Experiments with several factors (factorial designs)

8

- In this chapter we firstly introduce the concept and terminology associated with extension of the randomized design of Chapter 7 to include multiple factors (Section 8.1).
- We then discuss the concept of interaction between those factors (Section 8.2).
- We then add a note of caution against confusing levels and factors (Section 8.3).
- We introduce split-plot designs and Latin square designs and explain their attractions and limitations (Sections 8.4 and 8.5, respectively).
- We finish the chapter by discussing how the design concepts raised in this chapter link to statistical analysis (Section 8.6).

8.1 Randomized designs with more than one factor

Let's continue to think about the tomato plant project from Chapter 7. Imagine that (as well as the effects of plant feed) we are also interested in whether the application of insecticide has any effects on tomato plant growth. How should you go about exploring the effects of insecticide? The first option is to simply do what you did with plant feed and carry out a second experiment, again using a one-factor design with the presence or absence of insecticide as our experimental factor. Such an experiment would address the following question:

Does the application of insecticide affect the growth of tomato plants?

Whilst there is nothing wrong with this in principle, there is another possibility, which is to carry out a single experiment that looks at the effects of both plant feed and insecticide at the same time. Since we are exploring the effects of more than one factor, this fits the definition of a factorial experiment given in Chapter 7. How should we set up such an experiment? First we would allocate each of our tomato plant seedlings

at random to one of the four plant feed treatments ('control', 'low', 'medium', or 'high'). This is exactly what we did in Chapter 7. However, we would then take the seedlings allocated to one of the feed rates, say the control treatment, and randomly allocate them to one of the two insecticide treatment groups ('insecticide applied' or 'no insecticide applied'). We would repeat this for the other three plant feed treatments. Now we are varying two factors, the rate of plant feed application and whether or not we apply insecticide. This sort of design is called a *two-factor design* (or *two-way design*). The way that we have set it up means that it is also a *fully cross-factored* (or **fully crossed)** *design*. This means that we have all possible combinations of the two factors; or to put it another way, we have plants with and plants without insecticide in all of our plant feed treatment groups. In contrast, if for whatever reason we could only apply insecticide to plants in two of the plant feed treatment groups, but had plants without insecticide in all four plant feed treatment groups, then our design would have been described as an *incomplete design*. The statistical analysis of fully cross-factored designs is easy to do—in this case we would probably choose a two-way analysis of variance (ANOVA). In contrast, analysing incomplete designs is significantly more difficult. Avoid incomplete designs whenever possible. If you need to use an incomplete design for practical reasons (such as a shortage of samples) then get advice from a statistician first.

A **fully crossed** design means that all possible combinations of treatments of the factors are implemented. Thus if in a three-factor experiment, factor *A* has two levels, factor *B* has three levels, and factor *C* has five levels, then a fully crossed design will involve 30 different treatment groups (2 × 3 × 5).

For factorial experiments, your analysis will be a lot easier if you have a complete design.

8.2 Interactions

We have arrived at a fully cross-factored, two-factor design, with four levels for the first factor (plant feed), and two for the second (insecticide). This will give us eight treatment groups in total (4 × 2), and, as long as we have the same number of plants in each treatment, and more than one plant in each treatment too, our design will be balanced and replicated. You can easily imagine how you can get three-factor, four-factor, and even more complicated designs. In general, it is best just to imagine these; for experiments that you actually have to do, avoid making them so complex that you get overwhelmed either collecting the data or trying to analyse it.

If simplicity is the key to good design, then you might ask, 'Why do a two-way factorial design in the previous case; why not perform the two separate one-factor designs?' The main attraction of using a two-way design is that it allows us to answer multiple questions within the same experiment. In this case the main questions would be:

Do the plant feed treatments affect the growth of tomato plants?

Do the plant feed rates differ in their effects on the growth of tomato plants?

Does the application of insecticide affect the growth of tomato plants?

Does any effect of insecticide on growth depend on the rate at which the plants are being given plant feed?

The first three questions are about what statisticians refer to as the **main effects** of each factor (the overall effect of the plant feed and the overall effect of insecticide). The final question is about the **interaction** between the factors, what statisticians call the *interaction effect*, and this can only be answered with a factorial design.

In many biological studies, the most interesting questions are specifically about the interactions between factors rather than their main effects. For example, we might hypothesize that older people respond more slowly to a given drug therapy than younger people. If we put this into statistical terms, we are suggesting that the effect of one factor will depend on the level of another; there is an interaction between the two factors. The important thing to stress here is that, if your hypothesis is about the interaction, you must test for that interaction. And if you need to test the interaction, you need to use a design that allows you to test the interaction. This sounds obvious but it is easy to go wrong if you are not careful; we explore these issues in more detail in Box 8.1 and Statistics Box 8.1.

So our more complex two-factor design has allowed us several questions, including questions about interactions between factors that would not be possible with the simpler one-factor design. It has also saved us from some difficult decisions. Thus, suppose we had already completed our single-factor plant-feed study from Chapter 7 and are now planning a separate experiment to look at insecticide. An obvious question is should we be using plants that we also give plant feed, and if so, at what rate should they be fed? By carrying out a single two-factor experiment we avoid this decision, as we are looking at the effect of insecticide under all four plant-food conditions at once. Indeed, this also means that if we do see an overall effect of insecticide in our study we can be more confident about how general any effect of insecticide is likely to be, since we are testing it under four different plant feed conditions (the same can obviously be said for effects of plant feeding, since they are being examined with and without insecticide). This issue is discussed more extensively in Box 8.1. And there is one final advantage: our two-factor study has almost as much power to test for the effects of plant feed as a one-factor study with the same total number of plants. It also has almost as much power to test for insecticide effects as a one-factor study of insecticide with the same total number of plants. Thus, our two-factor study can examine both the effect of plant feed and the effect of insecticide using about half the number of tomato plants as would be required if we were to carry out separate experiments for each factor.

On hearing this argument, you are probably tempted to add more and more and more factors to your study. We would caution against this: if you vary too many factors simultaneously, your experiment will necessarily be very large, requiring a large number of subjects and a large amount of effort. Also, interpreting interactions between more than two factors can be challenging.

Let us imagine that we want to investigate whether dependent factor *A* is influenced by independent factors *B* and *C*. If the effect of *B* on *A* is affected by the value of *C*, or equivalently if the effect of *C* on *A* is affected by the value of *B* then there is an **interaction** between the factors *B* and *C*. Alternatively if the value of *A* is affected by the value of *B* in the same way regardless of the value of *C*, then we can say that there is a **main effect** due to factor *B*. Similarly, if *A* is affected in the same way by the value of *C* regardless of the value of *B*, then there is a main effect due to *C*. If there is an interaction, then by definition there are no main effects. If there is no interaction then both, one, or no main effects might be present.

Q 8.1 Let's illustrate the problem of interpreting an interaction involving three factors. Imagine that in the tomato plant experiment we explore the effects of tomato plant variety (two levels), food applied (two levels: whether a plant food was added to the growing medium at low or high dose), and insecticide treatment (two levels: whether an insecticide was sprayed on the plants or not) on growth rate. Our subsequent statistical analysis informs us that there an interaction between all three factors; can you describe biologically what this means?

BOX 8.1 Interactions and main effects

In the main text we claimed that one advantage of carrying out experiments with two or more factors was that this would allow you to look at both the *main effects* of the factors and also their *interactions*. Understanding what these actually are is fundamental to understanding these more complex designs. However, these concepts are also some of the most misunderstood statistical ideas. In this box we will explore these concepts in a little more detail in the hope of showing that they are both easy to understand and extremely useful.

Let's go back to thinking about the effects of feeding treatments on growth rate in different varieties of tomato plants. Imagine that we want to investigate two feeding regimes, high and low food, and we want to compare two varieties of tomato, which we will call varieties *A* and *B*. This will give us four different treatments (both varieties with both feeding regimes). Figure 8.1 shows several of the possible outcomes of an experiment like this.

Let's begin with the situation in Figure 8.1a. Here we can see that for both varieties of tomato, the high-feed treatment leads to faster growth than the low-feed treatment. However, if we compare the two varieties of tomato when they are fed identically, we see no difference. Statistically, we would say that there is a significant main effect due to feed treatment, but no main effect due to tomato variety. Biologically, we would say that growth rate is affected by how much you feed the tomatoes, but not by particular tomato variety.

The situation in Figure 8.1b is a little more complicated. Here again, we see that for both varieties the high-feed treatment leads to faster growth than the low-feed treatment. Although now, if we compare the two varieties under the same feed treatment, we find that variety *A* grows slower than *B*. Statistically, we would now say that we have significant main effects due to both feed treatment and tomato variety; biologically, we would say that tomato growth rate is affected by how much you feed them and also by the tomato variety.

Figure 8.1c looks very similar, but differs in one very important respect. Again, we can see that for both varieties, high feeding increases growth rate. Similarly, if we look at the growth rate of the two varieties when they are given lots of food we see that, as in the previous case, variety *A* grows slower than *B*. However, if we compare the varieties at the low-feed level, we see no difference between them. What does this mean? This is what statisticians refer to as an interaction. Biologically, it means that the effect of the variety of tomato on growth rate is different under different feeding regimes, or to be a bit more technical, the effect of one factor (variety) depends on the level of the other factor (feeding level). In our example, the difference is quite extreme; there is no difference between the varieties at one food level, but a clear difference at another.

In Figure 8.1d we see a less extreme case. We do see a difference between the varieties under both high- and low-food regimes, but under low food the difference between the varieties is much smaller. Statistically, this is still an interaction—the effect of one factor still depends on the level of the other—or to put it

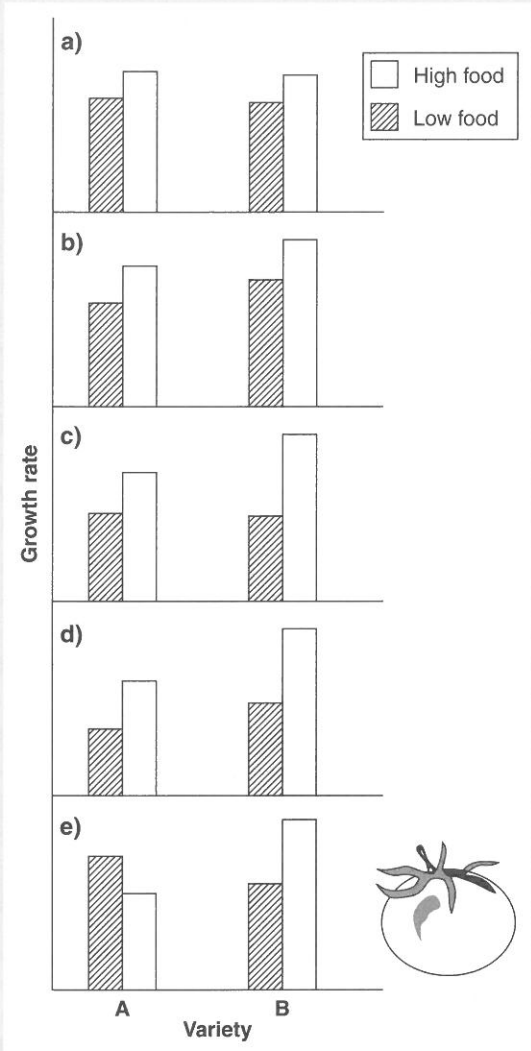


Figure 8.1 The growth rates of tomato plants in a two-factor experiment. One factor is plant variety (with two levels: *A* and *B*); the other factor is feeding regime (two levels: *high* and *low*). In (a) there is a main effect of feeding regime only (that is, there is no difference in growth rate between the two varieties when fed at the same level). In (b) both feeding regime and variety have separate main effects on growth rate. In (c), (d), and (e) there is an interaction between the two factors, such that the effect of one is different for different levels of the other.

biologically, the effect of variety on growth rates depends on which food treatment we consider.

We might even see a third type of interaction, as shown in Figure 8.1e. Here the growth rate of variety *A* is slower than *B* when they are given lots of food, but faster

than *B* when they are given only a little food, so not only does the size of the difference in growth rates depend on the food treatment, but so does the direction of the effect.

So why should we care about interactions? One reason is that a significant interaction can make it very difficult to interpret (or to even talk about) the main effect of a treatment. Imagine we got the result in Figure 8.1e. What is the main effect of food level? In one variety it is positive, and the other negative, so it is meaningless to talk about the main effect of food level. It only makes sense to talk about the effect of food if we state which variety we are referring to. Similarly, in Figure 8.1d if we want to talk about the difference between the varieties we need to state which feeding regime we are comparing them under. We can't really talk about an overall difference between the varieties in any meaningful way. Even in Figure 8.1b it is hard to say exactly what the main effect of feeding is because the size of the effect is different for the two varieties. So we can say that increased feeding has an overall positive effect on growth rate, but that the size of the positive effect depends on the variety. Only in situations like in Figure 8.1a can we make a confident statement, such as the overall effect of increased feeding is to increase growth rate by a specified number of units. Another way to think about this is that our factorial design has allowed us to say more about the generality of our treatment effect. If we had only carried out the experiment on a single variety, then any effect we detected might be very specific to the particular variety that we chose. By including a second variety, we can start to ask questions about how general any effects of the treatment are. If the treatment has the same effect for both varieties (i.e. there is no interaction), then we can be more confident that the effects of the treatment are likely to be more general. Of course, if we had included more varieties then we could be even more confident. On the other hand, if we detect a strong interaction between treatment and variety, then it tells us we must be very cautious in extrapolating our results to other tomato strains. Interpretation of interactions can sometimes be tricky, and we discuss this further in Statistics box 8.1.

STATISTICS BOX 8.1 An example of an experiment to test for an interaction

Suppose that you are interested in sexual selection in sticklebacks. You might hypothesize that because males invest heavily in sexual displays and competing for females, they may be less good than females at coping with parasitic infections. One possible prediction of this is that any deleterious effects of parasites will be greater on male sticklebacks than on females. This is a hypothesis about an interaction. We are saying that we think that the effect of parasitism will depend on the sex of the fish. In other words the effect of the first factor (parasitism) will depend on the level of the second (sex). By now, you will immediately see that one way to test this hypothesis would be to have a two-factor design, with infection status (parasites or no parasites) as one factor, and sex (male or female) as the other.

You might then measure the weight change in fish during the experiment as an estimate of the cost of parasitism. As an aside, you might be wondering why you need the fish with no infections—wouldn't it be enough to simply have infected males and females and compare their change in weight in a one-factor design? The problem with this is that we would have no control. If we found that fish changed in weight, we wouldn't know whether this was due to the parasites or something else (for example, the feeding regime). So, to measure the effect of parasitism, we need to compare fish with and without parasites. Now if we performed the experiment this way, it would be a relatively simple matter to test for the interaction between sex and infections status, and if the interaction was significant, this would support our prediction (but remember that a significant interaction could also arise if parasites had a bigger effect on females than males, so it is essential to look at plots of the data to make sure that the trend is the way round that you predicted).

Now this (hopefully) seems straightforward to you, but we claimed earlier that people often make mistakes, so what do they do wrong? The most common mistake is to treat this experiment as if it were two single-factor experiments done on the males and females separately. The researcher might analyse the difference between males with and without parasites, and find that males with parasites lose significantly more weight. The analysis could then be repeated on the females, and may perhaps find no significant difference between the weights of the two types of female. From this the researcher would then conclude that there is a difference in the effect of parasites on males and females. This seems logical, but it is wrong. The reason it is wrong is that the researcher has only tested for the main effects of parasitism, but not the interaction between parasitism and sex. By doing separate tests, the researcher has looked at the difference in male fish caused by parasites, and at the difference in female fish caused by parasites, but has not directly compared the difference in male fish to the difference in female fish. You may be thinking that we are being needlessly fussy, but in Figure 8.2 we show how this can make a difference. Figure 8.2 shows the results of an experiment like the one that we have just described. You can see that the weight loss of male fish is greater when they are parasitized than when they are not, and a statistical comparison of these means would show a significant difference between them. In contrast, for the females, the statistical test says there is no significant difference between the parasitized and unparasitized individuals. However, if we look at the figure, we can see that mean values for the females are exactly the same as those for the males. What is going on? Why is the male difference statistically significant, but the female difference not? Well there are several possible reasons, but one is that maybe the sample sizes for males and females are different. Suppose we had only half as many females as males, this would mean our statistical power to detect a difference between the female fish would be much lower than for the male fish, leading to us observe a significant difference in male fish but not in female fish. However, this is hardly good evidence for concluding that there really is a difference in the effects of parasites on males and females. Looking at the graphs, common sense

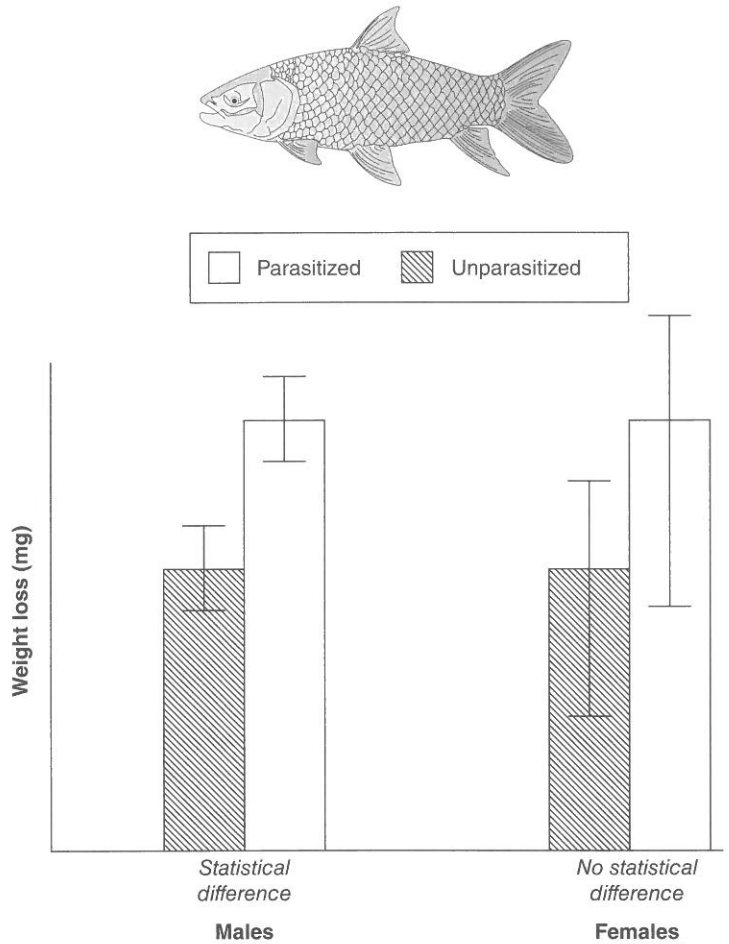


Figure 8.2 The mean and standard error of weight loss in fish, segregated by sex and parasitism status. We see that for both male and female fish the mean level of weight loss is greater for parasitized fish than unparasitized ones. Indeed, the mean weight loss for parasitized fish is the same for both sexes. This is also true for unparasitized fish. What is different between the sexes is the standard error (a measure of the effect of within-sample variation on our confidence that our sample mean is a good estimate of the population mean). For both parasitized and unparasitized fish, variation in weight loss seems much more important in females than males. The power of a statistical test is reduced when the effect of such variation is higher, so a test might find a significant effect of parasitism status on weight loss in males but not in females. However, as plotting the means makes clear, this difference between two separate tests should not automatically be interpreted as suggestive that parasitism status and sex interact in their effect on weight loss.

tells us that there is absolutely no good evidence for a difference in the way males and females respond to parasitism and if the researcher had tested the interaction directly using an appropriate statistical test, this is the conclusion that they would have drawn. This is not just a problem caused by unequal sample sizes; it can arise for any number of reasons causing greater intrinsic variation in some groups than others. In the end the only way of being confident that you have (or don't have) an interaction is to test explicitly for that interaction.

Notice in this example that sex is not a randomly allocated factor (nor was variety in our tomato plant example). That is, individuals are either intrinsically male or female; we do not randomly allocate subjects their level of the factor 'sex'. Despite this, we can still carry out the experiment and perform the analysis in exactly the same way as if the factor were randomly allocated. The only difference comes when we interpret our results. Because we did not allocate sex randomly, we have to be on our guard against confounding factors. If we find an interaction or main effect involving sex, we need to be on our guard that the driving factor might not be sex itself but something that correlates with sex but which we have not included in our statistical model. For example, it might be that it is not sex that has a strong effect but body mass, and males tend to have higher body mass than females.

Q 8.2 An eminent professor hypothesized that plants from old tin mine sites will have evolved increased ability to deal with tin poisoning. In a laboratory study, plants collected from mine sites grew bigger than plants from non-mine sites when grown in soil containing high tin levels. From this, the professor concluded that his hypothesis was correct. A bright young research student questioned this conclusion on the grounds that there was no evidence the increased growth rate had anything to do with tin; maybe the mine plants had evolved to grow bigger for some other reason. The professor repeated his experiment using tin-free soil and found no difference in the size of plants from the two groups. The professor now concluded that his hypothesis was certainly correct. Do you agree?

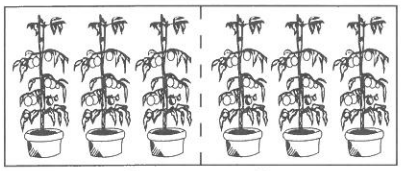
Often hypotheses we are interested in testing involve the interaction between factors. If you are interested in such a question, then you must ensure that your experimental design and statistical analysis allow you to directly explore this interaction.

8.3 Confusing levels and factors

One confusion that can arise with complicated designs is the distinction between different factors and different levels of a factor. You need to avoid such confusion in order to analyse and interpret a study correctly. In the tomato example, it is easy to see that the different feeding rates are different levels of a single factor. In contrast, what if we had performed an experiment with five different brands of plant feed? Now it is less clear; are the plant feeds different factors or different levels of the same factor? If we had performed this experiment and had a treatment for each plant feed, then we must think of the five plant feed treatments as five levels of a factor called plant feed type, and we would have a one-factor design. But suppose we then carried out an experiment with only two of the plant feeds and set up the following four treatments: no plant feed, plant feed *A* only, plant feed *B* only, plant feed *A* and plant feed *B* together. Now it is better to view the plant feeds (*A* and *B*) as two separate factors with two levels (presence or absence), and analyse the experiment as a two-factor design. In Figure 8.3, we show a number of possible designs with their descriptions to help you get used to this terminology.

Take care to avoid confusing levels and factors.

1.



A B

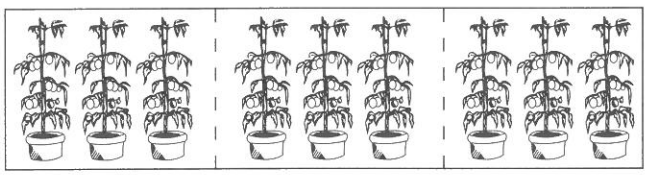
Fertilizer

Replicated 1-factor design with 2 levels of the factor (fertilizer type).

This can answer the question:

a) Do the fertilizers differ in their effect?

2.



Fertilizer **Pesticide** **Control**

1-factor design with 3 levels of the factor (type of cultivation).

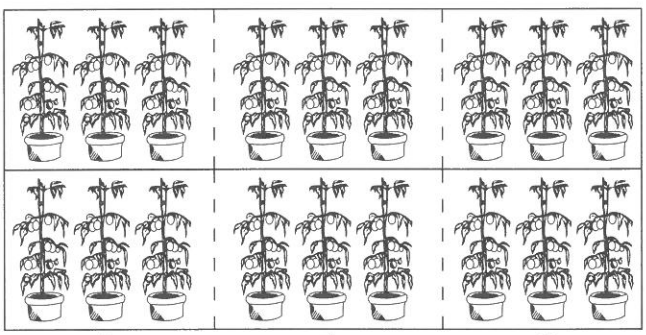
This can answer the questions:

a) Does fertilizer affect plant growth?
b) Does pesticide affect plant growth?
c) Do fertilizer and pesticide differ in their effect on plant growth?

3.

Pesticide

No Pesticide



A **B** **C**

Fertilizer

2-factor design with 3 levels of the 1st factor (fertilizer type) and 2 of the 2nd factor (pesticide use).

This can answer the questions:

a) Do the fertilizers differ in their effect on plant growth?
b) Does pesticide affect growth rate?
c) Does the effect of pesticides depend on the type of fertilizer?

Figure 8.3 Different experimental designs allow different questions about the system under study to be explored. In this case, we have three different experimental designs that allow exploration of the influence or influences on growth rate of tomato plants. The more complex the design, the more complex the range of questions it can be used to address.

8.4 Split-plot designs (sometimes called split-unit designs)

The term **split plot** comes from agricultural research, so let's pick an appropriate example. You want to study the effects of two factors (the way the field was ploughed before planting and the way pesticide was applied after planting) on cabbage growth. Three levels of each factor are to be studied. We have at our disposal six square fields of similar size.

The full randomization method of conducting this experiment would be to divide up each field into, say, six equal sections then randomly allocate four of the 36 sections to each of the nine combinations of ploughing and pesticide application (see Figure 8.4 for a picture of what the experiments would look like). The split-plot method would be to allocate two fields at random to each ploughing type, and then plough entire fields the same. We would then take each field and randomly allocate two of the six parts of that field to each pesticide treatment. Surely, this is a worse design than full randomization—why would anyone do this?

In a **split-plot design**, we have two factors and experimental subjects that are organized into a number of groups. For one factor (the **main-plot factor**) we randomly allocate entire groups to different treatment levels of that factor. We then randomly allocate individual experimental subjects within each of these groups to different levels of the other factor (the **sub-plot factor**).

The answer is convenience, and sometimes this convenience can be bought at little cost (although the statistical analysis of split-plot designs is more involved). In practice it is difficult to organize for different parts of a field to be ploughed in different ways; our split-plot approach avoids this nuisance. In contrast, applying pesticides, which is done by a person rather than a machine, can easily be implemented in small areas. The price we pay for adopting a split-plot design is that it is much better able to detect differences due to pesticide use (the **sub-plot factor**), and the interaction between herbicide and ploughing, than it is to detect differences due to ploughing (the **main-plot factor**). This may be acceptable if we expect that the effects of ploughing will be much stronger (and so easier to detect) than those of pesticide application, or if we know that ploughing will have an effect but what we are really interested in is the interaction between ploughing and pesticide use.

Split-plot designs are not restricted to agricultural field trials. Imagine exploring the effects of temperature and growth medium on yeast growth rate. The experimental units will be Petri dishes inside constant-temperature incubators. It would be natural to use temperature as the main-plot factor in a split-plot design, assigning a temperature to each incubator, then randomly assigning different growth mediums to the dishes inside individual incubators. Be careful here not to forget about replication. If you care about the effect of temperature at all, then you need replication with more than one incubator for each temperature. It is a very dangerous argument to claim that two incubators are absolutely identical, and so any differences in growth between them can only be due to their different temperature settings. Even if you bought them from the same supplier at the same time, it could be that the difference in their positions within the room means that one experiences more of a draught when open than the other, one of them may have antibacterial residues from a previous experiment that the other doesn't have, or one could have a faulty electrical supply that leads to greater fluctuations in temperature. Be very, very careful before claiming that two things in biology are identical (as they almost never are), and always replicate.

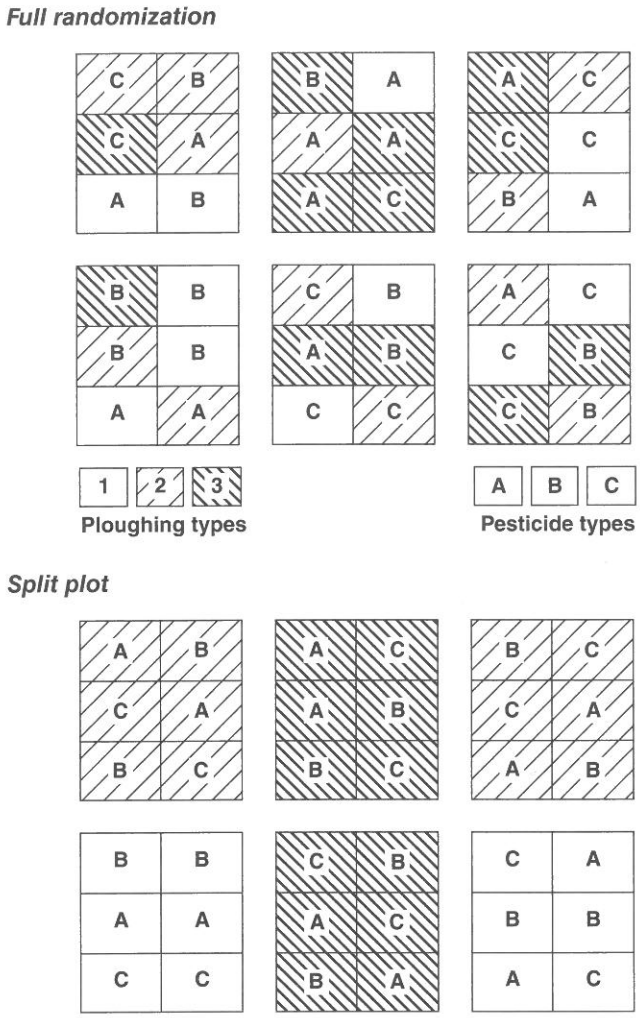


Figure 8.4 We have six fields each divided into six sections. We have two factors in our experiment: ploughing technique (with three levels: *1*, *2*, and *3*) and pesticide application (with three levels: *A*, *B*, and *C*). In the fully randomized treatment, we allocate four sections each to one of the nine (three times three) different combinations of the two factors. Allocation of a section to one of the nine treatment combinations is done entirely at random without consideration of which field a section is in. In the split-plot approach, entire fields are allocated a ploughing level. That is, two fields chosen at random are allocated to a given ploughing regime, and all the sections in a given field experience the same ploughing as each other. Ploughing is called the main-plot factor. We then take each field in turn and allocate the split-plot factor to individual sections within that field. We randomly allocate two sections within the field to each of the three methods of pesticide application. The split-plot approach may be attractive when it is practically very inconvenient to plough different parts of one field in different ways.

In both these cases the split-plot design is attractive because the spatial scales at which we can easily impose experimental variation differ between the two factors. A split-plot design can also be attractive if it is simply easier to impose variation in one factor than another. Imagine we want to explore the effects of three different levels of water temperature and four different lengths of previous fasting on the activity of crabs. We have at our disposal 48 crabs and a single large experimental aquarium. Each crab is randomized to a combination of water temperature and previous fasting period, and we have four replicate crabs for every combination. To help preserve independence of subjects we introduce crabs into the aquarium singly and observe them for thirty minutes each. If we completely randomize the order in which we assay crabs then we will frequently have to change the water temperature. This is inconvenient because it takes 24 hours to achieve this in our experimental system. To reduce this inconvenience we would split the crabs into six groups where all the crabs in a group have the same water temperature and there are two replicates of each fasting period per group. We then randomly order the eight crabs within each group, randomly order the six groups, and assay the 48 crabs in that order. Temperature has become the main-plot factor in a split-plot design: we have significantly reduced the number of times we will have to change the water temperature over the course of our experiment, and so shortened the duration of the experiment considerably.

Q 8.3 We argued earlier that replication is important. What would you do in the yeast-growth experiment described here if you only had access to three incubators and you wanted to explore three different levels of temperature?

Split-plot designs should only really interest you in an experiment with multiple factors where one factor can be allocated easily to groups of subjects, but is practically awkward to assign to individual subjects.

8.5 Latin square designs

Like the split-plot designs just discussed, the Latin square design is another alternative to a fully randomized factorial design. The big attraction of the Latin square approach is in reducing the number of experimental units required. Suppose that we wanted to evaluate the quality of apples from orchards subject to four different pesticide regimes, so the factor of interest has four different levels. We require experts to evaluate the amount of insect damage in a box of apples taken from each orchard and estimate the percentage of apples that are fit for sale to humans. Since evaluation is time consuming, we require four experts. Further, to find sufficient space for our experiment, we used four different farms (each of which has several orchards). We might reasonably expect both the identity of the expert and the farm that the apples came from to impact on the score that a box gets, as well as any effect of pesticide treatment. Thus, since we have three factors in our study, each with four levels, we would ordinarily recommend that the number of subjects be a multiple of 64 (43). This ensures that the experiment is balanced and has at least one subject for each combination of levels of the three different factors. Adoption of a Latin square design allows us to reduce this number of subjects considerably, such that we only need a multiple of 16 (42). This is done by allocating pesticide treatments to orchards within farms and boxes of apples to expert evaluators according to an appropriate Latin square (see Figure 8.5). Such an approach can considerably reduce the size (and thus cost) of an experiment.

Some examples of Latin square designs

3 x 3

	1	2	3
1	A	B	C
2	B	C	A
3	C	A	B

4 x 4

	1	2	3	4
1	A	B	C	D
2	B	A	D	C
3	C	D	B	A
4	D	C	A	B

	1	2	3	4
1	A	B	C	D
2	B	C	D	A
3	C	D	A	B
4	D	A	B	C

	1	2	3	4
1	A	B	C	D
2	B	D	A	C
3	C	B	D	B
4	D	C	B	A

	1	2	3	4
1	A	B	C	D
2	B	A	D	C
3	C	D	A	B
4	D	C	B	A

The first blocking factor (A) is represented by the columns and the second blocking factor (B) by the rows of these tables.

The treatment levels of factor C are represented by capital letters in the tables. Note the symmetry of these designs: each row is a complete block and each column is a complete block.

Figure 8.5 To make a Latin square, we first of all produce an $N \times N$ square grid, then place N distinct symbols (letters in our case), in the cells of this grid, such that each symbol appears once and only once in every row and every column (analogous to Sudoku puzzles). This grid is then used to allocate experimental units. For example, if we take the first (leftmost) of the 4×4 grids this could be used for the apple experiment described in the text. We want to investigate the effect of four pesticide treatments, that we label *A*, *B*, *C*, and *D*. We use four farms (represented by the rows 1 to 4 of the grid), and four experts (represented by the columns 1 to 4). Each expert will receive one box of apples from each of the different farms. Expert 1 will receive a box from farm 1 of apples grown under pesticide *A*, a box from farm 2 grown under pesticide *B*, a box from farm 3 grown under pesticide *C*, and a box from farm 4 grown under pesticide *D*. Expert 2 will receive a box from farm 1 of apples grown under pesticide *B*, a box from farm 2 grown under pesticide *A*, a box from farm 3 grown under pesticide *D*, and a box from farm 4 grown under pesticide *C*. In all, only 16 subjects (four orchards on each farm) will be required, whereas a traditional full-factorial experiment would have required 64 orchards.

However, there are restrictions on when Latin square designs can be applied. These designs are highly constrained in that the number of levels of each of the factors must be the same. For example, imagine we are interested in comparing the effects of three different diets on the growth rates of puppies. We are concerned that the breed of a puppy and the size of the litter it came from might be confounding factors. Now because diet has three levels (the three separate diets), then to use a Latin square design we need to impose that the other two factors also have three levels, so we would use three different breeds of dog and select subjects from three different litter sizes. This is quite a restriction. Let's say that to get the number of replicates you need, you have to use seven different breeds of dog. You could get around this problem if you could divide these seven breeds into three groups. We would only recommend this trick if there is a truly natural division (e.g. into gun dogs, lap dogs, and racing dogs, in our example). If you cannot achieve this symmetry, then you must use a fully randomized factorial design instead. An important assumption of the Latin square design is that there are no interactions between the three factors. In our example, we are assuming that the effects of breed, litter size, and diet act completely independently on the growth rates of puppies. Our knowledge of animal physiology suggests that this may well not be true. In cases like this, where you think interactions could be important, again we would recommend a factorial design instead, where interactions can be tested. However, in cases where you can justify the assumption of no interactions from your prior knowledge of the underlying biology, and where you can design a useful experiment so that all the factors have the same number of levels, the Latin square does present a very efficient design, in terms of reducing the numbers of replicates needed. In practice, we find that most scientists rarely find themselves in circumstances that justify use of this design.

The Latin square design is an alternative to a fully randomized factorial design using fewer experimental units. However, it can only be used to ask a very restricted set of scientific questions.

8.6 Thinking about the statistics

In any study, the experimental design that you choose and the statistics that you use to analyse your data will be tightly linked. Now that we have outlined some of the most common experimental designs that you will come across we want to finish this chapter by demonstrating this link with an example.

Suppose you are in charge of a study to evaluate whether a range of herbal extracts are of any use as antibacterial agents. To do this you decide to measure the growth of *Escherichia coli* bacteria on agar plates containing the different extracts, and also on control plates. You have nine extracts to test. How do you proceed?

Obeying our advice from Chapter 1, you decide to think about the statistical tests that you will use to analyse your data before you start collecting any. Your study has one factor: the different extracts. Thus, your study will be based around a one-factor design.

After discussion with a statistician (or reading a statistical textbook), you decide that the type of analysis appropriate for this study is a one-factor analysis of variance (or one-way analysis of variance). Do not worry if you have never heard of this test before, and have no idea how to actually do it. In fact, in these days of user-friendly computer programs, the actual doing of the test is the easy part—it is making sure your tests are appropriate for your data that can be tricky. The important thing from our point of view is that this test, as with all statistical tests, has a number of assumptions that must be met if it is to be used appropriately. One important requirement is that for each level of the factor you need at least two measurements (i.e. some replication). In your case this means that for every extract (and a suitable control), you need to make at least two experimental measurements of growth rate. A second requirement is that the data collected must be a measurement (i.e. measured on an interval or ratio scale, see Statistics Box 8.2 for details), not an arbitrary score. So if you chose to measure the average colony diameter of 100 randomly chosen colonies per plate, the analysis would be appropriate; but if instead you simply assessed growth on the plate on a 5-point scale based on categories (like 1 = no growth, 2 = very little growth, and so on), you could not use this analysis. There are several other requirements of this test, and whilst we will not go through them here one by one, that is exactly what you would do if you were in charge of this study. If you can meet all the requirements, all well and good, but what if you can't? First you need to consider the consequences of breaking a particular assumption. In general, failing to meet the requirements of a test will reduce your confidence in the result (and give ammunition to the Devil's Advocate). Failing to perfectly meet some assumptions may have little effect on the reliability of the test, and in this case you might decide to proceed with caution. However, this will often not be the case and failure to meet some requirements will entirely invalidate the test (which is not to say that you won't be able to get your computer to carry it out, just that the results would be untrustworthy). Detailed discussion of these different situations is well beyond the scope of this book, and if you find yourself in this situation we would strongly advise reading a good statistics book such as one of those in the Bibliography, or speaking to a statistician. If after this you decide that you cannot meet critical assumptions, you are left with two options: either to redesign the study to meet all the assumptions or to look for another suitable test with different assumptions that you can meet (bearing in mind that such a test may not exist, or may be considerably less powerful than your original choice).

Once you have decided on the test that you will use you can also begin to think about the power of the experiment you are proposing, and the kinds of sample sizes that will be required to make your study worth doing. Unless you know how you will analyse your data in advance it is impossible to make sensible estimates of power (since the power will depend on the test that you choose—see Chapter 6 for more on power).

Whilst these procedures might seem very long winded and slightly daunting, especially if you are not yet confident with statistics, we hope you can now begin to see clearly how thinking about the statistics when you think about your design, and long before you collect data, will ensure that you collect the data that allows you to answer your biological question most effectively.

STATISTICS BOX 8.2 Types of measurement

Different statistical tests require different types of data. It is common to split data into the following types:

- **Nominal scales**: a nominal scale is a collection of categories into which subjects can be assigned. Categories should be mutually exclusive but there is no order to the categories. Examples include species or sex.
- **Ordinal scales** are like nominal scales except now there is a rank order to the categories. For example, the quality of second-hand CDs might be categorized according to the ordinal scale: poor, fair, good, very good, or mint.
- **Interval scales** are like ordinal scales except that we can now meaningfully specify how far apart two units are on such a scale, and thus addition or subtraction of two units on such a scale is meaningful. Date is an example of an interval scale.
- **Ratio scales** are like interval scales except that an absolute zero on the scale is specified so that multiplication or division of two units on the scale becomes meaningful. Ratio scales can be either continuous (e.g. mass, length) or discrete (number of eggs laid, number of secondary infections produced).

Our advice is that you should try and take measurements as far down this list as you can, as this gives you more flexibility in your statistical analysis. Hence, if you are recording the mass of newborn infants then try to actually take a measurement in grams (a ratio measurement) rather than simply categorizing each baby as either 'light', 'normal', or 'heavy'; as this will increase the number and effectiveness of the statistical tools that you can bring to bear on your data. Of course, by taking our advice and deciding before data collection what statistical tests you will carry out, you will already know the type of data you require.

Think about the statistics you will use before you collect your data.

Summary

- For factorial experiments, your analysis will be a lot easier if you have a complete design.
- Often hypotheses we are interested in testing involve the interaction between factors.
- If you are interested in such an interaction effect, then you must ensure that your experimental design and statistical analysis allow you to directly explore this interaction.
- Take care to avoid confusing levels and factors.

- Split-plot designs should only really interest you in an experiment with multiple factors where one factor can be allocated easily to groups of experimental units, but is practically awkward to assign to individual experimental units.
- The Latin square design is an alternative to a fully randomized factorial design using fewer experimental units. However, it can only be used to ask a very restricted set of scientific questions.
- Think about the statistics you will use before you collect your data.