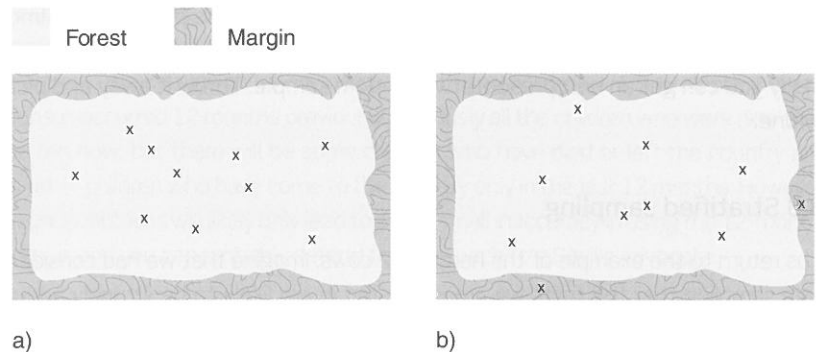


Figure 4.6 Comparison of random and stratified sampling schemes. If we place traps at random as in a) then it is quite likely that none of our 10 traps end up in the 20% of the forest that we classify as margin habitat. With stratified sampling, we allocate 2 of our 10 traps randomly in this marginal area and the remaining 8 in the interior as in b). This maximizes the chance that our traps provide a representative sample of the invertebrates in the whole woodland.

It is possible to stratify on more than one variable. In the 'Scottish children' example we could stratify on geographical area as well as ethnicity, so that of the 960 White girls we ensure that (for example) 8% of those come from the city of Edinburgh.

Stratified sampling requires that you have good information on the distribution of the variable that you want to stratify on. There is no danger in stratifying carefully on a variable that turns out not to be influential—if ethnicity turns out to not be a good predictor of the height of ten-year-olds then our sample is not any less valid because we chose to sample by stratifying on that variable. But, if the variable is influential and we stratify inappropriately then we can end up with a biased sample. Say that marginal sites in the forest in fact represent 7% of the land area, but we mistakenly thought that they made up 17% of the land area: if those sites end up making up 17% of our sample sites then our sample is not a good representation of the forest, because too much weight is given to marginal sites. Thus, you should only think of stratified sampling when (i) you have identified an influential variable that you would like to make sure is appropriately represented in the sample; and (ii) you have good information on the distribution of that variable across the statistical population. If not, then stick to simple random sampling.



We use *stratified sampling* to ensure that the sample accurately reflects the population for a trait (or traits) that might be highly influential on the qualities we want to measure on subjects in our sample.

4.3.4 Cluster sampling

Imagine that we had used simple random sampling or stratified sampling to identify 1000 girls and 1000 boys in Scotland to be in our sample. These individuals will be scattered right across the country. Collecting their heights will take a lot of organization and a lot of driving. With a bit of organization by phone we can arrange to measure geographically close children on the same day. Having measured one child, we get into the car and program our satellite navigation device to help us drive to the next individual; we can then sit in the car and wait with our flask of coffee until the time we have arranged to measure them. Overall, we will spend a lot more time driving and drinking coffee than we will measuring children. This may be worth the effort, but sometimes we can still get a good sample and cut down on organization and logistics by sampling not single individuals but natural clusters. This is the essence of cluster sampling. It doesn't have any theoretical advantages over simple random sampling, but it can sometimes give you just as good a sample at much reduced effort. But it requires that individuals fall into natural clusters.

Ten-year-old Scottish children do fall into natural clusters. They can naturally be grouped into schools. On average a primary school might have 30 ten-year-old children so rather than randomly selecting 2000 individual children from across Scotland in a simple random sample, in cluster sampling we might use simple random sampling to select 70 primary schools and measure all the children in each school. This will greatly reduce the hassle of data collection: we will have to do a lot less driving. Is this still a

representative sample of Scottish ten-year-olds? Probably. The average height of girls in a school will be affected by things like the ethnic and socioeconomic breakdown of the local area, so we need to be sure that we average across this variation; by sampling 70 schools we might reasonably expect to do this. But we must also be careful that when we adopt cluster sampling we do not subtly but significantly change the population from which we sample. By sampling at school level we will no longer include children who are taught at home rather than at school in our sample. In fact only 800 children (of all ages) in Scotland are taught outside the school system, so this change in sampled population should not concern us much. But, as a generality, we need to be careful of changing the underlying population (and so changing the specific question we are investigating) when we adopt cluster sampling for reasons of convenience.

It is possible to imagine hybrid sampling schemes that involve layers of sampling. For example, if we decided that 2000 children felt like a good number to aim for but 70 schools was a little small then we could select 100 schools then within each school use simple random sampling to select ten ten-year-old boys and ten ten-year-old girls. Notice that very small schools might have fewer boys and girls in total than you need for your sample, and you might decide either just to sample all of them (and visit more schools); or to avoid such schools. The best choice will depend on trading off practicality with any concern of introducing bias by excluding some types of school. This question of whether to sample more schools or more pupils within each school is another manifestation of the concept of subsampling discussed in Section 11.2.1.

Imagine that you wanted to ask a different research question that required you to allocate children in your survey to different treatments. Say you might be interested in investigating which of two methods was most effective at teaching children a foreign language; or you might be interested in which of two toothpastes was more effective at preventing tooth decay in children. You could either allocate whole clusters (whole schools) to a treatment, or allocate individuals to a treatment. This decision will come down to practicality and risk of interaction between individuals. For the teaching method it would be much, much easier to implement if allocation was done at a whole-school level and all of the children in one school were allocated to the same method. This would also help with avoiding concerns stemming from interaction between individuals. Since the children will interact with each other outside the classroom, sharing knowledge, we can see that if half the children in the school were taught one method and half the other, then the observed difference between children on the two schemes might well be reduced because when the children interact with each other the children experiencing the better teaching method help raise the performance of those on the other method. However, for the toothpaste study we think the chance of children sharing toothpaste is pretty low and so each child can be more readily considered as independent, and there is no great organizational challenge associated with assigning individual children in the same school to different toothpaste types (particularly if clever design of packaging (see Section 2.3.1 on placebos) reduced obvious differences in appearance between the two types).

Lastly, we recommended that clusters (schools) be selected by simple random sampling where each school has an equal probability of being in the sample. You will

also see cluster sampling implemented where the probability of each cluster being in the sample is weighted by the size of the cluster. In our case that would strictly be by the number of ten-year-olds in the school, but the total number of children at the school is probably easier to obtain information on and would do just as well. Weighted sampling is a little trickier to implement but by weighting by cluster size you can see that we get back to the underlying assumption of simple random sampling of individuals, that all individuals have an equal probability of being in the sample. We mention this idea of weighting only because you might come across it in published papers. For your own work we do not see a compelling reason to add this extra mathematical complexity (and thus extra room for a mistake) to your method of sampling.



Cluster sampling can be a good way to get a good sample at reduced effort when individual subjects are naturally grouped into clusters.

4.3.5 Convenience sampling

Especially early in your scientific career, when you don't have a lot of time and money to devote to projects, you will often find yourself doing another type of sampling called *convenience sampling* (also called *accidental*, *grab*, or *opportunity sampling*). Bluntly, this sort of sampling is not as methodologically pure as those already discussed, but (as the name suggests) it is often just an awful lot more convenient. It might not be as appealing to the purist as the methods discussed already, but it can still be done really well (or really badly!) You would like to estimate the difference between the heights of ten-year-old Scottish boys and girls, but you simply don't have the resources to travel across the country and conduct sampling in any of the ways discussed previously. This doesn't mean that you cannot carry out a useful study. It occurs to you that there is a local zoo that attracts families and school groups from across Scotland (and beyond), and you approach this zoo and obtain permission to try and recruit children into your study. At the zoo you station yourself somewhere where visitors pass by, and when a group with at least one child that you think might be ten years old passes by you approach an adult in the group (likely a parent, school teacher, or youth group leader) and explain the nature of your study and whether the child might be appropriate for your study (correct nationality and age). If they are suitable and all appropriate parties are willing then you add this child to your sample. In doing this you can quite quickly build up a substantial sample that will let you explore the question of interest to you.

However, two issues you must be aware of are *selection bias* and *external validity*. It can be easy to introduce selection bias in convenience sampling and you must guard against even the appearance of this. Selection bias occurs when some suitable subjects are more likely to end up in your sample than others in a way that biases your results. Set yourself careful protocols for selection that will help you avoid selection bias and explain to others that you have avoided bias. In this case, the issue is that you are going to approach groups that might have a child who is ten with a view to recruiting that child. The real danger for this study is that your assessment of whether a child looks like they might be ten could in part be based on their height, the very factor you are interested in measuring. It would be very unfortunate if you let suitable subjects

pass by unrecruited because you guessed they were younger than ten because they are unusually short for a ten-year-old. The way around this would be to approach any group where it is even slightly possible that a child might be ten, so you steel yourself to approach any group whether at least one child looks as if they are definitely older than five but less than fifteen. This will cause you extra effort and occasionally some embarrassment, but will help you to avoid bias.

We should mention another type of bias related to humans: *recruitment bias*. Until now we have assumed that all individuals approached to be in your study agree to be in the study, or equivalently that there is no relation between the trait we are interested in (height) and willingness to be in the study. If we break this assumption then we have recruitment bias. Here we have to be careful; you could imagine that a child who is self-conscious about their height because they perceive it to be unusual might be less willing to take part in the study. Again, it is worth thinking through a protocol that can help avoid recruitment bias and demonstrate to others that you have avoided recruitment bias. In this case, it would be worth emphasizing to the potential subject that (i) they were approached because they looked like they might be ten and not because there was anything unusual about their height; (ii) that you won't record their name; (iii) that you will measure their height in a way that means other people will not see the value; and (iv) you will never disclose their value for height to anyone. There is no guarantee that such steps will entirely eliminate recruitment bias but they should help reduce this concern considerably.

External validity is the extent to which the study can be generalized to the population of interest. The previous methods have very strong external validity because all members of the population could potentially be in the sample. In our child-heights example, all Scottish schoolchildren might have been included in the sample; when we use a convenience sample at only one location you have to make more of an effort to make people believe that your study has external validity. In the child-height study it might be worth asking participants about their ethnicity and about where in Scotland they come from. If you can show that your sample does actually have a wide spread geographically, and has an ethnic distribution that is similar to the population of interest (all Scottish ten-year-olds), then this should help convince others about the external validity of the study. External validity is another way of thinking about how appropriately you can generalize from a study—see Statistics Box 3.1 for more on this. (It won't surprise you to find that there is also a piece of experimental design jargon called *internal validity*. This is about the validity of any inferences you make about cause-and-effect relationships, and so is about avoiding confounding factors—see Sections 1.4.2 and 3.1.3.)



Particularly in studies involving humans, there is a blizzard of terminology associated with biased sampling: selection bias, volunteer bias, non-response bias, ascertainment bias, detection bias, information bias, response bias, assessment bias, observer bias, recall bias, allocation bias, and attrition bias. See Section 4.3.5 of the supplementary information for brief definitions if ever you encounter any of these terms in other reading. Go to: www.oxfordtextbooks.co.uk/orc/ruxton4e/.



Convenience samples are much more convenient to obtain, and can still be useful if you take care in the design of your data collection.

4.3.6 Self-selection


Self-selection is a real limitation to the usefulness of the phone polls beloved of newspapers and TV stations. Generally, people are encouraged to phone one number if they, for example, prefer Coke to Pepsi or another number if they prefer Pepsi. Now, what can


you conclude if the Pepsi number receives twice as many calls as the Coke number? First you must think about the population that is sampled. The population that sees the newspaper article with the phone number is not a random cross-section of the population of people who drink cola. People do not select newspapers in a newsagent at random, so immediately you are dealing with a non-random sample of the population. But things get worse. People who read the newspaper do not all respond by picking up the phone, indeed only a small fraction of readers will do so. Can we be sure that these are a random sample of readers? In general, this would be very risky; generally only people with a very strong opinion will phone. So we can see that these phone polls are very poor indicators of what the wider public believe, even before you ask questions about whether the question was asked fairly or whether they have safeguards against people voting several times. Our advice is to be very careful when interpreting data that anyone collects this way, and, if you are sampling yourself, make sure you are aware of the consequences of this sort of self-selection.


In human studies a certain amount of self-selection is almost inevitable. If you decide to ask random people in the street whether they prefer Coke or Pepsi, then people have the right to decline to answer, hence they self-select themselves into or out of your study. Your experimental design should seek to minimize the extent of self-selection, and then you must use your judgement to evaluate the extent to which the self-selection that remains is likely to affect your conclusions.

Self-selection is not confined to human studies. For example, if we are seeking to measure the sex ratio of a wild rodent population then it would seem logical to set out live traps and record the sex of trapped individuals before releasing them. However, this will give an estimate of the sex ratio of that subset of the rodent population that are caught in traps, not necessarily the general population. It could be that some sections of the population are less attracted to the bait used in the traps than others or less likely to be willing to enter the confined space of the trap than others. Hence, the estimate of sex ratio that we get is affected by the fact that individuals may to some extent self-select themselves into our study by being more likely to be trapped. If an individual's likelihood of being trapped might be related to its sex then the sex ratio of the trapped sample might be quite different from that of the wider population.

Let's now move on to the next chapter, where we examine replication a little more closely so as to warn you of some potential pitfalls.

 **Q 4.2** What do we mean by asking the question 'fairly'? Can you give examples of unfair ways to ask the question?

 **Q 4.3** What can you do to explore how serious self-selection is for your estimate of sex ratio in the rodent population?

 Sometimes some self-selection is unavoidable, but try to minimize it, then evaluate the consequences of any self-selection that you have left.

■ Summary

- Whenever we carry out an experiment, we are trying to find ways to remove or reduce the effects of random variation, so that the effects that we care about can be seen more clearly.
- Replication involves making the same manipulations and taking the same measurements on a number of different subjects.

- Replication is a way of dealing with the between-individual variation due to the random variation that will be present in any life science situation.
- In the life sciences, we are almost always working with a sample of experimental subjects, drawn from a larger population. To be independent, a measurement made on one individual should not provide any useful information about the measurement of that factor on another individual in the sample.
- In a simple random sample we list all members of the population of interest, then independently and randomly select a fixed number of them to be in the sample—all have equal probability of being included.
- Simple random sampling (especially for larger sample sizes) almost always gives a representative sample. But once or twice in your career you might have to reject a sample as unrepresentative.
- If you want a sample to be highly representative of the population distribution of a given trait then stratified sampling can achieve this.
- If individual experimental subjects naturally cluster then selecting clusters rather than individuals might be a lot more convenient.
- Often we end up convenience sampling. Generalizing from a convenience sample can be a challenge, but you can meet that challenge with careful sample design.