# What you need to know about sampling

## General points

- The purpose of sampling is to make inferences about the *population* it is drawn from.
- Sampling involves selecting a number of *units of interest* from the population and collecting data for these units.
- Be clear on what is the unit of interest. Think of each unit as a row in a table of data. If you are mist-netting birds, the unit of interest may be:
  - each individual bird, if you are interested in breeding condition or morphometrics, or
  - the netting session, if you are interested in bird density or species diversity.

  (Biologists are often sloppy with terms, describing a netting session as a sample when it's actually a sampling unit, and the sample consists of several netting sessions.)
- Be clear too on what is the population of interest, and be sure it corresponds to the sampling scheme. For example, if you aren't going to sample infants, exclude them from the population. If you can't survey limestone outcrops, define the study area to exclude them.
- The sample data will be used to calculate the likelihood, ie, the probability of the data given a specific model and parameter values. The basic assumption is that the data are random draws from some specific distribution, often binomial, Poisson or normal.
- *Haphazard* sampling is not the same as random sampling; you should have a proper method for selecting units, using random numbers, rolling dice, drawing names from a hat, etc.
- In a simple random sample, all units in the population have equal probability of being included in the sample. With complex sampling methods, probabilities are not equal, but they are known and are included in the analysis. We can still use the laws of probability.
- Sampling is independent if drawing one unit has no affect on the probability of drawing other units (see Systematic Sampling below). Lack of independence often means that the units in the sample are more similar to each other than units in the population, resulting in a narrower posterior distribution (or narrower confidence interval in frequentist analysis).
- Sampling is usually done *without* replacement, but sampling *with* replacement is also possible (and is often useful with spatial sampling).
- Sampling is different if you are interested in modelling a relationship between a response variable and explanatory variables. Here you need to include units representing the full range of the explanatory variables. For example:
  - Effect of fertiliser on crop yields: take separate (random) samples from plots with fertiliser and plots without.
  - Effect of altitude on bird density: take a series of (random) samples at different altitudes.

  You should still randomly select units within each group – unless you are sure that no other factors have an effect.
- If your sample is not selected randomly, you cannot make inferences about the whole population. The study may still provide useful *insights* rather than conclusions, and the product can be presented as a *case study*.

## Possible sample designs

**Simple random sampling** – use random numbers or a similar randomizing procedure (haphazard is not random!)

**Cluster sampling** – the population is divided into non-overlapping clusters (primary sampling units) of units of interest (now 'secondary sampling units'). The sample consists of randomly-selected primary units and *all* the secondary units inside them.

**Multistage sampling** – similar to cluster sampling, but a *subsample* of secondary units in each primary unit is taken.

**Systematic sampling** – a first sample unit is selected at random, then other units are selected with a fixed space between them. For example, selecting the first quadrat at random, then placing quadrats at 100m intervals north, south, east and west; or rolling a die to select the first person to interview as they leave a site, then interviewing every $10^{th}$ person to exit.

Selection of units is *not* independent, and units in the sample are generally less similar than in the population, resulting in a broader posterior probability distribution. This can be dealt with by including autocorrelation in the model, but it can get complicated.

**Stratified sampling** – the population is divided into non-overlapping strata, and a random sample is drawn from each stratum. Stratified sampling reduces the variance of the overall estimate, provided the units within each stratum are similar. With post-stratification, a simple random sample of units is taken and these are allocated to strata on the basis of data gathered during the survey; only possible if the proportions in each of the strata are known a priori.

**Double or two-phase sampling** – a larger (random) sample is selected first and an auxiliary variable observed for this, than a subsample is drawn and the main variable of interest observed for this.

- Auxiliary variable could be cheap and quick surrogate of variable of interest, eg. habitat suitability assessed by remote sensing as auxiliary variable, and occupancy determined by ground surveys in the subsample.
- Auxiliary variable could be the basis for stratification, with the subsample being selected from strata within the large sample.

**Line-intercept sampling** – used for large, stationary objects, eg. groups of chimp nests or wolf tracks in snow, where probability of inclusion depends on size. Size is not known a priori.

**Adaptive sampling** – useful for rare species of plants (possibly animals) which occur in clusters larger than the plots sampled. During the survey, when a plant is observed, neighbouring plots are added to the sample; if one of these also has the plant, its neighbours are added, and so on.

- The initial sample may be a simple random sample, stratified sample, systematic or strip sample.
- Treating the data as a simple random sample will give wrong answers; methods which allow for adaptive sampling must be used.
- A practical problem is that the number of plots which will have to be surveyed is unknown in advance, so planning and budgeting are difficult.

## Further reading

If you want to use one of the more complicated sampling designs to estimate a population parameter, a good resource is:
Stephen Thompson, *Sampling*, $3^{rd}$ edition 2012, Wiley.

## Some sampling situations

1. Morphometric data for bats from a harp-trapping survey in 2 different habitats: what are the samples? are the measurements independent? how to estimate variances (and CIs)?
2. Riverside survey of proboscis monkeys or crocodiles, can't do whole length of river, so need to divide into sections using natural landmarks and take a sample: how to do this? Coda: if surveying 1 side only, is it okay to toss a coin to decide which side to survey?
3. Want to know how many active hunters live within 4 km of the boundary of a protected area, also how many of these hunters use traps (and there are too many villages in the area to be able to visit all of them): how to design the survey?
4. You want to estimate the number of strangler fig plants in a forest; figs are fairly rare, and tend to occur in groups. How would you plan the survey?