

Probability sampling is important in the social sciences because it allows us to take information from a 'small' number of 'sampling units' (e.g. people) and represent the population of interest with a known level of statistical 'confidence'. However, for these analyses to be accurate, the sampling must fulfil strict criteria.

**Random selection from the population of interest** – This implies truly random selection – *you cannot pick the sampling units yourself!* Random number generators are used to do this. It is vital that you are selecting units from a good 'sampling frame', i.e. a list of all possible units in the population of interest. If this list is not complete in any way, your sample could be biased.

**Sample size considerations** – Working out the sample size required to identify relationships in the data depends on several different considerations. For example:

- The expected strength of relationships between variables;
- The number of variables wanted in a statistical model;
- The precision of measurement instruments (i.e. the accuracy of questions in a survey); and
- The size of the population from which the sample comes.

Table 1 shows some population sizes along with the minimum sample size and maximum number of variables these samples would allow<sup>1</sup> in a statistical model<sup>2</sup> with a categorical response variable. When the population is small, you need a larger proportion of the population in your sample to obtain accurate estimates, and the number of variables that can be included in a statistical model is limited. As a rule of thumb, the number of variables allowed in a model will be 10% of the sample size. Therefore, if you want to include more variables, you must increase your sample size. It is also common practice to over-sample so that missing data will not reduce the sample size below the minimum required, e.g. if you expected a response rate of 50% and the population size was 500, you may as well attempt a census.

Table 1: Population, sample size, and maximum model variables<sup>1</sup>

Population size	Categorical response	
	Min. sample size (% Pop.)	Max. model variables <sup>2</sup>
100	80 (80%)	8
500	218 (44%)	21
1,000	278 (28%)	27
10,000	370 (4%)	37

**Independence of sampling units** – Knowing whether your sampling units are independent is vital when analysing sample data. If we took a random sample of 1000 pupils at 50 schools and related their improvement in academic achievement to school characteristics, we would have to take into consideration that there are only 50 schools, not 1000 – i.e. analysing the data at the level of the pupil would give incorrect estimates of the relationship between school characteristics and academic improvement. This is an example of where hierarchical/multi-level statistical models would be required to produce accurate analyses.

**Missing data** – There are two main types of missing data: **a)** a missing unit – this occurs when all data for a particular sampling unit is missing (e.g. when a person does not respond to a survey) and **b)** a missing item – this occurs when a particular item of data is missing (e.g. an answer to a question on a survey). The problem with missing data is that it is often not random (people have reasons for not responding, for example) and can therefore bias your analyses.

Some ways to combat this problem include:

- Keeping a survey very short and piloting questions;
- Using incentives to encourage participation;
- Building a relationship with respondents to encourage participation; and
- Asking non-respondents why they did not respond (to determine if your analyses are likely to be biased).

It is often necessary to fill-in ('impute') missing items or reweight data before analysis. Imputation enables the full returned sample size to be used in models and reweighting causes the data to more closely match the population from which it came on key characteristics. *[NB: Imputation and reweighting methods can be complex and time consuming.]*

<sup>1</sup> Adapted from: Bartlett et al., "Organizational Research: Determining Appropriate Sample Size in Survey Research", Information Technology, Learning, and Performance Journal, Vol. 19, No. 1, Spring 2001

<sup>2</sup> If a categorical explanatory variable with 5 categories was included, this would count as 4 of the maximum model variables (4 dummy variables).