3. Sampling accuracy

In this section readers will:

(a) Be presented with a mathematical definition of sampling accuracy.
(b) Examine accuracy in original and transformed ("normalized") populations.
(c) Make observations on the growth pattern of accuracy with varying sample size.
(d) Verify the inverse relationship between variability and accuracy.
(e) Determine population-specific accuracy boundaries.

3.1 Definition

Let us assume a finite population of N elements $y_1, y_2, ..., y_N$ with a minimum value y_{min}, a maximum value $y_{max} \neq y_{min}$ and mean μ. We also consider a sample of n elements with sample mean m. A relative index of proximity of the sample mean m to the population mean μ (briefly referred to as *accuracy* A in the paper), is defined by the following formula:

$$A = 1 - \frac{|m - \mu|}{R} \qquad (3.1)$$

where R denotes the population range $y_{max} - y_{min}$.

The above definition is in accord with the classical approach of determining a minimum allowable difference between a true population parameter and its estimator (see Cochran, 1977; Thompson, 1992; for discussion), except for the introduction of the population range into the item describing absolute error.

3.2 Normalizing the target population

Generally, the range R of a population is not known but this will not affect the study of accuracy if we consider the original population mapped onto the standard interval [0,1] through the transformation formula:

$$u_i = \frac{y_i - y_{min}}{R} \tag{3.2}$$

It is evident that by its definition through (3.2) the resulting *normalized* population $u_1, u_2, ..., u_N$ will have elements between and including 0 and 1.

3.3 Sampling accuracy in normalized populations

It will be shown that the accuracy of any sample from the original population as defined in (3.1) is equal to the accuracy of its mapped equivalent taken from the normalized population.

Proof:

In the normalized population $u_1, u_2, ..., u_N$ all elements will be between and including 0 and 1 and the mean will be:

$$\mu_u = \frac{\mu - y_{min}}{R} \tag{3.3}$$

Any sample of n elements $y_{k_1}, y_{k_2}, ..., y_{k_n}$ with mean m is mapped onto a normalized sub-set $u_{k_1}, u_{k_2}, ..., u_{k_n}$ with mean:

$$m_u = \frac{m - y_{min}}{R} \tag{3.4}$$

Since all normalized elements are between 0 and 1 the range of a normalized population is 1. By using expression (3.1) to formulate the accuracy A_u of sample $u_{k_1}, u_{k_2}, ..., u_{k_n}$ and by taking into account (3.3) and (3.4), we find:

$$A_u = 1 - \frac{|m_u - \mu_u|}{1} = 1 - \frac{|m - y_{min} - \mu + y_{min}|}{R} = 1 - \frac{|m - \mu|}{R} = A$$

hence the proof of the proposition.

The fact that sampling accuracy remains unchanged when a population is normalized by means of transformation formula (3.2) permits us to study the accuracy with regards to normalized populations only.

From this point on it is assumed that all population parameters and sampling approaches are referring to normalized populations. In this manner the accuracy A will be simply defined as:

$$A = 1 - |m - \mu| \tag{3.5}$$

By its definition (3.1) it also follows that accuracy A has a lower value of zero and a maximum of 1.

Numerical example

Consider the population of 11 elements 0, 1, 2, …, 10 with mean 5. By selecting the sample (2, 6), the population mean is estimated by the sample mean 4. By applying formula (3.1) we find that the resulting accuracy is:

$$A = 1 - \frac{|m - \mu|}{R} = 1 - \frac{|4 - 5|}{10 - 0} = 0.90$$

Next we normalize the population by applying formula (3.2). It easy to verify that the normalized elements are: 0, 0.1, 0.2, …, 1 and the population mean is 0.5.

The previous sample (2, 6) is mapped on the normalized sample (0.2, 0.6) with mean 0.4. By applying the same formula (3.1) for sampling accuracy we find:

$$A = 1 - \frac{|m - \mu|}{R} = 1 - \frac{|0.4 - 0.5|}{1 - 0} = 0.90$$

Which verifies numerically that the accuracy of any sample from the original population as defined in (3.1) is equal to the accuracy of its mapped equivalent taken from the normalized population.

3.4 Accuracy plots

Let us assume a normalized and finite population of size N and a series of successive random samples with sizes 1, 2, 3, …, N. In each sample the population mean will be approximated by a sample mean with accuracy A defined as in (3.5). By plotting A against sample size the resulting graph will show a fluctuating accuracy curve of hyperbolic shape (first plot of Figure 3.1). In this example sample size is expressed by the ratio n/N so that both the horizontal and vertical axis are scaled from 0 to 1. Notice that the curve does not start from 0 but from $1/N$ which is the smallest sample proportion.

Accuracy plots are easier to view and analyze if sample proportion is expressed by the ratio $\log n / \log N$ rather than n/N. In this case the

curve takes an exponential shape starting from 0 (equivalent to the ratio $\log 1 / \log N$). This is shown in the second plot of Figure 3.1.

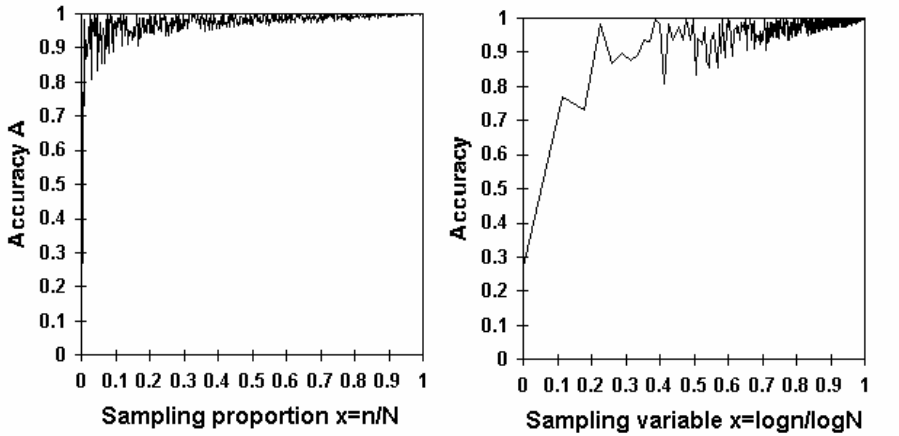

Figure 3.1. Accuracy plots from a normalized and finite population of size N. Accuracy values correspond to successive random samples with sizes 1, 2, 3, …, N. Notice the different shapes of the two plots depending on the expression used for sample proportion.

A striking feature of accuracy growth is its sharp increase near the small samples and its much slower and stabilized shape beyond a certain "critical" sample size. It will later be shown that for finite populations this critical size corresponds to $\sqrt{N}$.

3.5 Accuracy and variability

It is easy to prove that in finite normalized populations the sample size achieving a *minimum* allowable accuracy A_{min} with a given probability is given by:

$$n = \frac{1}{\frac{(1 - A_{min})^2}{z^2\sigma^2} + \frac{1}{N}} \tag{3.6}$$

Expression (3.6) is based on the classical approach for determining safe sample size (see Thompson, 1992; p. 32). In this approach a pre-set maximum allowable difference d between the estimated mean and its true value is established, as well as a small probability α that the error will not exceed that difference. Sample size is then determined as:

$$n = \frac{1}{\frac{d^2}{z^2\sigma^2} + \frac{1}{N}} \tag{3.7}$$

where z is the upper $\alpha/2$ point of the standard normal distribution and σ^2 the population variance. Expression (3.6) derives from (3.7) by taking into account that in normalized populations the *maximum* allowable error d will be between 0 and 1 and it can therefore represent the difference $1 - A_{min}$.

3.6 Population-specific accuracy boundaries

Expression (3.6) can be used to formulate a population-specific lower boundary function for sampling accuracy at varying sample size. Solving for A_{min} we obtain:

$$A_{min}(n) = 1 - z\frac{\sigma}{\sqrt{n}}\sqrt{1 - \frac{n}{N}} \tag{3.8}$$

The above expression indicates that with varying sample size the resulting accuracy will be expected to be found above the curve formed by $A_{min}(n)$ at a probability level determined by z.

Figure 3.2 illustrates two examples of population-specific accuracy boundaries. The following parameters were used in evaluating expression (3.8):

N=1000.

n=1, 2, …, N.

z=1.96.

In both examples σ is the standard deviation of the normalized population.

Accuracy values and boundary functions are plotted against the sampling variable $\log n / \log N$. With few exceptions all accuracy values, whether resulting from small or large samples, are above the lower boundary defined by (3.8).

A weak point in the above process is that such accuracy boundaries can seldom be used as *a priori* guidance for achieving sampling accuracy at a desired level. Expression (3.8) constitutes only a *population-specific* accuracy boundary since the variance of the target population is assumed to be known. Generally this is not the case at the initial stage of a sampling programme, thus defeating the purpose of setting-up accuracy boundaries on an *a priori* basis. However, *global boundaries* can instead be constructed through the use of two specific populations for which σ^2 can be computed in advance and this will be the main subject of the next section.

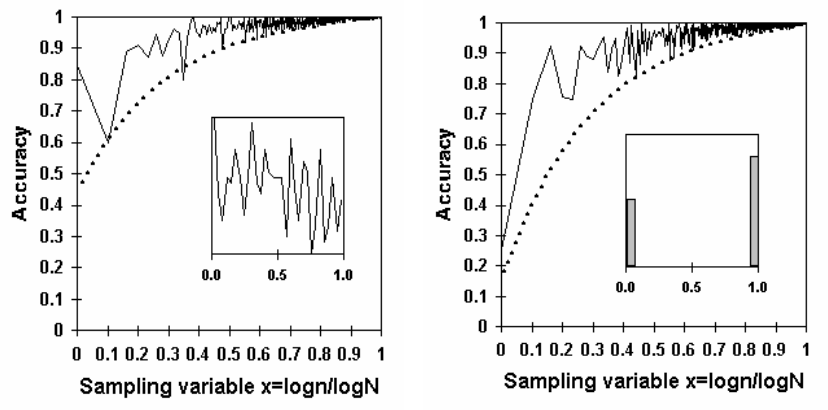

Figure 3.2. Fluctuating sampling accuracy and population-specific accuracy boundaries (dotted line) for two finite and normalized populations of size N=1000. The first population is flat, whereas the second is concave and binary.

SUMMARY

At this point readers should be familiar with the mathematical definition of sampling accuracy and its relation to sample size. The following points have been emphasized:

(a) In this handbook sampling accuracy is defined as a relative index of proximity between the actual population mean and an estimate resulting from a sampling operation.
(b) Accuracy remains unchanged if the target population is normalized.
(c) Accuracy values follow a standard growth pattern with sample size.
(d) It is possible to formulate accuracy boundaries when the population variance is known or can be guessed at.
(e) Property (d) is not very useful because it requires *a priori* knowledge about the target population, which normally cannot be obtained.
(f) There is a clear need for *global (i.e.general)* accuracy boundaries that are independent of the population variance and depend only on the population size.