Sampling (4 pages; 7/8/17)

In Statistics, the group of people or items to be studied is known as the **population**, and we may be interested in making inferences about **parameters** of the population.

Example A: The mean age of pupils in Year 7 at a particular school.

Example B: The proportion of people in a particular parliamentary constituency who intend to vote Labour.

Example C: The proportion of tyres produced by a factory that are faulty.

Ideally, a list would be available of all people or items in the population, and a complete set of data obtained, to enable the required parameter to be calculated. A population **census** is taken every 10 years for this purpose.

In most cases, however, this is not possible or feasible. Reasons for this include the following:

(a) No list exists (e.g. the population of men with shoe size 11 or more).

(b) It would be too costly or time-consuming to create the list.

(c) In the case of Example C, the tyres may be spoilt in the process of testing, and so we could not test every tyre.

Instead, a sample is taken, and a **statistic** obtained, which provides an estimate for the population parameter that we are interested in. An example of a statistic would be the sample mean, with the corresponding parameter being the population mean.

It is important that the sample is representative of the population; i.e. that it is not **biased**.

The easiest way of ensuring that the sample is not biased, is to make sure that it is **random**. A sample is random if every possible choice of sample is equally likely. It is not always possible to achieve complete randomness; in which case, we try to choose our sample in a way that minimises bias.

The different types of sampling method are described below, and classified according to whether or not it is possible to list every member of the population (such a list is called a **sampling frame**).

(A) Where it is possible to list every member of the population

Simple Random Sampling

By this is meant that every possible sample (of the chosen size) is equally likely to be chosen. It follows from this that every member of the population is equally likely to be chosen.

Systematic Sampling

Here we may have all the members of the population listed in a spreadsheet, for example. We select every kth member in the spreadsheet, having chosen the 1st one at random from the 1st k members. We should ensure that there is no hidden bias, due to the way that the data have been ordered. For example, if it consists of classes in a school, with one class following another, it could be the case that the more able pupils have been listed first in each class.

Stratified Sampling

It may be known that the population consists of a number of **strata** that have different characteristics (relevant to the investigation). Thus for Example B, the voting habits of younger and older people may be expected to differ. We then take simple random samples within these strata, where the sample sizes are in proportion to the size of the strata in the overall population.

(B) Where it is not possible to list every member of the population

Opportunity Sampling

This is where we take advantage of an existing collection of members, rather than seeking out a sample from the wider population. For example, a sample of criminals may be taken from a prison. This type of sampling can be prone to bias, but may be useful as a pilot exercise.

Quota Sampling

This is often employed by market researchers stopping people in the streets.

As with stratified sampling, the desired proportions of different categories (e.g. defined by a combination of age, gender, income etc.) are first of all decided on. Then the researcher is often free to choose people from these categories, until the required quota are achieved. Thus Quota Sampling usually involves an element of Opportunity Sampling.

This approach is highly likely to involve bias. For example, researchers may favour people who appear to be amenable to answering questions.

Cluster Sampling

This is often employed when carrying out nature studies. To investigate, say, the population of ants in the UK, a moderate number of sites (the 'clusters') may be chosen around the UK, and then samples taken from these sites. This approach assumes that the sites all have similar characteristics. It would not be appropriate for assessing voting intentions amongst the human population!

Further issues that should be considered are as follows:

(i) Have the right questions been asked? (For example, "Which party do you trust the most?" may not produce the same response as "Which party will you vote for?")

(ii) Is there likely to be a tendency for respondents to lie or modify their answers, if questions are worded in a particular way? (e.g. "Do you drink excessively?")

(iii) Is the sample large enough to be statistically significant? This issue is addressed when we carry out hypothesis tests, later on.

(iv) Is the sample **self-selecting**? For example, shoppers in an expensive store are unlikely to be representative of the whole population.