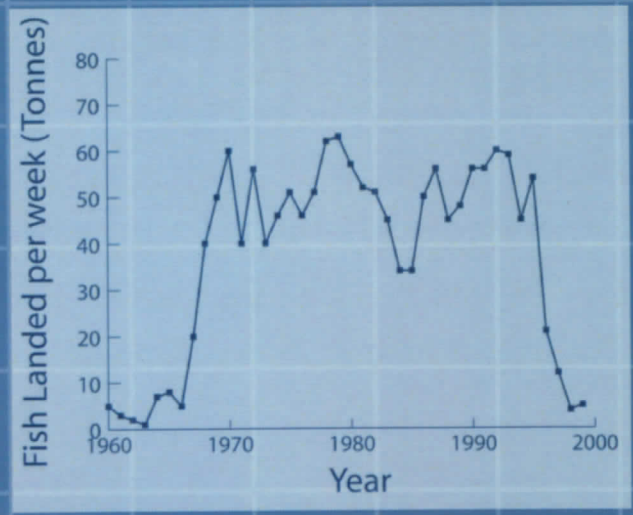
Introduction to Statistics for Biology

THIRD EDITION

Robin H. McCleery
Trudy A. Watt
Tom Hart



4

Planning an Experiment

> Organizing is what you do before you do something, so that when you do it, it is not all mixed up.
>
> **A.A. Milne**

This chapter is split into two sections and deals with how to design an effective experiment. First, we will discuss the principles of sampling to show how different sampling regimes affect the results we obtain. Then, we will put this knowledge to use by discussing how to design your own experiment.

4.1 Principles of Sampling

4.1.1 First, Catch Your Worm!

In Chapter 2, we measured ten worms and considered how to describe and analyse them. How we actually obtain these data is not a trivial matter. There are practical as well as statistical issues. Let us consider the challenge.

First, there are practical problems: if we go out during the day, worms will be buried, and we will need to dig them up. How deep should we go? Also, we run the risk of cutting some of them, which we would then not be able to use. Would there be a bias towards cutting and discarding longer worms? It would probably be better to go out on a wet night with a torch and collect them into a bucket while they are on the surface. Is there a bias in which worms come up? All of these practical considerations might influence which individuals we pick.

How do we decide which ten worms to pick up from the many thousands available? If we collect them from the footpath because it is convenient, we might have individuals that can move through compacted soil and might therefore be smaller than average. We want a sample of ten worms that is *representative* of the *population* of thousands of worms in the field. Deciding

on a method to pick ten worms is called **sampling**. This chapter will describe alternative methods of sampling, together with their strengths and weaknesses, and then discuss the parts of a good experimental design.

4.1.2 The Concept of Random Sampling

We have already discussed this in some detail in Chapter 1, Section 1.2.4. The purpose of random sampling is to ensure that each individual in the population of interest has an equal chance of being selected (no bias towards certain individuals). We usually take a random sample (in this case, of ten observations). This may be unrepresentative because, by chance, no observations have been selected from a distinctive part of the population. Consider what we might infer about the heights of athletes in general if our sample site was a basketball court as opposed to the jockey's enclosure at a race track. These are obvious biases, but we may not know the biases of our population before we sample them, so we need to take general precautions.

4.1.3 How to Select a Random Sample

Let us consider a worked example of how to generate a random sample. For simplicity we will imagine that a small square contains 400 bushes, conveniently arranged at 1 m intervals in 20 rows and 20 columns. Figure 4.1 below shows the weight of fruit (g) from each bush. We need to work out how to select a random sample without knowing anything about the distribution of the population.

How could we do it badly? One of the easiest ways is to walk around and sample the first ten bushes we see. Using this method, we would probably have a bias towards the biggest, most conspicuous, bushes. Our sample of bushes (and probably the yield) would be bigger than the population average, so from our sample we would infer that bushes have a larger yield than they actually do. Also, our bushes would probably all be on the edge, so we may sample bushes that are not competing as much for water and space. This shows how problematic sampling can be, and why time spent planning is not wasted.

To select a position at random we could use the last two digits from the random number button on our calculator. If this gives 0.302 and then 0.420, this would identify the position in row 2 and column 20 (2 m down from the top of the field, and 20 m from the left). If we select a random sample of 16 positions at which to harvest the crop and measure its yield, we could get the selection shown in Figure 4.1. Note that, by chance, no positions have been selected from the top right or bottom right of the field, where crop yields happen to be low. So, this sample is random, but unrepresentative.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	8	7	12	9	6	8	12	14	7	4	5	4	3	7	4	2	4	6	8	7
2	16	18	9	10	7	9	14	10	6	3	4	8	5	6	5	3	3	4	6	6
3	22	**15**	17	14	12	13	16	9	4	5	6	5	7	7	6	5	4	5	7	9
4	19	16	13	10	8	9	12	**9**	8	7	9	**10**	9	8	3	3	7	10	13	10
5	16	12	12	9	5	7	10	12	6	8	7	6	6	5	6	4	6	7	9	10
6	19	16	11	7	8	**10**	14	10	16	15	17	16	18	17	19	22	24	**20**	19	15
7	22	18	16	19	12	19	17	12	18	**21**	20	19	25	22	24	27	20	17	22	14
8	25	14	16	12	13	15	16	14	23	25	28	33	30	27	31	25	22	20	24	29
9	20	17	14	**16**	17	19	20	24	25	27	32	40	42	35	35	37	29	28	27	35
10	**25**	22	19	25	29	32	36	30	35	40	**41**	47	45	40	32	33	27	27	31	30
11	29	30	27	32	37	42	45	45	43	51	53	52	48	39	42	**37**	33	29	27	20
12	31	35	38	45	47	44	49	51	**50**	58	56	50	41	43	30	27	29	23	21	19
13	28	33	30	**37**	40	39	42	46	51	48	44	**40**	37	30	35	22	21	17	14	16
14	26	29	32	35	35	31	29	32	40	37	35	31	16	19	20	11	12	10	13	12
15	22	28	31	29	28	27	25	27	16	22	19	16	8	4	2	3	6	5	7	10
16	21	22	28	26	31	**27**	22	24	**20**	14	10	8	6	6	4	7	4	3	1	5
17	17	26	30	31	31	27	25	33	12	8	7	6	6	4	2	1	0	0	0	0
18	17	24	31	33	27	22	19	17	14	10	9	7	3	0	3	1	0	4	2	0
19	19	19	21	17	15	16	16	19	16	13	10	12	8	3	1	0	3	2	5	3
20	23	**21**	17	14	13	19	23	17	12	11	6	9	7	3	1	0	1	1	2	4

FIGURE 4.1
Random sample of 16 bushes. Bold numbers indicate samples.

4.1.4 Systematic Sampling

In systematic sampling, sample units are chosen to achieve maximum dispersion over the population. They are not chosen at random but regularly spaced in the form of a grid (Figure 4.2).

The first point of a systematic sampling grid should be chosen at random so that we do not generate a systematic bias, for example, always miss out the edges of a site. Once this is chosen (for example, column 15, row 2 in Figure 4.2), the other sample points are chosen in a fixed pattern from this point, and so they are not independent of each other.

Much use is made of systematic sampling. For example, every 10th tree in every 10th row may be measured in forestry plantations. As long as the number of sample units is high, there is little risk of coinciding with any environmental pattern that might affect tree growth (e.g., drainage channels), and the data are often treated as if they were from a random sample.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	8	7	12	9	6	8	12	14	7	4	5	4	3	7	4	2	4	6	8	7
2	16	18	9	10	7	9	14	10	6	3	4	8	5	6	5	3	3	4	6	6
3	22	15	17	14	12	13	16	9	4	5	6	5	7	7	6	5	4	5	7	9
4	19	16	13	10	8	9	12	9	8	7	9	10	9	8	3	3	7	10	13	10
5	16	12	12	9	5	7	10	12	6	8	7	6	6	5	6	4	6	7	9	10
6	19	16	11	7	8	10	14	10	16	15	17	16	18	17	19	22	24	20	19	15
7	22	18	16	19	12	19	17	12	18	21	20	19	25	22	24	27	20	17	22	14
8	25	14	16	12	13	15	16	14	23	25	28	33	30	27	31	25	22	20	24	29
9	20	17	14	16	17	19	20	24	25	27	32	40	42	35	35	37	29	28	27	35
10	25	22	19	25	29	32	36	30	35	40	41	47	45	40	32	33	27	27	31	30
11	29	30	27	32	37	42	45	45	43	51	53	52	48	39	42	37	33	29	27	20
12	31	35	38	45	47	44	49	51	50	58	56	50	41	43	30	27	29	23	21	19
13	28	33	30	37	40	39	42	46	51	48	44	40	37	30	35	22	21	17	14	16
14	26	29	32	35	35	31	29	32	40	37	35	31	16	19	20	11	12	10	13	12
15	22	28	31	29	28	27	25	27	16	22	19	16	8	4	2	3	6	5	7	10
16	21	22	28	26	31	27	22	24	20	14	10	8	6	6	4	7	4	3	1	5
17	17	26	30	31	31	27	25	33	12	8	7	6	6	4	2	1	0	0	0	0
18	17	24	31	33	27	22	19	17	14	10	9	7	3	0	3	1	0	4	2	0
19	19	19	21	17	15	16	16	19	16	13	10	12	8	3	1	0	3	2	5	3
20	23	21	17	14	13	19	23	17	12	11	6	9	7	3	1	0	1	1	2	4

FIGURE 4.2
Systematic sample of 16 bushes.

In social surveys, every 50th name on the electoral roll might be selected as a person to be interviewed. This is very convenient. However, it is important to be aware of the possibility of bias. If flats were in blocks of 25 and were all occupied by couples, we might only interview people who lived on the ground floor!

Systematic sampling is very efficient for detecting rare events because we are less likely to miss one such event (perhaps a fallen tree within a wood or a molehill in grassland) than if the sampling has a random element. Also, because we are more likely to include both very small and very large individuals, the mean of a homogeneous population is often close to the true mean but has a large standard error.

4.2 Comparing More Than Two Groups

So far we have learned how to compare means of two groups. Very often, however, we wish to compare the mean performance of several different categories. For example:

- Do four fertilisers differ in their crop yields?
- Do three species of plant differ in their ability to colonise loam?
- Do five drugs differ in their ability to cure disease?

In this section we will see how to design experiments that will answer such questions. Let us start by designing an experiment to investigate the effect of fertiliser on yield of barley. We will have three treatments (A, B, and C). The term treatment is used to describe what the crop receives — so we are going to design an experiment with three treatments.

4.3 Principles of Experimental Design

4.3.1 Objectives

It is good practice to write down the background to your experiment. This consists of why you are interested in the problem and your general objective. It is good to include any important aspects of the biology of your experimental organism, or any practical considerations that may affect the experimental layout. Do you need to leave space between plots to avoid cross-contamination? Do you need to leave access space for watering or feeding? It is best to identify any biological factors that will affect your layout early on.

4.3.2 Replication

Replication gives us an indication of the variation of results and, in this way, an idea of the accuracy of our estimates. Suppose you are a policeman attending an accident. If you arrive on the scene and five separate witnesses tell you that the car was blue, you would be much more confident that the car was blue than if there were only one witness. Similarly, suppose four witnesses say blue, and one says green. You would still probably conclude that the car was blue, but with less certainty than before. Replication has also given you an idea of the variation. If you had spoken to the first witness (green) and stopped there, you would have concluded the wrong color and have no idea about the spread of opinion.

This concept is illustrated graphically below (Figure 4.3). Note how increasing the number of replicates dramatically changes the relationship and the inferences you would draw.

So, how do we replicate each treatment in a planned experiment (Figure 4.4)?

It is no good just splitting a field in three and applying a different fertiliser to each third (Figure 4.4a). Perhaps the natural fertility of the soil is higher at the bottom of the slope, so whichever fertiliser is allocated to that position will appear to be even better than it really is in comparison with the others. There is only one replicate of each treatment so how can we tell whether the

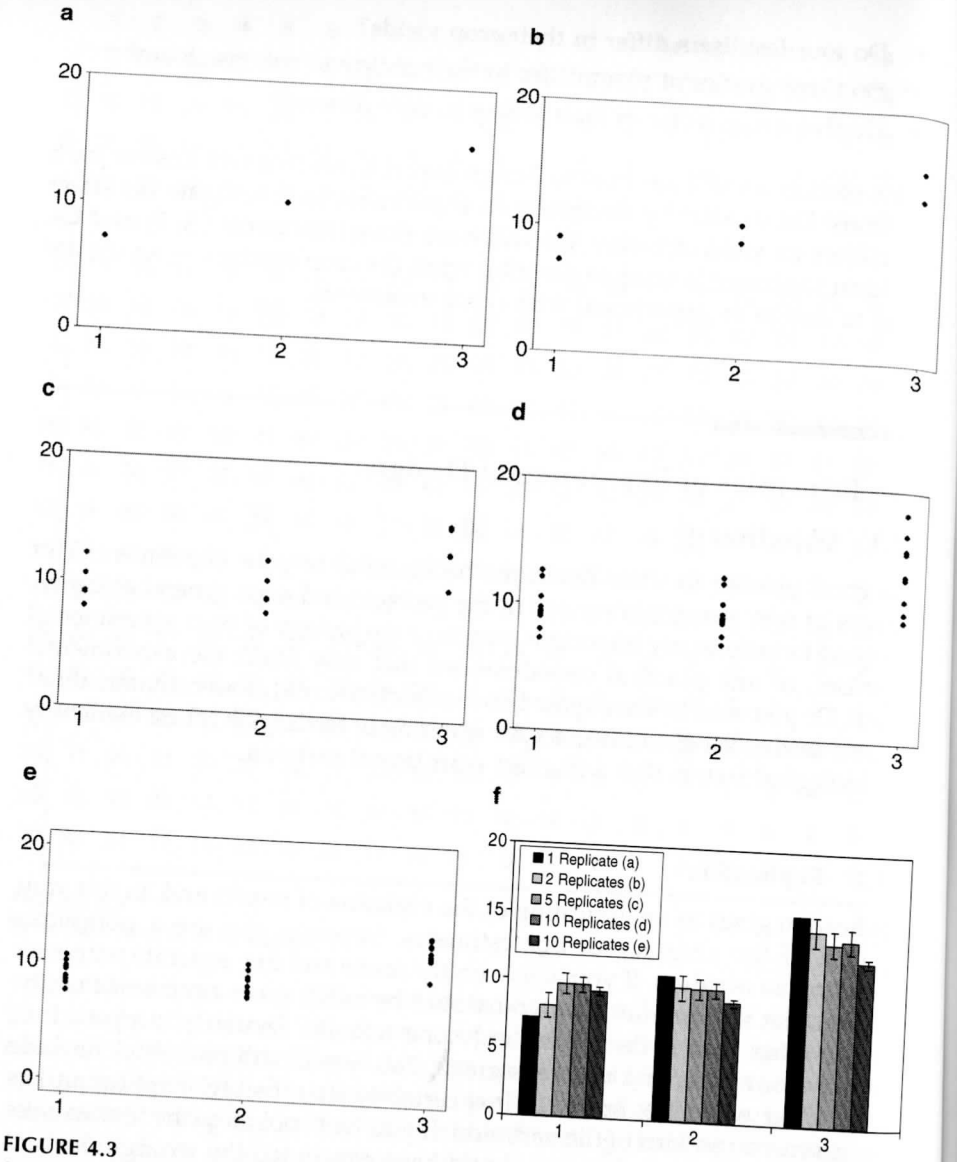


FIGURE 4.3
(a–d) As the number of replicates increases, the shape of the relationship changes. We can have more confidence in the estimate with more replication. (e): A data set with the same group means but smaller spread. (f) The relationship between replication and the standard error. In each case, the standard error decreases with increasing replication, as shown by the error bars getting smaller. Standard error cannot be calculated for only one group, so the error bar is missing for this data series. As shown by the difference between (e) and (f), less spread gives a smaller standard error, which indicates that we have a better estimation of the mean.

effect is due to the treatment, or to soil differences between plots? However, if we divide the field into 12 parts or experimental units (usually called plots) and allocate each fertiliser treatment (Section 4.3.3 discusses how to randomly

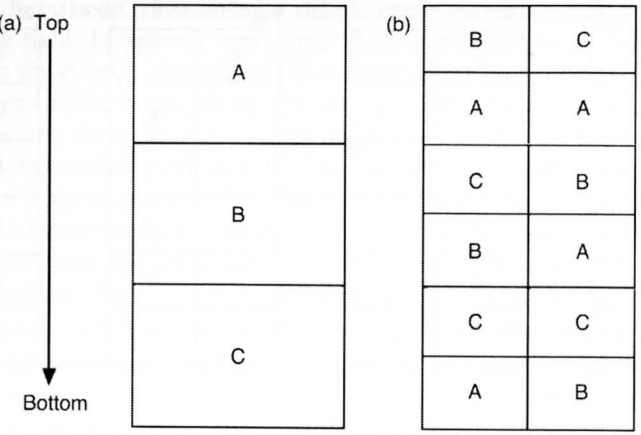


FIGURE 4.4
(a) Allocating treatments without replication. (b) An experiment with four replicates, each randomly allocated to a position on the field.

allocate treatments) to four of them, we will improve matters (Figure 4.4b). We now have four replicates of each treatment.

We can compare the yields from the four plots receiving organic fertiliser treatments. Any variation in these yields is caused by random variation in conditions across the site. The same is true for variation in the four yields from treatment B and, separately, from treatment C. So, replication has given us three separate estimates of the background or random variation in the yields of plots receiving the same treatment.

The greater the replication we have (perhaps six plots of each treatment instead of four), the more independent pieces of evidence we have about the fertilisers' effects. This means that we are able to detect smaller real differences in yield between the populations from which we have taken our samples. We can achieve increased precision in our estimates of the population mean yields and of differences between them by having more replication. A more technical discussion on choosing how much replication is needed to be found in Appendix B.

4.3.3 Randomisation

The four replicate plots of each treatment must be allocated to positions in the field at random to avoid the unknown biases we discussed before (Section 4.1.3). This is achieved by numbering the 12 plots from 1 to 12 (Figure 4.5a). We then use the random number button on the calculator (or use random number tables) to select 4 of these plots for treatment A (e.g., the numbers 0.809, 0.312, 0.707, 0.836, and 0.101 allocate this treatment to plot numbers 9, 12, 7, and 1; we ignore the value 36 from 0.836 because it is greater than 12) (Figure 4.5b). Then we select four more numbers for treatment B, and treatment C must go on the four plots that remain, which will be random because we randomised A and B.

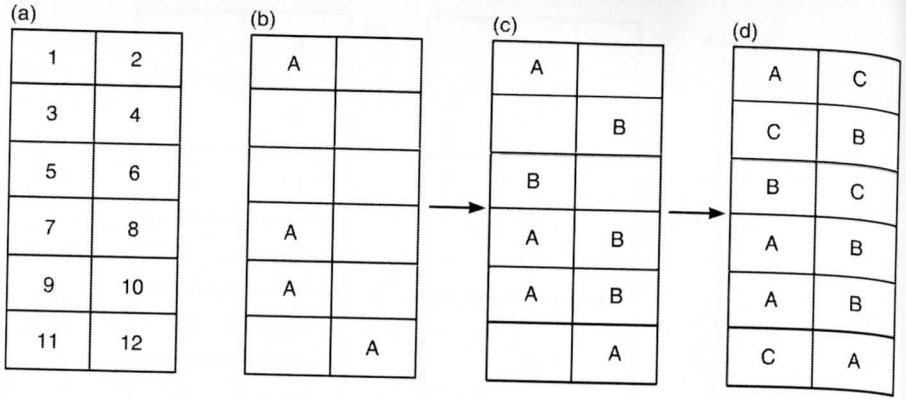


FIGURE 4.5
(a) Numbering of plots to enable random allocation of treatments (b–d) Allocating each treatment to a plot in turn.

As with selecting a sample from one population (Chapter 2), we can think of randomisation in allocating treatments to experimental units as an insurance policy. It protects us from obtaining estimates of treatment means that are biased, that is, consistently higher or lower than the population mean. This could arise because all the plots for one treatment happened to be in the corner of the field where the crop was damaged by frost. The replicates of each treatment must be interspersed (mixed up) over the site, and randomisation is a good way of achieving this. Most statistical analyses assume that you have randomised experimental units to the treatments. Very rarely, randomisation may produce an arrangement of plots in which, say, the four replicate plots of one treatment are grouped together at the bottom of the slope. Statistically, this would be fine because randomisation works to ensure that experiments in general are bias free. However, we are only interested in this particular instance of the experiment, which will cost us time and money. In the case of randomisation producing a clear bias, it would be prudent to rerun the randomisation process and reallocate plots. Only do this in very clear cases, however. Humans are very good at detecting patterns, and as a consequence, we tend to think we see clusters in patterns that are actually truly random.

4.3.4 Controls

Should we have a control in an experiment? A control is the name given to a treatment in which nothing is applied to the plot. We can then see what changes take place naturally during the experiment. Statistically, we treat this as an extra treatment. A variation on this idea is a procedural control. For example, in a drug trial, we might inject three different drugs into people (the treatments), and the control would be injecting saline. This means that we are not comparing people who have had an injection with people who have not.

So far we have been considering a simple experiment in which fertilisers are applied to a field. Imagine a more complicated experiment in which we are interested in the effect of sheep grazing at different times of year on the number of wildflower species in grassland. We can fence off each plot to make sure that the sheep are in the right place at the right time each year for, say, 5 years. However, the grassland present at the beginning will contain a range of species characteristic of its location, soil type, and previous management (for example, always grazed with sheep or was growing maize 2 years ago). So, if we carried out the same experiment on another site 50 km away, the species present would be rather different. Thus, having control plots on each site is useful. It tells us what happens on each site in the absence of any grazing treatments. We can think of it as providing a standard comparison between the two sites.

4.3.5 Blocking

What we are doing in experimental design is making it easy to partition variation into what we are interested in and what we are not. We will cover this in greater detail in Chapter 5, but in the meanwhile, it is important to introduce the concept of **blocking**. In Section 4.3.3 we looked at randomisation as a way of avoiding experimental biases. Another way to do this is by blocking, and the two techniques are routinely used together. Whereas randomisation is used to control for *unknown* biases or variation, blocking is brought in to control for *known* or *likely* differences between replicates.

With the completely randomised distribution of treatments earlier in this chapter, it is quite likely that we will allocate two replicates of a particular treatment to the field margin at the top of the slope and none to the margin at the bottom. Therefore, our results for each fertiliser could be influenced as much by the sample location as the effect of fertiliser. If we were to carry out such an experiment many times, effects would even out because, perhaps next time, the same treatment might be overrepresented at the bottom of the slope. However, if we have only the resources for one or two experiments, we need to find a way of overcoming known variations like this. Look at the two following graphs; see how our interpretation changes if we know that the higher values in each treatment come from wetter soil. If we can find a way of including this information in the analysis, then the error (the spread of the results) will clearly be much less (Figure 4.6).

If we know that the soil is dry at the top of our field and wet at the bottom, then we should block for this rather than randomise. We have four replicates, so it makes sense to divide the field into four blocks and represent each of our three treatments in each block. We should still randomise the position of treatments within block, but the main (known) bias is already dealt with by block. The logic of blocking is that rather than averaging out a bias, we are investigating it like a treatment, so that we work out the effect it has and then discount it. How we analyse a design that includes a blocking factor will be covered in Chapters 5 and 6.

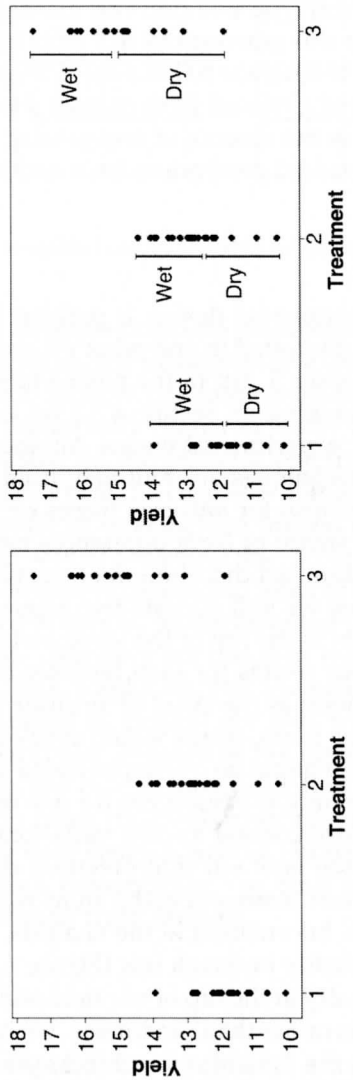


FIGURE 4.6
How our understanding may be further improved by including the information that soil moisture varies across the field.

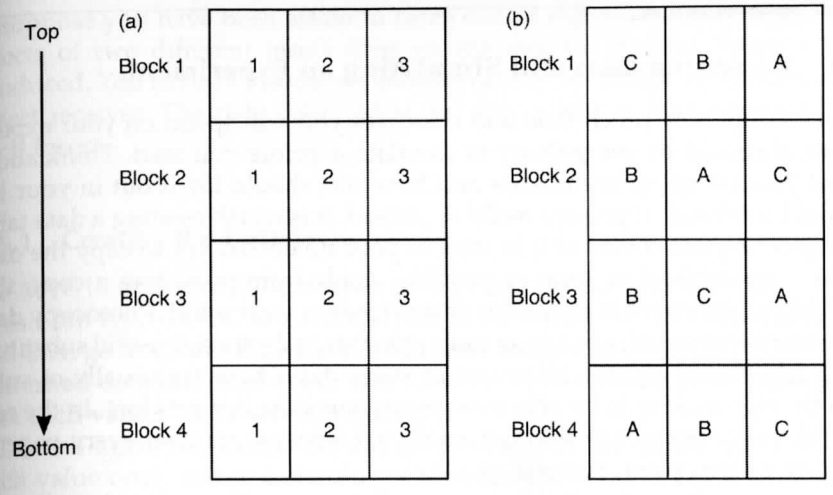


FIGURE 4.7
A randomised complete block design; each treatment is present in each block, but the order within each block is randomised.

A **randomised complete block design** (Figure 4.7) and reduces random variation and improves our ability to detect differences between treatments. Each quarter of the experiment is a "block," and each block contains a complete set of all the treatments within it. Each block is selected so that the conditions are even (or homogeneous) *within* it but they differ *between* one block and another. So, if the preceding diagram represents a field the top block might be on slightly sandier soil, and the block at the bottom might be more shaded because of a hedge. These differences should not be too great, as blocking is just a way of getting rid of confusing variation and will not give reasonable results unless the whole experiment is conducted on a reasonably uniform site.

How should we allocate the four treatments to the three plots within each block? This must be done using random numbers (as in Figure 4.7), but here we would generate random numbers from 1 to 3 to ensure that each treatment has an equal chance of occurring on each plot. We number the four plots in each block from 1 to 3. Then we use the random number button on our calculator. If the last digit is 2, we allocate treatment A to plot 2; if number 1 appears next, we allocate treatment B to plot 1. This leaves plot 3 for treatment C. Repeat this process for the next block.

Blocking should be used wherever there may be a trend in the environment that could affect the feature in which you are interested. For example, in a glasshouse heating, pipes may be at the rear of the bench, so one block should be at the rear and another at the front of the bench. Even in laboratories and growth cabinets there can be important gradients in environmental variables that make it worthwhile arranging your plots (pots, trays, petri dishes, etc.) into blocks. It is common to block feeding experiments with animals by putting the heaviest animals in one block and the lightest ones in another block. This helps to take account of differences in weight at the start of the experiment.

4.4 Recording Data and Simulating an Experiment

Considering how much time and resources you will spend on your experiment, it should be compulsory to simulate it before you start. Think about what you are going to measure and how you should lay it out in your lab or field notebook. If you are really organised, it is worth creating a data table and printing out copies of it to stick in your notebook. Try to copy the data onto a spreadsheet as soon as possible. Aside from providing a copy, this will highlight mistakes while the data is fresh in your mind. Photocopy data sheets whenever you can; if you cannot do this in the field, a useful substitute is to take digital photos at the end of every day. These are usually of sufficiently high quality to be able to recreate your data if it gets lost. In the real world, pages get wet or torn out, so do not forget to number every page or subject (worm, plant, bacterial plate).

A huge number of undergraduate projects cannot be analysed at all or at least in the way that was anticipated. This may be due to design, or loss of data points. Data points are lost by plants or animals dying, experimental plots getting disturbed, or by someone clearing out the fridge in the lab. Murphy's Law is alive and well in most experiments. Although in principle we should be able to get the design correct, it is a good idea to create dummy data using random numbers, and use these to try out your proposed analysis. If the computer does not let you run the analysis you want, it may be that you have insufficient replication or some other problem that you could fix before starting the experiment. If this analysis works, try deleting randomly selected data points to see how robust your design is. If the loss of one or two data points prevents you from running the test you want, you might want to reconsider the design. In addition to the factor of interest, take note of unusual observations that might influence the results. You might be able to account for them later, or it could lead onto another study.

The take-home message is that time spent planning is never wasted. Discussing your experimental design with an experienced researcher will identify most pitfalls and probably throw up more ideas to incorporate. Having a clear idea of what you are doing and exactly what to record saves a lot of time in the field or lab. *If nothing else, carry out a dummy analysis to see if you can analyse your experiment according to your plans.*

4.5 Simulating Your Experiment

You can create dummy data and analyse them for very complicated experiments. However, we will show you how to do this with a simple experimental design which can be analysed with a test we have already covered (the two-sample t-test).

Suppose you have been asked to carry out an experiment to compare the effects of two different insect diets on the amount of frass (insect feces) produced. You have 18 insects available, and you can choose which diet each insect receives. The data collected is the dry weight of frass produced by each insect.

4.5.1 Creating the Data Set

We have 18 sites and 2 treatments, so we need the data in Table 4.1.

You can type this in, or get MINITAB to do it for you. Label the columns, and then go to **Calc>Make Patterned Data> Simple Set of Numbers> Store patterned data in:** Insect **From first value:** 1 **To last value:** 18 **In steps of:** 1 **List each value:** 1 **List whole sequence:** 1

Do this again for Treatment, using first value 1, last value 2, and listing each value once, and you should get the data provided in Table 4.1. This is fine, but we clearly have not randomly allocated the sites to each treatment. We can do this by making use of a trick in MINITAB: **Calc>Random Data>Sample from Column(s)>Sample** 18 **rows from column(s):** Treatment **Store samples in:** Treatment. Make sure that the sample with replacement box is **not** checked. This randomises treatment with respect to insect number, and you should get something like Table 4.2. Note that if you do this yourself, you should get something similar, but not exactly the same because of the randomisation process.

TABLE 4.1

Insect Number and Treatment for Insect Experiment

Insect	Treatment
1	1
2	1
3	1
4	1
5	1
6	1
7	1
8	1
9	1
10	2
11	2
12	2
13	2
14	2
15	2
16	2
17	2
18	2

TABLE 4.2

Treatments Randomised to Insects

Insect	Treatment
1	1
2	2
3	1
4	2
5	2
6	1
7	1
8	2
9	1
10	2
11	1
12	1
13	2
14	2
15	2
16	2
17	2
18	1

TABLE 4.3

Simulated Data for the Insect Experiment

Insect	Treatment	Frass Weight
1	1	12.9013
2	2	12.6835
3	1	10.2610
4	2	9.7528
5	2	11.3427
6	1	14.1971
7	1	11.6366
8	2	13.8637
9	1	12.5817
10	2	13.8720
11	1	11.4658
12	1	11.4404
13	2	11.1036
14	2	10.9567
15	2	12.9425
16	2	10.3909
17	2	12.4214
18	1	12.4716

Finally, you need to simulate the data (the things you will record in the experiment). Do this by: **Calc>Random Data>Normal>Generate** 18 **rows of data Store in column(s):** Frass weight **Mean:** 12 **Standard deviation:** 1.0.

If you already have an estimate of the mean and standard deviation of whatever you are measuring, include it for realism, but it does not matter too much.

The simulated data are shown in Table 4.3.

4.5.2 Analysing the Simulated Data

Carry out a two-sample t-test as described in Chapter 3, and you should get something similar to the following output. The two groups are not significantly different, but this is hardly surprising, as we have just simulated the data using the same mean for each group. What this does tell us is that we can analyse the data in the way we intended and that there are no nasty hidden surprises.

```
Two-Sample t-Test and CI: Frass Weight, Treatment
Two-sample t for Frass weight

Treatment  N   Mean  StDev  SE Mean
        1  9  12.08   1.29     0.43
        2  9  11.96   1.38     0.46

Difference = mu (1)  mu (2)
Estimate for difference: 0.120598
95% CI for difference: (1.217092, 1.458289)
t-Test of difference = 0 (vs. not =): T-value = 0.19
P-value = 0.850 DF = 15
```