Chapter 11 Taking measurements

Additional self-test questions

Q11.1 In the book, we consider a situation where we have seven samples from a control group and seven samples from a treatment group. We said: 'These 14 samples should be measured in a random order (or possibly in an order that alternates control and treatment samples)'. What are the arguments for measuring at random versus alternating individuals between the two groups?

Measuring the 14 individuals in random order is the classical option. There is one argument you could make against it: that a very small fraction of random samples would not be appropriate. For example, imagine that we labelled the control group individuals 1–7 and the treatment group 8–14. We then asked a computer programme to generate a random permutation of the numbers 1–14 for us and it gave '14, 12, 9, 10, 13, 1, 8, 11, 3, 6, 4, 2, 6, 7'. This does not feel like a great permutation, because we measure almost all the treatment individuals before any control individuals. We have discussed cases like this in the book. In a situation like this you should reject this permutation (before taking any measurements) and ask the computer for an alternative one. The only problem with this philosophy is that there is no hard and fast rule for what is an acceptable permutation. If the permutation was '14, 2, 9, 10, 13, 1, 8, 11, 3, 6, 4, 12, 6, 7', the treatment individuals still hold six of the first eight positions, but the case for obviously rejecting it is less strong than before. Such permutations that seem 'not quite mixed up enough' will be very uncommon (and increasingly so as sample sizes increase), but they can happen. Alternating would be a way to avoid this problem. The only practical argument against alternation is that it would give you very unreliable results if your measurement tool or technique had some systematic but intermittent fault that only occurred every second measurement. In our case such a fault might kick in for all the control measurements but none of the treatment ones. This does sound highly unlikely, but not utterly impossible. So in truth, we would not be concerned by either approach.

We don't deny that alternating is easier to implement, but not by very much, so we don't think that on its own is a valid reason for opting for alternating.

Q11.2 During an angling competition on a small lake, you want to record the weight of each pike caught and the time it was caught, to allow you to look for an effect of time of day on the behaviour of different sizes of fish. How would you minimize inaccuracy and imprecision?

To avoid imprecision, we are going to need good-quality scales, say accurate to 5 g or 1 g over a range up to 20 kg. Just as important, you must make sure you have a clear methodology for how the fish is weighed. This is particularly important if you need to employ several field assistants to cover the whole lake (remember there are ethical issues related to how long the fish has to wait on land to be measured; anglers will want to land it immediately to avoid losing it). Importantly, the scales should be cleaned regularly to avoid build-up of dirt and/or mucous from the fish. To avoid inaccuracy (bias), you (or your field assistants) should do the measuring, not the anglers. Indeed, you should probably do the measuring out of sight to the anglers and without releasing your results to them immediately, to avoid them seeking to influence you. The scales should be calibrated regularly with a range of standard weights throughout the day (perhaps after every time you clean them).

Q11.3 You want to compare activity of chimps in Berlin Zoo and the Bronx Zoo in New York: how will you ensure consistency of measurement?

You should certainly make sure that measurements are taken at the same time of year in similar weather conditions and when nothing unusual (veterinary interventions, say) has been occurring at either site. Perhaps you should arrange for the two enclosures to be videotaped. These tapes could then be analysed by people in quick succession (avoiding problems of drift); indeed tapes could be spliced so that the observer does not measure all of one site before turning to the other. This approach avoids the standardization issues involved with using one person to measure in one place and a different person to measure in the other. However, you might want to get several people (independently) to score behaviour from the tapes. You might struggle to make the tape watchers blind to the fact that there are two sites (because of differences in the appearance of the enclosure and the chimps) but you should be able to blind them to the actual identities of the sites.

Q11.4 Imagine that as part of a citizen science project you organize members of the public to walk designated same-length transects around a major city at the same time and record the birds that they see. How best should you obtain consistency?

You need to specify a walking pace, and perhaps encourage people to practise so they all walk at the same pace. You should perhaps offer some advice to standardize on clothing also (encouraging muted colours for example). You should make clear that people must do this alone, and not use their mobile phone and things like that (consistency of appearance and behaviour). You should offer guidance on whether to use binoculars or not, whether to use bird calls, or whether they must actually see a bird in order to record it as present. You

must draw up a standardized data sheet for recording the birds people see and agree on actually how that sheet would be filled in, and preferably trial this; this trialling could be done as you work with volunteers to standardize their bird-recognition abilities. This can be done by giving a lecture on bird recognition, followed up by going out in pairs or small groups in pilot studies to pool knowledge, compare ideas, and work towards agreed standardization.

Q11.5 Explain floor and ceiling effects in your own words. Why are they a problem in experimental design?

A floor effect occurs when all the subjects score the lowest possible value, regardless of the treatment group they are in. In a ceiling effect they all get the highest score. If we were looking for an intelligence difference between male and female students and looked at their score in simple questions (like 'what is 5 x 4?'), we would fall foul of a ceiling effect. Because our test is inappropriate, there could be a difference in intelligence between men and women but our test cannot pick it up. We need to design a test where we expect there to be some between-individual variation. There is a bit of a trade-off here, in that too much between-individual variation reduces statistical power to detect a difference, but a floor or ceiling effect makes detecting a difference near impossible. So you want some between-individual variation, but not too much.

Q11.6 In your own words, explain why it would be good for a person measuring experimental units to be blind to the treatment group that a particular unit belongs to.

This helps to avoid any accusation of conscious or unconscious bias.

Q11.7 When would this be impractical?

It becomes very difficult when the groups are very different from each other. If the person is tasked with weighing male and female sparrows, it is difficult for them to directly handle the birds and be unaware of the sex of an individual.

Q11.8 If blinding would be desirable but impractical then what mitigating steps can you take?

The key thing is to avoid subjectivity that requires a measure of judgement; there is much more scope for bias if someone is looking at birds and scoring them as 'fat' or 'lean' rather than weighing them on electronic scales. It may also be possible to blind them to the hypothesis under test. If they do not know the full picture, there may be less scope for introduction of bias. Of course, in the case above you could avoid this problem by having one person measure all the male birds and another person weigh all the females; but this would be unwise.

Q11.9 A new vegetarian dog food claims to give improved coat condition within two weeks. How would you test this claim?

You are going to randomize individuals to a treatment or control. The key thing is that the control dogs must also have their diet changed (just not to the vegetarian food), or else any difference you might pick up might be due to change of diet rather than change to a particular diet.

Q11.10 In the experiment above, how would you measure coat condition?

We think what matters here is owners' perception rather than any mechanistic measure of coat quality. We would ask all the owners to meet up at the start of the trial and score all the dogs other than their own on coat condition; then the owners meet up again and repeat the exercise after the trial. You would need to take steps to make sure the owners do not talk to each other (during the trial or when they all meet up) about which group their dog is in. Perhaps owners should be blind to this. Rather than scoring all the dogs, you might have thought to ask them just to pick the five with the best coats, but this would be a less powerful test; it could be that the food has an effect but you don't pick it up because there are just five dogs in the group that started off so far ahead of the rest that although the food changes things it does not change these five from being the front runners. If owners are very variable in their choices, you might get an expert (say, a dog groomer) to score all the dogs before and after. They would have to be blind to which group each dog is in.

Q11.11 Exam books are designed so that the name of the student is unavailable to the marker. Discuss the reasoning behind this in terms of blind procedures.

The procedure removes the perception that the marker might be biased by such factors as personal dislike for a particular student or an expectation that a certain student is gifted.