Random Samples and Sampling Techniques

Random Samples

When drawing a sample in order to estimate population parameters, such as the mean and standard deviation, we require that every member of the population has an equal chance of being selected.

But any process of systematic selection has the potential to reduce this element of chance.

As an example of random sampling consider drawing lottery balls from a lottery machine. Every ball has an equal chance of being drawn, a point which the organisers of the competition are eager to point out.

But in this case there may be no more than 100 balls, and achieving true randomness of selection is more or less attainable. Yet this is not as simple as one might imagine – the weights of the balls must match exactly, as must the dimensions, and this is a problem of engineering. Putting names on a piece of paper into a hat is not truly random as the size and weights of the pieces of paper differ.

In other words it is extremely difficult to achieve a truly random sample and when the population is large, random sampling becomes practically impossible. In these situations some form of systematic sampling must take place. For example, taking every 10^{th} name from a list such as a telephone book.

The introduction of systematic sampling always raises questions as to the reliability of the sample as a basis for the determination of a population parameter, such as the mean or variance. However, some methods are obviously less reliable than others. One is expected to be able to detect obvious cases where the sample is not sufficiently random; however, a detailed knowledge of sampling and survey methods is not required.

Example

The manager of a shoe store wishes to determine whether he is effective at his job. He asks his 10 most trusted colleagues, all of whom he has recently promoted, whether they are satisfied with his leadership style. Comment on the manner in which the sample was chosen and suggest a more suitable way of obtaining a sample.

Solution

Clearly, the manager's friends are likely to respond positively. He should place all his subordinates' names into a hat and ask those that he draws out at random.

As the example illustrates the problem here is to avoid glaring cases of bias. There are methods of sampling that distort the outcome. Systematic sampling need not be a source of bias in this sense. These questions require judgement.

A formal definition of a random sample of size n is as follows:

> A random sample of size n taken from a parent population X is a set of observations $X_1, X_2, \ldots X_n$ such that each observation is determined by chance and hence has the same probability distribution as X (the parent population) and each pair of observations is independent.

Independence means that knowledge of one observation does not affect the probability of another observation – it is another way of saying that each observation is determined by chance.

Some terms

A *sampling frame* is a representation of *all* the items available to be sampled – for example, a list of all the names of people in a particular population. However, often no sampling frame exists. For example, it is not possible to name and list all the fish in the Atlantic Ocean, so no sampling frame exists to begin with when sampling such fish.

The proportion of the available items that is actually sampled is called the sampling fraction.

$$\text{Sampling fraction} = \frac{\text{Sample size}}{\text{Population size}}$$

The aim of sampling is to estimate a population parameter, such as the mean or standard deviation, from information drawn from the sample. A value derived from a sample is called a sample statistic, or just *statistic*. A statistic that is used to estimate a population parameter is called an *estimate*. The difference between an estimate and the population parameter is called the *sampling error*. The aim of effective sampling is to reduce the sampling error as far as possible- by choosing an appropriate sample size, and by eliminating bias.

Sources of bias

There are various ways in which a sample may become unrepresentative of the population, and they cannot all be listed. However, some familiar situations are

(1) When the question asked is not actually appropriate to the matter under investigation. For example, if you want to determine the most likely person to be elected at an election, then you should not ask, "Who will make the best MP?" but rather "Who will you vote for?"

(2) If the question is framed in a certain way it may invite a certain kind of response. For example, the question, "Do you consider yourself to be a pleasant person?" is likely to draw the answer, "Yes!"

(3) Sometimes people do not like being questioned and they deliberately give the wrong answers. Bias can also be introduced by the identity of the person asking the questions – for example, the managing director should not collect the answers to the question, "Is the management here effective?"

(4) The sample must be representative of the population. Questionning MPs on a woman's issue will create bias, since there are an unduly high proportion of male MPs. MPs are not representative of the people as a whole, even though they represent them in Parliament!

Sample size

The sample size must be large enough to create confidence in the estimate drawn from it. However, this is a subject for further mathematical treatment, and the size of the sampling fraction required depends on the degree of confidence required.

Census

A census is when every member of the population is observed or measured. A census is used if the population is small and extreme accuracy is required. However, in most cases where information about a population is required, a sample is taken instead.

Sampling

The advantages of sampling are that sampling is (a) cheaper and (b) still representative of the whole population

The disadvantage of sampling is that it introduces uncertainty due to natural variation and bias

There are several kinds of sample

Simple Random Sample

A sample of size n is called a simple random sample if every other possible sample of size n has an equal chance of being selected.

Example

In lottery ticket sampling you put the ticket into a container and withdraw the required sample or equivalent.

In order to implement simple random sampling you must have a sampling frame. When such a frame exists the simple random sample is often found by the use of random numbers.

Systematic Sampling

Systematic sampling occurs when a sample is obtained by choosing at regular intervals. First the population is placed into an ordered list. Then the sample is chosen by choosing items from the population at regular intervals. Systematic sampling is used when a population is too large for random sampling.

The advantages of this type of sampling is that it is simple to use and suitable for large numbers

The disadvantage is that it is only random if the ordered list is truly random

Stratified Sampling

In this case the population is divided into groups or categories that are mutually exclusive. No individual can be in two groups. The groups are called strata.

This is used for large populations and where there are natural divisions. Each stratum is sampled randomly. If clear strata are present then it can be more accurate than simple random sampling.

When the number in the sample for each strata is proportional to the size of the corresponding sub-population, then the sample is called a proportional stratified sample.

If the strata are not clear then they may overlap.

Cluster Sampling

This is a form of stratified sampling. The population is divided into sub-groups, but instead of taking, say, one member from each sub-group only one, or a few, of the sub-groups are sampled. In other words just one, or a few, "clusters" are chosen. This is done for practical reasons. For example, if you want to know something about Canadian polar bears it makes sense to chose one community of polar bears from all the separate sub-populations, but you have to assume that that the community of bears is representative of all the bears.

Sampling with and without replacement

There is a distinction between sampling with and without replacement.

Sampling with replacement occurs when once a item has been chosen it is "returned" to the population, so that it may be chosen more than once. For example, choosing numbered balls from a bag. A ball is drawn, if the ball is then replaced into the bag, then you have sampling with replacement; if the ball is not replaced into the bag, then the balls are being sampled without replacement.

Unrestricted random sampling - is the term for when simple random sampling is used with replacement.

When simple random sampling is used it is understood to be without replacement.

Quota sampling

Quota sampling is a commonly used non-random method. For example, an interviewer interviews a quota of people who fit a certain criterion.

Small versus Large Samples

Large versus small samples

Sample bias is the tendency for the sample statistic to differ from the population parameter that it estimates. In general, as the sample size increases the sample bias decreases.

Therefore, an argument in favour of the use of large samples is that the larger the sample the smaller the sample bias.

However, there are arguments in favour of small samples.

Firstly, if there is a significant participant variable that is affecting the results, a large sample size will decrease its impact. A small sample, will show up the effect of a participant variable, and result in an improved design of the experiment. Therefore, small samples should be used as these improve experimental design. Thus, large samples disguise weaknesses in the *design* of an experiment.

It is often stated that an optimum size for a sample is 25 to 30 subjects. If this does not result in a significant difference, then either the hypothesis is false or the design is faulty and requires modification.

Small samples are also cheaper to fund, and take less time. It is also sometimes the case that it is difficult to find sufficient numbers of respondants. The data from smaller samples may be analysed more quickly.