Statistical literacy guide
A basic outline of samples and sampling

Last updated: October 2006
Author: Ross Young

There are two types of samples – scientific and non-scientific samples. The best statistical samples are those that are conducted scientifically.

Non-scientific samples are those where the cases to be sampled are selected for their typicality or availability, and it is not clear from the results of surveys using non-scientific sampling how they can be generalised to a wider population. For example, we wish to survey blood donors about the reasons why they choose to donate blood. If we simply interviewed all those blood donors giving blood when we visited the blood centre, then this would be known as a non-scientific sample.

In **scientific sampling**, the probability of any person being selected as part of the sample is known because, in order to select a scientific sample, a 'sampling frame' is needed. A **sampling frame** is a list of all individuals in the population to be surveyed. A common sampling frame used in official statistics is a list of addresses drawn, for example, from the electoral register or the Postal Address File. Samples drawn on the basis of a sampling frame are known as **probability samples** since the probability of inclusion for any individual or household is known, and therefore it is possible to generalise or make inferences about the wider population from the results of the survey.

There are several ways in which probability samples may be selected. The best or 'purest' form of probability sampling is the **simple random sample**, in which every person in the population has a known and equal chance of selection. Simple random samples are often generated by computer from address files or telephone directories, and it is commonly used in market research or telephone sales campaigns. However, cost and convenience considerations mean that researchers often use a modified form of random sampling, where individuals in the population have a different (albeit known) probability of inclusion. For example, we may choose to sample every 5th person in a list of names (**systematic sampling**), or sample different groups in the proportion they exist in the population as a whole (**stratified sampling**), or limit our sampling frame geographically by selecting particular sampling points such as parliamentary constituencies or postcode sectors (**cluster sampling**). Alternatively, we may use a combination of all these sampling methods in a single survey.

It is good practice, when presenting the findings of a survey, for details of the sampling method to be provided, usually as an appendix to the survey report.

However, the results of sample surveys may not reflect the situation among the whole population. For example, if we undertake a survey of household income using simple random sampling and visit 100 houses, we would have 100 results and we could calculate the average (mean) household income among these households. Because we have only taken a sample, we suspect that the mean household income from our survey may be different from the mean income of the whole population because we have not interviewed the whole population and there was a chance factor involved in who we decided to sample. This chance factor is referred to 'sampling error' and, when considering the results of sample surveys, it is important to measure the extent of sampling error.

Sampling error will depend on two factors – the size of the sample and the extent of variation in the indicator we are measuring among the population we are sampling.

The larger the sample size, the less chance there is of selecting cases with extreme values of the indicator we are measuring (e.g. household income), and the better chance we have of approximating the true value among the whole population. In other words, surveys with larger sample sizes (e.g. 1,000 or 15,000 people) are more likely to tell us something interesting about the whole population and, conversely, surveys with small samples (e.g. 10 or 50 people) are less likely to produce results which can be extrapolated to the population as a whole. The results of large sample surveys are less likely to produce high levels of sampling error.

Again, it is good practice for details of the size of the sample to be provided when presenting the findings of surveys, alongside an estimate of the degree of sampling error to allow a calculation of how the survey's results may deviate from the population as a whole. For example, opinion pollsters often present their results as estimated vote shares for parties A, B, and C as being plus (+) or minus (-) *x* per cent. +/-x% is an estimate derived from a calculation of sampling error. The lower the sampling error the more certainty we should have that the results they publish reflect the true voting intention of the population as a whole.

Other statistical literacy guides in this series:

- What is a billion? and other units
- How to understand and calculate percentages
- Index numbers
- Rounding and significant places
- Measures of average and spread
- How to read charts
- How to spot spin and inappropriate use of statistics
- A basic outline of samples and sampling
- Confidence intervals and statistical significance
- A basic outline of regression analysis
- Uncertainty and risk
- How to adjust for inflation

This sheet is of a series of statistical literacy guides put together by the Social & General Statistics section of the Library. The rest of the series are available via the Library intranet pages.