Beyond complete randomization: blocking and covariates

9

Imagine we want to explore how a new drug affects blood pressure in humans, and we have a sample of people enrolled in our trial that we want to allocate to either the new drug or a placebo. There is nothing wrong with allocating individuals randomly to one of the two groups as discussed in Chapter 7. There are many confounding factors that might be linked to someone's blood pressure (e.g. age, sex, body mass index (BMI), diet, or occupation), but our randomization will average these out across the two groups. However, all these factors will add between-individual variation to our study, and so we may have to have quite a large sample size to get the statistical power that we want (see Chapter 6). If we don't have a good grasp of what confounding factors are likely to cause most of the between-individual variation (or the important confounding factors are difficult to measure on our subjects), then randomization or within-subject designs (discussed in Chapter 10) are our best approaches. However, if (for example), based on previous work, we expected that BMI might have a particularly strong influence on blood pressure, then alternatives might be open to us. We could reduce noise in our experiment by only using individuals from within a narrow range of BMI values, but this will affect how we can generalize from our study. If we only use subjects from the 'healthy' range of BMI values, then we lose our ability to inform about the effects of the drug on overweight and underweight people. Alternatively, we can use BMI as a blocking factor or covariate, and this chapter will introduce you to these techniques.

- We begin by explaining the concept of blocking in experimental design (Section 9.1).
- We then explore the range of things that you might be interested in blocking on (Section 9.2).
- We look at the potential costs of blocking to help you decide whether it is an attractive design for a given experiment (Section 9.3).
- We look at paired designs as a special form of blocking with only two subjects in a block (Section 9.4), and consider more generally how you might best select the size of blocks (Section 9.5).
- Finally, we introduce covariates (Section 9.6), and consider interactions involving covariates (Section 9.7).

9.1 The concept of blocking on a particular variable

If a given variable is expected to introduce significant variation to our results, then we can control that variation by **blocking** on that variable. This involves splitting our experimental subjects up into blocks, such that each block is a collection of individuals that have similar values of the blocking variable. We then take each block in turn and for each block distribute individuals randomly between treatment groups. It is thus a more complicated alternative to complete randomization. Blocked designs are sometimes called *matched-subject designs*, and hence blocking can be called *matching*.

Suppose that we are interested in whether the type of food we give to a greyhound affects its running speed. We have 80 greyhounds and want to test the effects of four food types. The 80 greyhounds are of different ages, and age is expected to have a strong effect on running speed. We could treat age as a **blocking** factor in the experiment.

First of all, we rank the dogs by age, and then partition this ranking so as to divide the dogs into blocks, so that those in a block have a similar age. This might be 20 blocks of 4 dogs, 10 blocks of 8, or 5 blocks of 16. This choice does not matter too much (but see Section 9.5 for further thoughts on selecting block size), providing the number of individuals in a block is a multiple of the number of treatment groups. We then take each block in turn and randomly allocate all the dogs from one block between the treatment groups, in exactly the way that we did in the fully randomized design. We repeat the same procedure with all the blocks. This approach of using blocks (sometimes called strata) as part of our allocation procedure is called *stratified block allocation* and contrasts with the complete randomization that we described in Chapter 4.

So what does this achieve? One way to think of it is this: in the randomized experiment we will find ourselves comparing an old dog in one treatment and a young dog in another. If we compare their running speeds and find a difference, part of this may be due to food, and part due to age (and the rest due to random noise). In contrast, in the blocked design we are making comparisons within our experimental blocks, so we will only ever be comparing dogs of similar ages. With differences due to age much reduced in any given comparison, any effects due to diet will be much clearer (see Figure 9.1). The blocking has effectively controlled for a substantial part of the effects of age on running speed. You can think of this blocked design as a two-factor design, with food type as the first factor, and age block as the second.

@ See Section 9.1 of the supplementary information for more on allocation strategies, including explanation of the terms *restricted random allocation, block allocation, permuted block allocation*, and *block randomization*. Go to: **www.oxfordtextbooks.co.uk/orc/ruxton4e/**.

Effectively what we are doing is dividing the between-individual variation into variation between blocks and within blocks. If we have blocked on a variable that does have an important effect on between-subject variation, then the between-block differences will be much greater than those within blocks. This is important, because in our final statistical analysis we can use 'blocks' as an additional factor in our ANOVA. This will allow us to test for between-treatment differences within each block (where between-subject differences are small) rather than across the whole sample (where between-subject differences would be larger). Or, as a statistician might say, we can control for variation amongst blocks before looking for differences due to treatment. Thinking again in terms of the questions our design allows us to answer, our randomized block design allows us to answer the following question:

Do food types affect running speed in greyhounds, once age has been controlled for?

Or put in a slightly different way:

Are age-corrected running speeds in greyhounds affected by the food types?

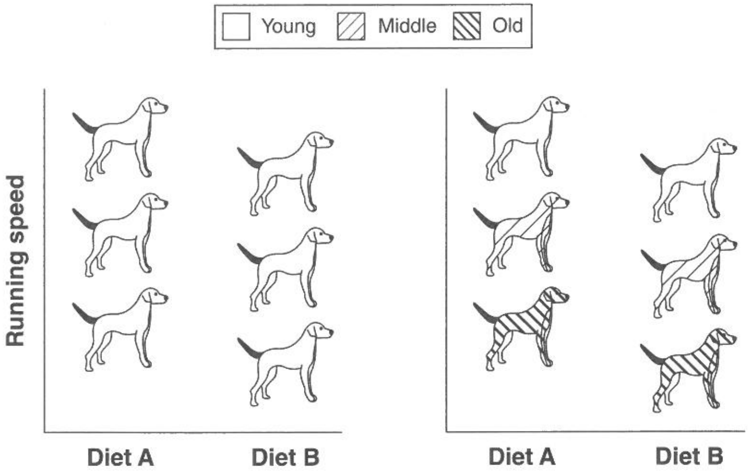

Figure 9.1 The left-hand panel illustrates a fully randomized design to explore the effect of diet on the running speed of dogs. Dogs are simply randomly allocated to one of two groups: one group experiences diet *A*, the other diet *B*. However, if we know the age of the dogs and expect age to have a strong effect on running speed, then we might instead opt for the blocked design illustrated in the right-hand panel. Here dogs are first allocated to three blocks on the basis of their age (young/middle/old). Each of these age blocks is taken in turn and the dogs within that block randomly allocated to either diet *A* or diet *B*. Subsequent analysis will then test for a difference in running speed due to diet between the two groups of dogs of each given age block. If a large proportion of the within-group variation in the fully randomized experiment was due to variation in age, then the blocked design may be more effective at testing for an effect of diet.

And (as should be no surprise to you by now) if we can remove some of the inherent variation from our study, we can increase our statistical power. You can find more on this in Box 9.1.

BOX 9.1 Variation, blocking, and power

Imagine that you are interested in whether the 'song' that a male fruitfly produces with his wing during courtship differs in its characteristics between populations of a given species. You collect individual males from two populations, and measure a particular component of their courtship song, the inter-pulse interval (IPI). This character is easy to measure and known to vary between species. You plan to measure 15 males from each population, but the procedure that you use to measure IPI is time consuming, and realistically you can only measure ten flies a day, and so you are going to have to measure flies over three days. Previous work has shown that courtship song can be affected by factors like atmospheric pressure which are likely to vary between days. Sensibly you decide on a randomized block design where you measure five flies of each population on each day. Figure 9.2a shows the results of the study. There is a suggestion that the IPI of one population is lower than the other, but there is also a lot of random variation amongst flies from the same population.

Figure 9.2b shows the same data plotted in a different way, separating the measurements by the day on which they were made. The first thing that should be clear from this graph is that there is much less variation between flies from the same population when measured on the same day than there is when the data from the three days are pooled (as in the first panel). The reason for this can be seen by

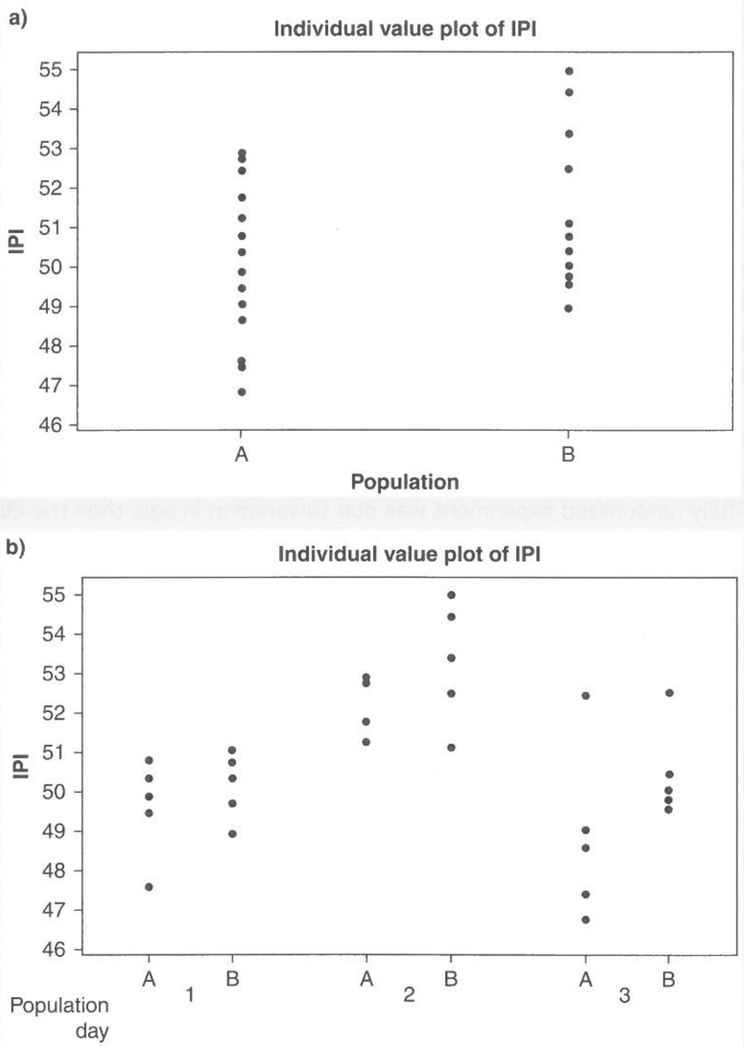

Figure 9.2 a) The IPI of individual flies from the two populations and b) the same data now broken down by day of measurement.

looking at the average IPI of the flies on each day. Overall flies on day 2 seem to sing slightly faster (i.e. with a smaller IPI) than flies on day 1, and the flies on day 3 may sing slightly slower. This day-to-day variation is part of the random variation in the pooled data, and is part of the reason that the data is so noisy. If we use *Day* as a blocking factor in our statistical analysis, then we are effectively removing the day-to-day variation, making any differences between populations easier to see. Or in statistical jargon, we are looking for population differences having controlled for day-to-day variation.

Imagine that we had carried out a power analysis prior to carrying out this study to estimate the power of our planned study to detect a difference in IPI of 1.5. The power of the study using the variation in 9.1a as our estimate of the random variation (that is, carrying out the statistical analysis without the blocking factor) would have given a power of 0.55, whilst the analysis including the blocking factor and using the within-day variation would give a power of 0.84. So if we ignore the effect of the blocking factor, we would expect to detect a difference of 1.5 only just over half of the time, but if we take the blocking factor into account, we increase our chance of detecting the same difference to over 80% of the time.

If you know and can measure some aspect of experimental subjects that is likely to explain a substantial fraction of between-subject variation then it can be effective to block on that factor.

9.2 Blocking on individual characters, space, and time

You can block on any factor (indeed on any number of variables) that you think might be important in contributing to variation between individuals that have been given the same experimental treatment. The only condition for a characteristic to be used as a blocking factor is that you can measure it in individuals, so you can allocate them to groups of individuals that are similar in that trait (blocks). We will refer to this as *blocking by individual characters*.

There is a second type of blocking that is commonly used in biological experiments that we will call *blocking in space*. Imagine that in our tomato experiment discussed in Chapter 8 we find that we can't fit all our plants into a single greenhouse, but instead need to use three different green houses. If there are differences between the conditions in the greenhouses that affect growth rates, then by carrying out the experiment at three different sites we are adding some noise due to site differences. We could simply ignore this variation and use a fully randomized design where every plant is allocated at random to a greenhouse irrespective of its treatment. However, this could be a very inefficient way to carry out such an experiment. We cannot simply have all of the individuals in a given greenhouse getting the same treatment, as this would introduce greenhouse identity as a confounding factor. A common solution to this problem is to block by greenhouse. What this means is that we allocate equal numbers of each treatment to each greenhouse, and note down which greenhouse a plant is grown in. We now have a two-factor design with feeding rate as the first factor, and greenhouse as a second 'blocking' factor. This means we can make comparisons between plants in the same greenhouse, comparisons that will be unaffected by any differences between greenhouses. Blocking by site is frequently seen in large agricultural trials (where several fields or even farms may act as the blocks), but could equally apply to any experiment where assays are carried out at different sites. Sites can be anything from fish tanks, to shelves in a growth cabinet, to labs in different countries.

The final kind of blocking that we will discuss is *blocking in time*. This is exactly analogous to blocking in space except that the measurements are carried out at different times rather than in different places. In our tomato example, if you only had one greenhouse and had to carry out the experiment in three batches to get the sample sizes you needed, then you could block on batch number (i.e. blocking in time) in an exactly analogous way to the situation we described of treating each of three greenhouses as a block (blocking in space). This may also be useful if you can't possibly assay your whole experiment in a single time period, and are concerned that the time when a particular experimental subject is measured might influence the measurement taken from it (see Chapter 11 for more on this).

The recipes described previously, where equal numbers of subjects from each block are assigned to all of the treatment types (or equal numbers of each treatment type are assigned to each block), are called *balanced complete-block* or *randomized complete-block* designs. It is possible to have different numbers of subjects within a block assigned to different treatment groups, or indeed to have treatment groups that do not get experimental subjects from every block. Such incomplete-block designs are much harder to analyse statistically. If you are forced into such a situation, perhaps by a shortage of experimental subjects, then our advice is to consult a statistician beforehand.

Q 9.1 Imagine that two scientists need to evaluate the quality of tomatoes grown in an experiment with three different greenhouses and four different treatment types. They are worried that the score a tomato gets might be affected by which of them measures it; can blocking come to their aid?

As well as intrinsic characteristics of subjects, it can also sometimes be useful to block on time and/or space.

9.3 The advantages and disadvantages of blocking

The advantage of blocking is in reducing the effect of intrinsic between-subject variation so that treatment effects are easier to detect. As such, it can potentially increase our statistical power. There are drawbacks too. Blocking using a characteristic that does not have a strong effect on the response variables that you eventually measure in order to compare treatments is worse than useless. Hence, do not block on a characteristic unless you have a good biological reason to think that the characteristic that you are blocking on will have an influence on the variables that you are measuring. The reason that blocking with an ineffective variable actually reduces your chances of finding a treatment effect is because the statistical test loses a bit of power when you block, because that power is used in exploring between-block differences. Normally this price is worth paying if blocking leads to reduced between-individual variation within blocks; but if it doesn't, you lose out.

Blocking also cannot be recommended if you expect high drop-out rates such that some treatment groups end up with no representatives from several blocks. This effectively produces an unplanned incomplete-block design, which, as we discussed earlier, can be very hard to analyse statistically. Generally, the benefits of blocking increase as sample sizes increase, hence we would not recommend blocking in situations where you are constrained to using rather fewer experimental subjects than you would ideally like. In such cases, complete randomization is usually better.

Don't block on a factor unless you have a clear expectation that that factor substantially increases between-individual variation.

Q 9.2 You have a sample of student volunteers on which you intend to explore the effect of four different exercise regimes on fitness. Are there any variables you would block on?

If you are concerned that blocking seems to violate the concept of independence of samples discussed in Chapter 5 see Section 9.3 of the supplementary information. Go to: **www.oxfordtextbooks.co.uk/orc/ruxton4e/**.

9.4 Paired designs

A common form of blocking is a **paired design**. For example, imagine that we want to look at the effect of an antibiotic injection soon after birth on the subsequent health of domestic cattle calves. We might use sets of twins as our experimental pairs, randomly assigning one calf from each set of twins to get the antibiotic. The attraction of this is that we can perform statistical tests (e.g. paired *t*-tests or *z*-tests) that compare within pairs. This has the same benefits as all forms of blocking: eliminating many potential confounding factors, because the two twin calves in a pair will be genetically similar, will be fed by the same mother, and can easily be kept together in order to experience the same environmental conditions. But notice one drawback: since we only perform our investigation on calves that were from a set of twins, can we be sure that our results will also apply to calves that were born alone? This type of extrapolation concern was dealt with in Section 6.4.1. Paired designs need not only use sets of twins. For example, a study looking at sex differences in parental feeding in seabirds might compare the male and female partners within a number of breeding pairs. Comparing within pairs controls for variation due to factors such as number of chicks in the nest and local foraging environment, both of which might be expected to contribute considerably to variation in feeding behaviour.

In **paired designs**, we divide the population into pairs of similar individuals (or use naturally occurring pairs) and randomly assign the individuals of each pair one to each of two treatment groups.

Paired designs can be attractive for similar reasons to blocking generally, but be careful to think about whether your choice of pairs restricts the generality of conclusions that you draw from your sample.

9.5 How to select blocks

In Section 9.1, we said that it did not matter much how you divided subjects into blocks. For example, if you had 80 subjects and four treatment groups, you could go for 20 blocks of 4, 10 blocks of 8, or 5 blocks of 16. Let's see if we can give you some clearer advice.

First of all, don't have any more blocks that you can justify if you are blocking on something that must be scored subjectively. For example, say the subjects are salmon parr and you are blocking them according to an expert's opinion on their state of health from external examination: given as a score on a 1–20 scale. In this case, it might be hard to convince yourself that those in the 18th and 19th blocks out of 20 are really different from each other. That is, it may be that the subjectivity involved in scoring fish on this fine scale adds noise to your study. If this is a valid concern, then you should use fewer blocks.

If you are ranking on something more objective, like the length of the parr, then you may as well go for as many blocks as you can, in order to reduce variation within a block due to length. However, our interpretation of the analysis of blocked experiments can be complicated if there is an interaction between the experimental factor and the blocking variable. For example, if we are interested in how quickly salmon grow on four different diets, straightforward interpretation assumes that although diet and initial length both affect growth rate, they do so separately. To put this another way, it assumes that the best diet for large fish is also the best diet for small ones. This may not be true. In order to explore whether it is or not, you need several fish on a given diet within each block. Hence, if you are concerned about interactions, and in general you often will be (see Chapter 8), then adopting 10 blocks of 8 would be a better choice than 20 blocks of 4.

Make sure you have replication of a treatment within a block if you are interested in interactions involving blocking factors.

9.6 Covariates

Let's return to our experiment where we compare the running speed of greyhounds fed different diets. Recall that we expect that differences in age will be an important source of variation between subjects' running performances, and the approach we took previously was to block by age categories. An alternative way to deal with between-subject variation due to age is to record the age of each dog alongside its speed. In our statistical analysis, we could then attempt to control for the effect of age, and so make any difference between the diets more obvious. In other words, by statistically removing some of the between-individual variation, we have increased the power of our experiment to detect any treatment effects. Box 9.2 shows how this works. This design allows us to answer the same question as the blocked design, namely:

> Do food types affect running speed in greyhounds once age has been controlled for?

Whether blocking by age or treating it as a covariate is the best way to proceed will depend on how the age is expected to affect running speed (see later in this section).

BOX 9.2 How controlling for a covariate can increase the power of the study

Suppose you are interested in whether treating a crop pest species of beetle with a pathogenic fungal agent has the potential to reduce their reproductive output. You carry out a simple fully randomized one-factor experiment. You allocate ten individual females to each treatment (fungus infection and control) and measure how many eggs they lay before they die. Figure 9.3a shows the kind of data you might get from such a study. Whilst there is a suggestion that there may be an effect of the treatment, it is clear from the figure that there is an awful lot of variation within each of the treatment groups, and this noise is going to make any effects hard to detect. Suppose that we carried out a power analysis prior to the study, using the amount of noise in this data as our estimate of variation, the power of the proposed study to detect a difference of, for example, ten eggs would be low (about 30%.)

However, we know that for many insect species the lifetime fecundity is strongly influenced by their body size, and so variation in body size within treatment groups may be accounting for some of this variation in fecundity. Figure 9.3b shows that our intuition was correct in this case; there is an obvious relationship between a female's weight and the number of eggs she lays in our study. And this allows us to control for this variation. Essentially, how this works is as follows: we fit a line through our data that relates the number of eggs laid to the weight values. One way to think of this line is as providing the predicted fecundity for females of any size. We can then use this line to ask, for any individual in our study, how many more or fewer eggs they produced than would be expected for a female of their size.

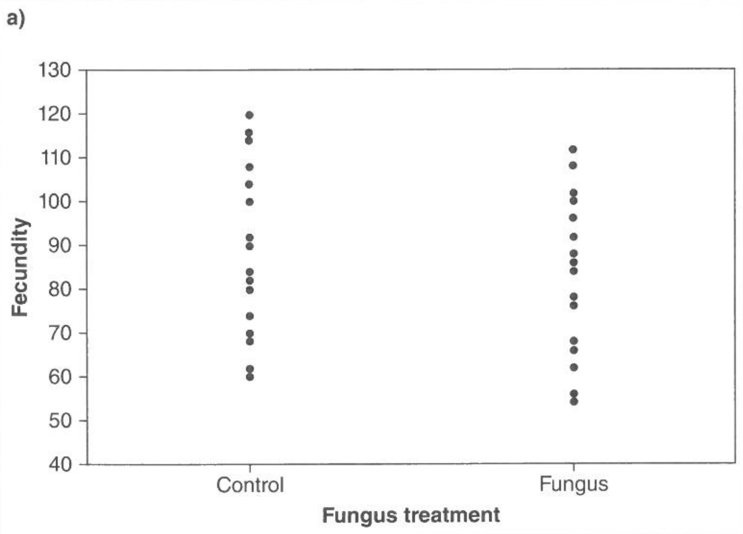

Figure 9.3 a) The lifetime fecundity of the females in the control and treatment groups.

This amounts to asking how far above or below the line does their actual fecundity value fall? Statisticians call this their residual fecundity; we can think of it as a measure of their fecundity corrected for their weight. Figure 9.3c shows these weight-corrected values plotted in the same way as the actual values in 9.3a. What is clear is that there is much less variation within treatments now that we have controlled for weight. For comparison, had we carried out a power analysis using the variation in this figure as our estimate of the amount of random variation, the predicted power of our study would have risen to close to 100%.

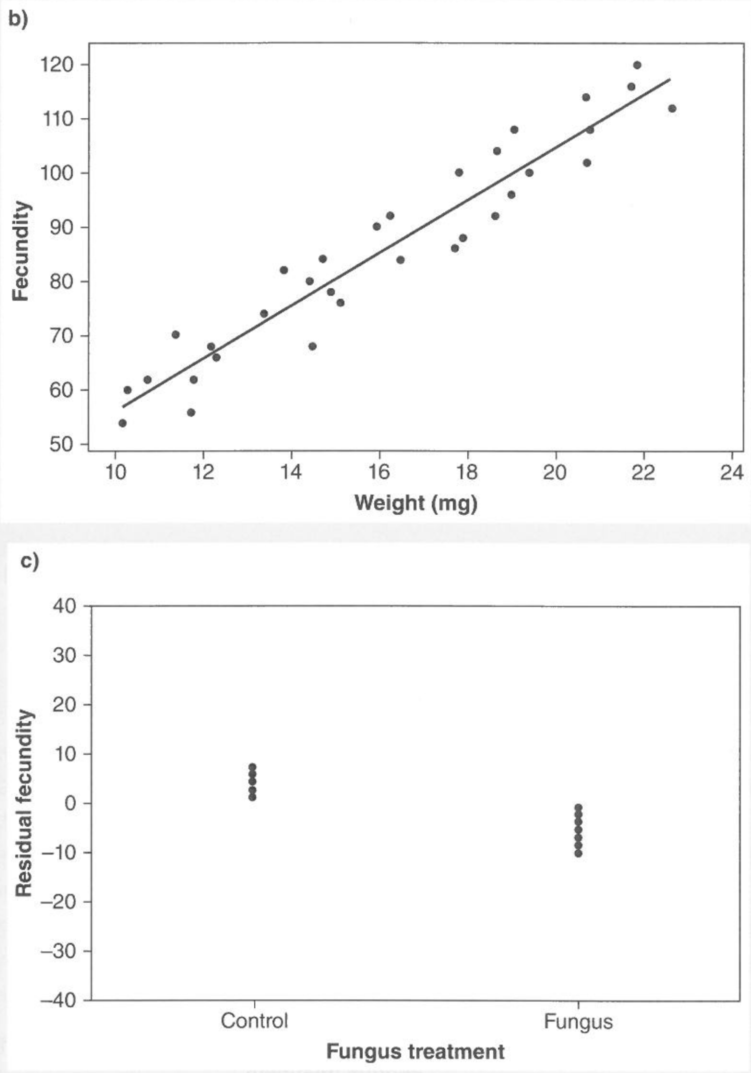

Figure 9.3 b) The lifetime fecundity of the same females plotted against their weight, and showing the line of best fit through the data. c) The residual fecundity of females after controlling for their weight.

As always we need to think about the statistics that we will ultimately use. In this case (where we measure a covariate), the design leads naturally to an ANCOVA (analysis of covariance) rather than an ANOVA. There are three requirements for such an analysis of covariance that are important to bear in mind.

The first requirement is that the covariant is a *continuous variable* like age. Continuous variables (such as height, weight, or temperature) are measured, can take a great range of values, and have a natural order to them (for example, 3.2 seconds is shorter than 3.3 seconds, and 5 metres is longer than 3 metres). In contrast, discontinuous (or discrete) variables involve a set of categories. They may only take limited values, and need not (but can) have any natural order to them. Typical discontinuous variables might be the colour of a seaweed, classified as *red*, *green*, or *brown* (no natural order), or the size of a tumour, classified as *small*, *medium*, or *large* (natural order). The distinction is not always clear cut, whilst classifying dogs as *young*, *middle-aged*, or *old* treats age as a discontinuous variable; the actual age of a dog in years might be regarded as discrete if only measured to the nearest year, but continuous if we allow fractions of years. If you need a reminder about different types of variable, have a look at Statistics Box 8.2. In our study, if we knew the exact age of our dogs then we could treat age as a covariate. But if we could only classify them as either *young*, *middle-aged*, or *old* then we should treat age as a blocking factor.

The second important assumption of ANCOVA is that we expect some form of linear relationship between running speed and age, so that speed either increases or decreases linearly with age. If we actually expected speed to increase between young and middle-aged dogs, but then decrease again as dogs become old, or to follow some other non-linear relationship, then a blocked design will be far more straightforward to analyse. Whether age is treated as a blocking factor or a covariate, your statistical test will be most powerful if the age distributions of dogs on the two diets are similar.

The third important assumption of ANCOVA is that any relationship between our continuous covariate and running speed is the same for our different treatment groups. In statistical terms, it requires that there is no interaction between our factor and our covariate. We will discuss this at greater length in the next section.

There is no reason to restrict yourself to one covariate. As a general rule, if there are characteristics of samples that you can measure and that you think might have a bearing on the response variable that you are actually interested in, then measure them. It's much better to record lots of variables that turn out to be of little use, than to neglect to measure a variable that your subsequent thinking or reading leads you to believe might have been relevant (but bear in mind our advice on focused questions in Chapter 2). Once your experiment is finished, you can't turn the clock back and collect the information that you neglected to take first time around.

Recording covariates can be a useful way to account for noise from measurable confounding factors, but blocking can often be more powerful unless your system satisfies two important assumptions about linearity of response and the continuous nature of the covariate.

9.7 Interactions between covariates and factors

In Section 8.2 we discussed interactions where the effect of one factor depends on the level of another factor. However, another important type of interaction that you will come across is between factors and continuous variables (covariates). An interaction between a factor and a covariate means that the effect of the factor depends on the level of the covariate. Figure 9.4 shows situations with and without such interactions. Both panels show the results of experiments to look at the effect of the brightness of a male stickleback's mating coloration (measured under standardized conditions) on his mating success when females are exposed to the males. Our experiment randomizes males to one of two experimental treatments: a tank illuminated by normal broad-spectrum light or a tank illuminated by a narrower band of frequencies giving red light. The two things that we expect to affect mating success are lighting treatment (a factor with two levels: normal or red) and a covariate (investment in pigmentation by the male, measured as a continuous variable: brightness). In Figure 9.4a, there is no interaction. Mating success is higher in normal light, but the lines are parallel, and this indicates that the effect of light type is the same at all levels of male brightness. Contrast this with Figure 9.4b. Here we see the relationship between mating success and brightness differs under the two lighting regimes (i.e. there is an interaction). The effect of light type now depends on the brightness of the fish concerned, with normal light increasing mating success of bright fish, but having no effect (or perhaps a slightly negative effect) on dull fish. Hence, you should be aware that it is possible to design experiments that allow you to test for interactions between a covariate and a factor as well as between two factors. In general, this will require that the values of the covariate overlap to some extent for the different levels of the factor. So we would want to avoid an experiment where all the fish tested in normal light had high levels of brightness, but all those tested in red light have low levels of brightness. Similarly, as we have already mentioned, if we want to use covariates to control for differences between individuals, such techniques are only really very useful (and definitely only straightforward) if there is a linear relationship between the thing that we are measuring and the covariate and if this relationship is the same for our treatment groups (there is no interaction between the covariate and the treatment).

➔ Covariates and discrete factors can interact; and such interactions will often require plotting the data to interpret effectively.

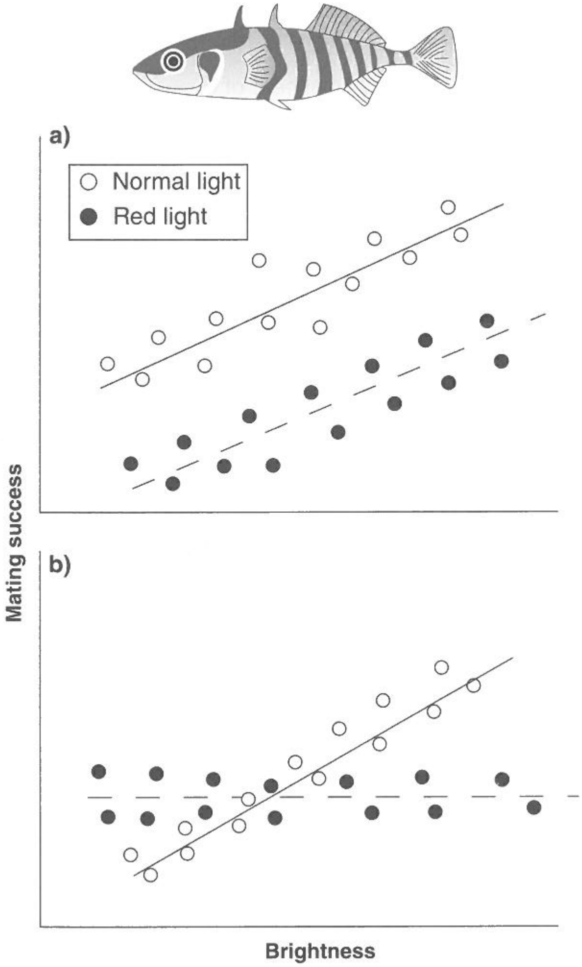

Figure 9.4 Two alternative outcomes for the results of an experiment to explore the effects of both a factor with two levels (normal or red lighting) and a covariate (brightness of the fish's exterior) on a male fish's mating success. In (a) both variables have effects but there is no interaction between these effects. That is, we can see that males are more successful if they are intrinsically brighter and/or if they are seen by females under full-spectrum normal light. But importantly there is no interaction: the effect of lighting regime is the same for all fish, no matter their brightness. In (b) there is an interaction. That is, we cannot say anything general about the effect of light without referring to a specific brightness level and we cannot say anything about the effect of brightness without referring to a specific lighting regime. Specifically, it seems that higher brightness is advantageous under normal lighting, but that brightness is immaterial to mating success under red light only. Another way to describe the interaction shown is that very bright individuals are more successful under normal lighting, whereas individuals of low brightness have less success under normal lighting than they do under red light.

Summary

- If you know and can measure some aspect of experimental subjects that is likely to explain a substantial fraction of between-subject variation then it can be effective to block on that factor.
- It can also be useful to block on time or space.
- Don't block on a factor unless you have a clear expectation that that factor substantially increases between-individual variation.
- Paired designs can be attractive for similar reasons to blocking generally, but think about whether your choice of pairs restricts the generality of conclusions that you draw from your sample.
- In general, make sure you have replication of a treatment within a block.
- For factors that vary continuously, including them as a covariate in your study may be a viable alternative to blocking, providing the trait in question is continuous and has a linear effect on the trait you are measuring in your study.
- Covariates and factors can interact.