Chapter 3 Selecting the broad design of your study

Answers to additional self-test questions

Q3.1 Using newspaper articles on whale strandings from back issues kept in a university library, a researcher finds that there are on average 24 reported instances of whale strandings a year around the Scottish coastline from 2000 to the present, but only an average of 13 a year from 1990–1999. They conclude from this that this increase is likely to be caused by increasing numbers of whales using Scottish waters due to climate change. Comment on the plausibility of this explanation in comparison to alternatives.

It seems unlikely that this size of increase can be put down to an increase in the discovery of strandings rather than an increase in strandings themselves. The Scottish human population size, or spread, or propensity to visit beaches hasn't changed that dramatically over the last 25 years. It is possible that the likelihood of a stranding being reported in the newspapers has increased because of a greater interest in environmental issues by the media, however we'd be surprised if this could explain such a spectacular increase. Thus, it probably is the case that there is genuinely an increased number of strandings. This need not suggest a change in the numbers of whales in Scottish waters; it may signify a change in behaviour (say feeding more regularly near to shore because of a change in food distribution), or it could be driven by the recent emergence or spread of a disease in the local whale population causing increased numbers of dead whales being washed ashore (although most dead whales sink). A disease could trigger a change in behaviour, as could reduced submarine (and thus sonar) activity around the West of Scotland with the end of the Cold War. Even if there is a rise in local whale numbers, this need not be to do with climate change, it could be about reduced persecution allowing the population to recover to previous levels, or increased persecution in Iceland, Faeroes, and Norway driving whales to seek shelter around Scotland.

Q3.2 Ecologists commonly use measures such as clutch size, feeding rate, and mass per unit length as indirect measures of fitness. What are the limitations of this, and why do ecologists persist in using these indirect measures in the face of these limitations?

Fitness is very challenging to measure directly. To compare the fitness of two female gannets, we would want to record the number of female offspring that each female produces during her lifetime that survive long enough to reproduce themselves. Given that

these birds can live for a decade and that offspring will disperse over hundreds if not thousands of kilometres, measuring fitness directly would be very, very challenging. This helps you see the attraction of indirect measures. Now, all other things being equal, if we find that in a given year the mean clutch produced by females of phenotype A is greater than the mean clutch produced by individuals of phenotype B, then this should lead to greater fitness for the genes underlying phenotype A. How good the indirect measure is depends on how reasonable the assumption of all other things being equal is. In our illustrative example, if there is a trade-off between annual clutch produced by an individual and its probability of surviving the winter, then our conclusion that maximizing clutch size maximizes fitness could be wrong. Hence, we use indirect measures of fitness because fitness is very hard to measure directly; however we must use our understanding of the general biology of a system to consider how reliable candidate indirect measures of fitness are likely to be.

Q3.3 Can you think of a scientific study where ethical considerations might drive you to using indirect measurements?

If we wanted to study how increasing amounts of alcohol in someone's bloodstream affects driving ability then it would be ethical to measure driving performance using a simulator rather than a real car.

Q3.4 Imagine that a study did find that people who preferred butter were better drivers than those that preferred margarine. Can you think of any hypotheses that could explain this finding? Which of these do you consider most plausible?

It seems very unlikely that this dietary preference influences driving ability directly. Butter or margarine makes up a very small part of our diet and contains little (in terms of nutrients or trace elements) that is not readily obtained from other commonly consumed foodstuffs. It is possible, but unlikely, that something present in butter but not in margarine has a beneficial effect on eyesight, alertness, or reactions. Similarly it is possible, but unlikely, that something present in margarine but not in butter has a detrimental effect on some aspect of physiological function that degrades driving ability.

Slightly more likely, but still unlikely, is that preference for butter or margarine is an indication of a wider difference in dietary preference between people, and this wider preference affects driving ability directly. For example, it could be that those that prefer margarine have a general preference for refined, heavily-processed food products over unprocessed foodstuffs. This could lead to a dietary deficiency that does lead to a decrease

in, say, alertness and so driving ability. This is more plausible than our first explanation, but still relatively unlikely compared to the answer below.

The most plausible explanation is a third variable effect. That is, that preference for margarine or butter has no effect whatsoever on driving ability, but is linked to some other variable that does affect driving ability. For example, women might be more likely than men to prefer butter to margarine and might be better drivers than men. Here the third variable is sex, which is related to both the dietary preference and driving ability. Other plausible third variables are age and socioeconomic group.

Q3.5 How would you the measure driving ability in a study like the one described above?

In the UK, the quickest and cheapest way would be to record the number of penalty points on a person's driving licence. This is inaccurate since a person's probability of having penalty points will be correlated with some but not all aspects of poor driving. Worse, it will be affected by the length of time for which they have held a licence, the amount of driving they do, and the type of driving they do (e.g. business or pleasure). Lastly, penalty points can in substantial part be a matter of good or bad luck.

It would be better to make each participant sit a driving test. This would be considerably more work. It would also be better to ask the examiner to provide a quantitative score of ability rather than just 'pass' or 'fail'. One person could fail because they coped poorly with a traffic situation that another person with identical driving ability does not experience on their test. This heterogeneity in driving situations experienced by the different people during tests leading to heterogeneity of scores is a problem. One way to cope with this would be to make every candidate sit a number of tests, but this is lots and lots of hard work. It might be better to use a simulator so that each candidate experiences identical sets of (simulated) traffic conditions. Again, we could get a qualified driving test examiner to assess performance, rather than attempting to do this ourselves. However, is the way they drive in these test conditions reflective of the way they would drive normally?

Q3.6 In the tail length experiment discussed in the book, we want to have a control group in which tail length is unaltered. Why then do we bother to cut the tails off then glue them back on in exactly the same position?

Other than tail length, we want as few things to be different between the three groups of birds as possible. It is just possible that the trauma of having the tail removed, or our ability to reattach the tail properly, has an effect on the subsequent behaviour of the male birds. It

is even just possible, if unlikely, that the smell of the glue influences female mate choice. By taking the trouble to perform these manipulations on all groups, we control for any of these possible confounding factors.

Q3.7 Discuss how you would test the hypothesis 'women find blue eyes more attractive than brown eyes' by correlational and manipulative means. Discuss the pros and cons of each and which you would adopt to address this question.

Correlation: We set up a cocktail party with 20 men, of similar age and occupation, ten with brown eyes and ten with blue. We also invite 20 women, each independently tasked with identifying the five most attractive men in the room. We then explore whether more blue-eyed than brown-eyed men were selected as the most attractive. Alternatively, this process could be done with a pack of passport-style photographs, asking women to score the attractiveness of each. This is quicker and easier, and removes added noise due to personality differences, height, build, clothing, etc.

Manipulation: In the cocktail party set-up, we could run several parties, with each man wearing clear contact lenses to some parties and coloured ones (designed to make them appear as if they have the opposite eye colour) to others. Different women are invited to each party, but we can explore whether the manipulation makes men appear more or less attractive. With the photograph set-up, we would use image analysis software to change the eye colour. A given individual's photograph would feature in each woman's pack of photos, but sometimes he would have brown eyes and sometimes blue.

The cocktail party set-up is a great deal of hard work and introduces possible confounding factors, so we would choose to use photos instead. The concern we have is potential third variables. For example, maybe women actually find dark hair attractive and having dark hair is correlated with having blue eyes. This concern would drive us towards the manipulative approach, but we would need to be very careful that we doctored photos carefully so as to produce realistic looking eyes. It would probably be worthwhile running a pilot project to see if women can detect whether or not a photo has been doctored. If we cannot produce a biologically-realistic manipulation then we will have to go for the correlational approach; we could record potential third variables such as hair colour and look for correlations between these and eye colour that might give rise to third variable effects.

Q3.8 The book suggests that women who go to university are less likely to marry than those who do not. However, the book argues that we should not conclude from this that studying at university in itself causes a reduction in a woman's propensity to get married. Explain this reasoning in your own words.

The initial observation comes from an un-manipulated correlational study, and it may be that a third variable can explain the apparent link between going to university and propensity to marry. For example, it may be that individuals from higher socioeconomic groups are both more likely to go to university and (quite separately) less likely to marry. Other potential third variables include ethnic group, religion, or geographic region.

Q3.9 Imagine a student is exploring the question of whether more or fewer birds are seen in a public park on rainy days than on sunny days. Their definition of a rainy day is 'if during my period of observation I see anyone in the park with an umbrella raised, then it's a rainy day; otherwise it is dry.' Discuss the appropriateness of this rule, and see if you can come up with an alternative which you feel is better.

In one sense, this is a good rule, because it is clearly defined. It is hard to imagine a case that cannot unambiguously be recorded as rainy or dry on the basis of this rule; there are no grey areas that would lead to you being unsure how to score a certain observational period. However, it suffers a problem in that although days can always be unambiguously defined as wet or dry, the categorizing of a day by this means may clash with common-sense definitions of wet or dry. It may be so wet that there is no one in the park to put their umbrella up! More likely, if it starts raining in summer time, no one puts their umbrella up because they haven't brought it with them. Similarly, strong winds may deter umbrella use. We don't think using people works at all; people can also put hoods up for all sorts of reasons and the definition of raincoats is fraught with potential ambiguity. You need a measure that has nothing to do with other park users. We think the key here is that ambiguous cases are going to be where there is simply a few spits of rain that doesn't really develop into anything and such rainfall is unlikely to have a big effect on birds; thus, recording days without substantial rain as dry days is probably OK. We're not sure we need a strict definition of substantial rain, since we'd almost all agree on all cases, but you might go for something like 'rain, sufficient to make the pavements noticeably wet or cause sufficiently frequent raindrops on a puddle that their number could not be counted with the naked eye'.

Q3.10 A driver in their twenties is three times as likely to be involved in a road traffic accident as a driver in their sixties. One explanation for this could be that people become safer drivers as they get older. Can you think of any likely third variable effects that could provide an alternative explanation for this observation?

The older age group will include large numbers of retired people whose driving is done mainly for pleasure; this will be much less true for the younger age group, a significant fraction of whose driving will be done for work purposes. This may mean that younger drivers spend a larger fraction of their time driving in rush-hour traffic or in poor weather conditions (when your boss would expect you at work, but you would not choose to go out for a pleasure trip). Further, business driving will often involve driving to somewhere that has to be reached by an appointed time, a pressure that is less common for pleasure trips. Further, it is possible that younger drivers simply drive more miles in a year than older drivers.

Q3.11 Can you think of an alternative explanation for the observation of Q3.10 in terms of reverse causation?

Simply, it may not be that driving ability improves with age, but rather that only good drivers live to old age. Under this argument, a given individual's driving ability could remain unchanged through life but we would still see a change in the mean ability across the population with age because poorer drivers are more likely to die in road traffic accidents or get banned from driving and so contribute to the population of younger drivers more than they contribute to the population of older drivers.

Q3.12 Which of the explanations above do you think is the most important factor explaining the three-fold difference in accident rates between these age groups?

The reverse-causation effect seems unlikely to be able to explain such a large difference. At most 4000 people a year die in road traffic accidents in the UK, and only the minority of these will be drivers. Perhaps the number of people who kill themselves, get banned, or scare themselves so badly that they voluntarily give up driving through their bad driving is of the order of 5000 a year. This is a very small fraction of the perhaps 40 million drivers in the UK. Hence, this seems unlikely to provide an explanation. It is more likely that some combination of the conventional explanation and the third variables explains the effects. Our guess is that the effect of difference in type of driving (business or pleasure) is likely to be the most important third variable. Although it is quite possible that people become safer drivers as they get older, it would seem surprising if they become safer to a sufficient extent to explain a three-fold reduction in annual risk without third variables being involved to some extent.

Q3.13 The book discusses the interpretation of a negative correlation between badger weight and number of parasites. The conventional interpretation is that the parasites lead to a reduction in body weight. Is a reverse causation explanation where low body weight leads to increased parasite burden plausible to you?

Not really. There are lots of third variable possibilities (hungry badgers might have a low body weight and might also eat otherwise unattractive food courses that increase vulnerability to parasites; unwell badgers might have low body weight and a compromised ability to defend themselves against parasites). But the only plausible reverse causation mechanism we can think of would be that parasites are particularly attracted to (or able to attach to) individuals of low body weight, and this seems unlikely.

Q3.14 If a carefully performed study using samples of sparrows caught around Edinburgh (where Nick lives) suggested that males had higher parasite loads than females, how far would you be comfortable generalizing from this sample? That is, to what extent would you expect the results from the sample to generalize to sparrows in Edinburgh, urban sparrows, UK urban sparrows, or UK urban birds?

If the samples were taken from Edinburgh, it seems like the population relevant to the project was sparrows in Edinburgh. If the sampling was appropriate, then the results should generalize to Edinburgh sparrows more generally. Strictly speaking, you should not consider their results to generalize any further. If we were interested in UK urban sparrows, then we should have sampled a range of UK cities. UK cities can be expected to differ in parasite prevalence: consider the very different climates of Edinburgh and Cambridge. By similar reasoning, you can see good biological reasons why you would not be comfortable generalizing to all Scottish sparrows or to all small passerines in Edinburgh. The moral is: think about the research question you are interested in, define your population and thus your sample accordingly, and be very cautious in speculating on other populations that you've not sampled.

Q3.15 A major scientific journal reported that school pupils who play a musical instrument performed better in a general test of memory power than those who do not. From this they concluded that playing a musical instrument improves memory power. Do you consider their conclusion premature, and if so how would you go about collecting more data to test this hypothesis?

The conclusion is premature because there could be a third variable at work. Most likely, being from a privileged background (or a 'good school') is related to both increased likelihood of playing an instrument and increased likelihood of trying harder in the test or

generally being good at tests. How to test this hypothesis by removing these third variables is challenging, as randomly assigning pupils to either learn an instrument or not learn an instrument is not really feasible: how can you ask someone to not learn an instrument? We think the best approach might be to test people before they begin to learn the instrument and again afterwards, to see if their memory score improves. Of course, this could be a simple effect of age, so we would also want a control group of children who don't learn an instrument to see if their scores improve also. We still haven't solved the third variable problem since people will self-select themselves into the instrument or control group. However, we think the best we can do is record all likely third variables and see if they differ significantly between the control and treatment groups. This is an example of where our ability to carry out the most effective experiment is hampered by practical and ethical considerations.