

## Session 1

### Chapter 1 & Chapter 2, Section 2.1

**1.1 Imagine the research task that you want to address is to estimate the fraction of the electorate who plan to vote in the next elections for your national government. You plan to do this by surveying a sample of the appropriate statistical population. Define this population in your own words.**

The population we are interested in is all those people who will be eligible to vote in the next set of national elections. Notice, this is not all the people in the country, since not everyone has the vote. It's not even all the adults, since not everyone is registered to vote and people in prison (for example) may not be eligible to vote. Strictly speaking, it's all the people on the voters roll (the official list of registered voters, also called the Electoral Roll) at the time of the next election, but the current voters roll may be a satisfactory approximation to this.

It may **not** be, however, because there may be a surge of people registering just before the elections, because of government campaigns, and the generally high profile of the elections in the media. Although these people may make up a small fraction of the electorate, they may have a disproportionate effect on the survey, because having taking the trouble to get registered they are more likely to use their vote.

However, by and large, I'd say the current list of registered voters is a good approximation to the population that we are interested in; this approximation is unlikely to influence our results by more than a percentage point or two, which is probably accurate enough for the likely use that our research will be put to.

**1.2 In the example of question 1.1, how might you choose a representative sample for the survey?**

Let's assume that the population is the current voters roll. In the UK, this is publicly available data. I think we could load all the names into a database, and assign each a number. We then get the computer to generate a sequence of random numbers from 1 up to the maximum number until we have the desired sample size. We must now contact each of these people to ask them the question. We might initially do this by telephone. There are several possible outcomes.

Firstly, they may answer our question, or decline to answer our question. If so, we have finished trying to sample them. Alternatively, we may be unable to contact them this way (perhaps they are ex-directory, or have no phone or are never in when we call). We could then follow this up with sending a letter or "doorstepping" them.

The latter must be a matter of last resort as this will be expensive and time consuming. I think we need such an active approach, as simply writing to the individuals may not be the best way of gathering the data. You might argue that people who do not reply to the letter are generally apathetic people who are less likely than the average person to use their vote.

Notice that you could simply give up on people if you can't reach them by telephone. However this could lead to the introduction of unintended bias in your sample. For example, retired people are more likely to be at home to answer the phone than people who work, and so could be over-represented in your study.

**1.3 In the example of question 1.1, what number of people would you aim to gather data on?**

I think it is unlikely to make much difference to the political party or the newspaper commissioning the survey whether the answer is 45% or 46%. So I think sample sizes like 10,000 would be overkill (despite being a small proportion of the total population). I think 1000 would still be on the high side. We can see from the discussion of question 1.2 that there is significant effort required to find a phone number and even more to doorstep. I think that how many people you approach depends on the method you use to approach them.

If you are simply mailing them and asking them to fill in a freepost form and post it back [and, as discussed above, I don't think this is a good method], then you might go for 500, expecting at least 100 [and more likely 200-300] to actually post a response back to you. If you are going for the "telephone call, then letter, then door-stepping" approach then the effort required for each person sampled increases substantially, but I think the fraction of people that actually give a response will increase from perhaps 40% to 80%. This suggests that a number like 200 or 250 would be reasonable.

I still think this would be a considerable amount of work, and my temptation would be to think about whether 150 or even 100 might be enough. I think less than 100 would be hard to justify.

**1.4 The book argues that if we were surveying the food preferences of different nationalities, then using five members of one family as five independent measurements of the food preferences of the nationality of that family would not be statistically valid. Explain the reasoning behind this in your own words.**

Five members of a family would not be independent measurements because they have shared experiences that they do not share with other members of their nationality. For example, it may be that the mother does almost all of the cooking and is particularly terrible at cooking rice. If this is true then we might expect that all family members report a low preference for rice dishes. Alternatively, it might be that because one family member is allergic to pulses, that family's meals never contain pulses and so again, we might expect all family members to report low preferences for dishes containing pulses. The family as a whole may have had a very memorable family holiday in India, as a result of which all family members report a high preference score for Indian dishes.

The rule you should use is if knowledge of how individual A responds makes you better able to predict how B will respond (from the range of possible responses), then A and B are not independent samples.

**1.5 Are there any variables you can think of where you might be able to justify using members of the same family as independent sampling units?**

It is very difficult to think of a variable where shared genetic inheritance and shared previous experience do not influence results. Even something like “favourite colour” could be affected by genetics or by granny’s unusual fashion sense. Perhaps something like aspects of fingerprint might be alright. However, in general we can assume that family members are almost never independent sample units.

### **1.6 Are there ethical considerations associated with the surveys described above?**

Absolutely! Firstly, there may be ethical issues associated with how the data in these surveys are used. You would hopefully be concerned if the surveys above were misused to support a racist agenda. There are ethical considerations with the storage of the data. For example, in the UK, if you keep data on anyone then you must comply with the provisions of the Data Protection Act. If you keep the data then you should be very wary of subsequently using it for any purpose other than the one that you originally collected the data for. For example, people might be comfortable cooperating in your voting intentions survey if they know that you are conducting it for a conservation organisation, but would be angry if you then sold all or some of your data to a multinational corporation.

There are ethical issues associated with collecting the data in the voting intentions survey. However we were careful to use publicly available data from the voters roll (for names and addresses) supplemented by directory enquiries (for telephone numbers). But wait, directory enquires provide information to normal users for normal purposes; we should check whether it is legal to use it for our more commercial purpose. There are ethical issues surrounding each of telephoning, mailing and (particularly) “doorstepping” (involving invasion of privacy, harassment, causing unnecessary stress), and we should think each of these through carefully.

### **1.7 An early draft of the book contained the following real example, but the authors were persuaded that it was both a flawed way of collecting data and ethically objectionable.**

A social scientist wanted to find out what papers people in a given street read. The problem is that people can be expected to lie when you ask them, saying that they read a serious “highbrow” paper when they actually read a less serious paper or no newspaper at all. The social scientist got around this by saying that he was piloting a paper recycling scheme, providing people with sacks and asking them to deposit all their waste paper in a sack which he collected at the end of the week. This allowed him to evaluate which papers were read by examining the papers placed in the bags. Having obtained this information, he took the contents of all the bags to a paper bank.

#### **Why is this unethical and scientifically flawed?**

The scientist is collecting information by deception. I hope you can see that this is only justified in very unusual circumstances. Information on people could now be stored without their knowledge that this data about them had been collected. It may seem like a harmless deception, but that argument feels like the thin end of a wedge. It is poor data because it collects papers that end up at home: it misses papers that are read on the bus

or train to work and then discarded. Also, because a paper ends up at home does not mean that all the members of that household have read that paper.

**1.8(a) Salmon parr eat less per day in winter than summer. Suggest at least three hypotheses that could explain this.**

- (i) The fish is primarily eating invertebrates drifting downstream and in the benthos, there may simply be a lower density of such prey available in winter than summer. Hence it may eat less each day because less food is available.
- (ii) They are probably visual feeders, and shorter day lengths, lower light levels, and heavier sediment loads in the water in winter all reduce the fish's ability to find prey (regardless of the prey's abundance)
- (iii) Flow rates are probably higher in winter, and so drifting invertebrate prey may be harder to catch, because they are faster moving.
- (iv) The fish may be less able to move at the speeds required to catch many prey at the lower water (and so body) temperatures experienced in winter.
- (v) Fish in winter may be concerned only with over-winter survival, rather than with growth and reproduction, and so their reduced feeding rate may simply represent reduced motivation to feed.

**1.8(b) How would you critically test between these hypotheses?**

- (i) This can be addressed in a field study. We simply need to sample the amount of food drifting down those streams with salmon parr in them at a range of times throughout the year. Exploring availability of benthic prey will be more challenging in the field, but not impossible.
- (ii) This probably needs investigation with fish in a flume in a laboratory, where we can change light levels, day length, and sediment load in a controlled way without introducing confounding factors such as temperature changes. These confounders would be a problem in a field study.
- (iii) Same arguments as (ii)
- (iv) Again, I think you'd need a laboratory study to separate this hypothesis from (ii) and (iii)
- (v) Now this one would be difficult to test directly, because motivation is a little difficult to measure. It is hard for us to know whether a fish does not respond to passing prey because it fails to detect it or because it detects it and decides not to attack. However, I think we can do this in the laboratory. If we keep temperature, light levels, sediment load, and prey availability the same and find lower attack rates (rather than capture rates) in winter than in summer, then this would suggest reduced motivation to feed in winter.

The problem with this is the artificiality that we impose. We need to keep temperature constant to eliminate (iv) but it could be that the fish uses temperature as a cue to switch to its wintertime motivational state. In the end, it may be that support for (v) can only be gained by elimination of alternative hypotheses.