Selection biases

Chris Wild, University of Auckland

In this course, we are using the word "selection" in a very broad sense to talk about the process whereby some things end up having data on them recorded and ending up in our data set while other things do not get data recorded about them.

So we talk about selection in the sense of *"what happened"* rather than the more normal English meaning of a choice made on purpose.

The selection-biases people tend to be most familiar with are the biases in polls and surveys from causes other than sampling (statisticians call them "non-sampling errors") and those are what we will discuss here.

In fact, biases or errors are generally biggest in the "non-scientific" polls or surveys that do not use sampling. Because they are so notoriously unreliable, it is better to think of the popular **non-scientific** (opt-in, or voluntary) **polls** conducted by television shows, magazines and websites as entertainment, not information.

The biases that arise this way are referred to as **self-selection bias**. The people who do contribute are often very different from those who don't and so are not representative of the general population. For example people with very strong opinions about an issue will tend to be heavily over represented. "Scientific" polls and surveys with **low response rates** are likely to give biased results for the same basic reasons (there it's called *non-response bias*).

With the above and what follows, we will not go into any details. The purpose of this article is just to alert you to areas where there are issues that you will need to find out about if you are going to conduct a serious survey. Or in the words or images used in ancient maps, an alert that "here be dragons."

There are a host of things that you might never think of that can have a significant **influence** on **the way people answer a question** such as information in the survey about why it is being done, differences in wording (you will probably be aware of *loaded questions*; see also this TV clip on YouTube), and questions that have gone before (which can bring something to the front of someone's mind that they would not have otherwise have considered);

There are many **social factors**. For example, unless we are totally convinced that our information will never be connected with us by any other person, providing information about ourselves is a social interaction so that concerns about how we will be seen by others can be more important than providing a true picture. Interviewer effects can be important. On the plus side, survey organisations know that there are some interviewers who people really tend to open up to - providing information on very personal things that they would never tell someone else. In terms of getting to the truth, and low non-response bias, such interviewers are gold. On the negative side, the social tendencies to tell people what you think they want to hear and a desire to present oneself in a good light may cause biases if the persona of an interviewer is somehow related to the issues being asked about.

Our last topic is errors connected with **inadequate** sampling frames. In cases where it is a practical possibility, survey samplers would sample from a list of everybody in the population of interest (the *target population*). Mostly they have to settle for the best approximation that can be practically implemented which provides the *sampling frame*. The approximation has often been something like an electoral roll (lists of everyone eligible and registered to vote). Or they will sample by sampling from the available (landline) telephone numbers coupled with some strategy to choose a person in the house that had that number. This used to work really well but now the fraction of houses without landlines has become substantial. So the mismatch between the sampling frame and the population of interest – the population we want to project our findings to – is the source of frame errors.

Statistics only justifies an extrapolation of findings from the sample taken to the population that was actually sampled from. The issues around inadequate sampling frames are a special case of **the general issue of** transferring or **generalising statistical findings beyond what is justified by the study design**.

Any extrapolation from the generalisation the study design permits is not justified by statistics. It has to be justified by something else. In the best cases this will be information from other sources, or perhaps some sort of theoretical, subject-matter understanding. But often people are just operating from some combination of "this is the best information we have/can get", "I can't see any reason why things would be different over there", and hope. And there may be large numbers of worms in those particular cans. But to reiterate, the further generalisations are not justified by the statistics.