Probability and Statistics: To p, or not to p?

Module Leader: Dr James Abdey

4.2 Random sampling

We have previously seen that the term **target population** represents the collection of units (people, objects etc.) in which we are interested. In the absence of time and budgetary constraints we conduct a census, that is a total enumeration of the population. Its advantage is that there is **no sampling error** because all population units are observed and so there is no estimation of population parameters. Due to the large size, N, of most populations, an obvious disadvantage with a census is cost, so it is often not feasible in practice. Even with a census **non-sampling error** may occur, for example if we have to resort to using cheaper (hence less reliable) interviewers who may erroneously record data, misunderstand a respondent etc.

So we select a sample, that is a certain number of population members are selected and studied. The selected members are known as elementary sampling units. **Sample surveys** (hereafter 'surveys') are how new data are collected on a population and tend to be based on samples rather than a census. Selected respondents may be contacted in a variety of methods such as face-to-face interviews, telephone, mail or email questionnaires.

Sampling error will occur (since not all population units are observed). However, non-sampling error should be less since resources can be used to ensure high quality interviews or to check completed questionnaires.

Types of error

Several potential sources of error can affect a research design which we do our utmost to control. The 'total error' represents the variation between the true value of a parameter in the population of the variable of interest (such as a population mean) and the observed value obtained from the sample. Total error is composed of two distinct types of error in sampling design.

- **Sampling error** occurs as a result of us selecting a sample, rather than performing a census (where a total enumeration of the population is undertaken).

  - It is attributable to random variation due to the sampling scheme used.
  - For probability sampling, we can estimate the statistical properties of the sampling error, i.e. we can compute (estimated) standard errors which facilitate the use of hypothesis testing and the construction of confidence intervals.

- **Non-sampling error** is a result of (inevitable) failures of the sampling scheme. In practice it is very difficult to quantify this sort of error, typically through separate investigation. We distinguish between two sorts of non-sampling error:
  - **Selection bias** – this may be due to i. the sampling frame not being equal to the target population, or ii. in cases where the sampling scheme is not strictly adhered to, or iii. non-response bias.
  - **Response bias** – the actual measurements might be wrong, for example ambiguous question wording, misunderstanding of a word in a questionnaire, or sensitivity of information which is sought. Interviewer bias is another aspect of this, where the interaction between the interviewer and interviewee influences the response given in some way, either intentionally or unintentionally, such as through leading questions, the dislike of a particular social group by the interviewer, the interviewer's manner or lack of training, or perhaps the loss of a batch of questionnaires from one local post office. These could all occur in an unplanned way and bias your survey badly.

Both kinds of error can be controlled or allowed for more effectively by a **pilot survey**. A pilot survey is used:

- to find the standard error which can be attached to different kinds of questions and hence to underpin the sampling design chosen
- to sort out non-sampling questions, such as:
  - do people understand the questionnaires?
  - are our interviewers working well?
  - are there particular organisational problems associated with this enquiry?

Probability sampling

Probability sampling techniques mean every population element has a known, non-zero probability of being selected in the sample. Probability sampling makes it possible to estimate the margins of sampling error, therefore all statistical techniques (such as confidence intervals and hypothesis testing – covered later in the course) can be applied.

In order to perform probability sampling, we need a **sampling frame** which is a list of all population elements. However, we need to consider whether the sampling frame is:

i. adequate (does it represent the target population?)

ii. complete (are there any missing units, or duplications?)

iii. accurate (are we researching dynamic populations?)

iv. convenient (is the sampling frame readily accessible?).

Examples of probability sampling techniques are:

- simple random sampling
- systematic sampling
- stratified sampling
- cluster sampling
- multistage sampling.

In this and Section 4.3 we illustrate each of these techniques using the same example from Section 4.1, i.e. we consider a class of 25 students, numbered 1 to 25, spread across five classes, A to E.

Simple random sampling (SRS)

In a simple random sample each element in the population has a known and equal probability of selection. Each possible sample of a given size, n, has a known and equal probability of being the sample which is actually selected. This implies that every element is selected *independently* of every other element.

Suppose we select five random numbers (using a **random number generator**) from 1 to 25. Suppose the random number generator returns 3, 7, 9, 16 and 24. The resulting sample therefore consists of students 3, 7, 9, 16 and 24. Note in this case there are no students from class C.

A	B	C	D	E
1	6	11	$\boxed{16}$	21
2	$\boxed{7}$	12	17	22
$\boxed{3}$	8	13	18	23
4	$\boxed{9}$	14	19	$\boxed{24}$
5	10	15	20	25

SRS is simple to understand and results are readily projectable. However, there may be difficulty constructing the sampling frame, lower precision (relative to other probability sampling methods) and there is no guarantee of sample representativeness.