# 5. Accuracy boundaries in small populations

In this section readers will be presented with an approach that concerns accuracy boundaries in small populations. Major topics include:

(a) Shortcomings of the probabilistic approaches described in Section 4 when target populations are of much smaller size.
(b) Advantages of algebraic-based over probabilistic-based accuracy boundaries in cases of very small populations.
(c) Practical criteria for determining the use of probabilistic and algebraic accuracy boundaries.

## 5.1 Example of probabilistic boundaries in small populations

Figure 5.1 represents the application of the probabilistic approach presented in Section 4 in the case of two small populations each with size $N$=100. Lower accuracy boundaries were constructed using formulae (4.7) and (4.9) at a probability level of 95 percent ($z$=1.96).

The plots illustrate a significant gap between fluctuating sampling accuracy and its predicted lower limits. This “safety” space becomes more exaggerated in very small populations, such as the days in a month, where $N$ can be as small as 28. It would thus seem that the probabilistic approach is “too pessimistic” in the case of small or very small populations and that safe accuracy limits can be obtained with much smaller samples than those indicated by the boundaries. This defect can partially be remedied by changing the value of $z$ according to the population size but this technique does not alter significantly the picture and adds considerable complexity to the construction of accuracy boundaries.

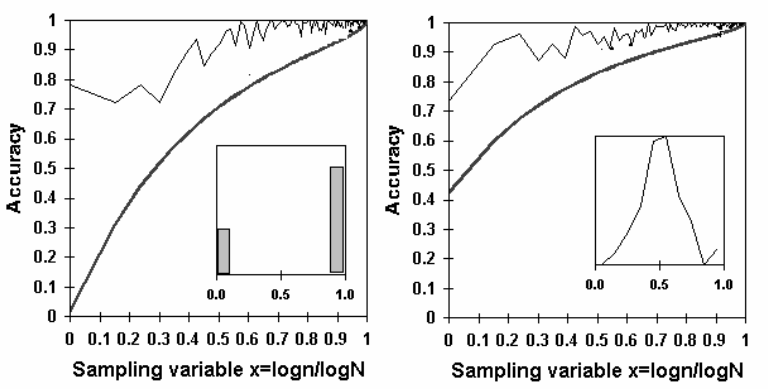


*Figure 5.1. Accuracy plots and probabilistic accuracy boundaries for small concave (0-1) and convex populations. Population size is $N$=100. Notice the excessive safety space between global boundary curves and fluctuating sampling accuracy.*

The reason for this shortcoming is that fundamental formula (3.6) (which constitutes the basis for formulating population-specific and global boundaries), assumes that sample means follow the normal distribution, an assumption that no longer holds when the population is too small.

## 5.2 Algebraic accuracy boundaries

Stamatopoulos (1999) has worked out an algebraic approach that seems to answer most of the questions raised in the previous section (see also References). Rather than applying the probabilistic formula (3.6), accuracy boundaries for small populations make use of an exponential function of the type:

$$G(x) = a_1 + a_2 N^{-kx} \qquad (5.1)$$

where the independent variable $x$ is the ratio $\log n / \log N$ and $N$, $n$ denote population and sample size respectively.

The three parameters $a_1, a_2, k$ are formulated on the basis of four basic variables denoted $\overline{W}$, $a$, $g$, $S$ which are computed as follows:

(1) Computation of $\overline{W}$ for concave populations.

$$\overline{W} = 1 - \log(1 + 0.5\, e^{\frac{1}{N}}) \tag{5.2}$$

(2) Computation of $\overline{W}$ for flat or convex populations.

$$\overline{W} = 0.75(1 - \frac{1}{N}) \tag{5.3}$$

(3) Computation of $a$.

$$a = \frac{2\overline{W}N^2}{(N-1)^2} - \frac{N+1}{N-1} \tag{5.4}$$

(4) Computation of $g$.

$$g = a + \frac{1-a}{N} \tag{5.5}$$

(5) Computation of $S$.

$$S = (1-a)\left(\frac{1}{\log N} - \frac{1}{N \log N} - \frac{1}{N}\right) \tag{5.6}$$

Once $\overline{W}$, $a$, $g$, $S$ have been evaluated, the three parameters $a_1, a_2, k$ of expression (5.1) are computed as follows:

$$a_1 = g - \frac{(1-S-g)^2}{2S+g-1} \tag{5.7}$$

$$a_2 = \frac{(1-S-g)^2}{2S+g-1} \tag{5.8}$$

$$k = -\frac{2}{\log N} \log \frac{S}{1-S-g} \tag{5.9}$$

To be noted that $a_1, a_2, k$ are only a function of the population size $N$ since the values of the four basic variables $\overline{W}$, $a$, $g$, $S$ depend only on $N$.

Figure 5.2 illustrates the application of the above approach on two small populations with size $N$=30. The first population is concave with 0 -1 elements while the second is convex. The dotted line represents the probabilistic curve drawn according to the concepts described in Section 4. Comparison between the two boundaries reveals that the probabilistic approach tends to become unduly “pessimistic” when the target populations are very small.

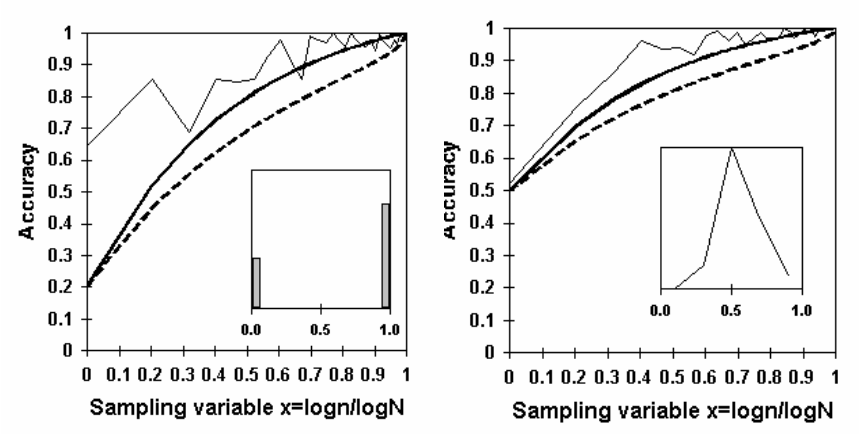


*Figure 5.2. Accuracy plots and algebraic and probabilistic boundaries (dotted line) for two small populations with size $N$=30.*

## 5.3 Properties of algebraic boundaries

The properties of algebraic boundaries as defined in (5.1) are similar to those defined in the probabilistic approach in Section 4.

(a) For $x$=0 the intercept of function (5.1) and the vertical axis $A$ is a value between 0 and 1.

In fact, $G(0) = a_1 + a_2 = g$ and by considering expressions (5.2), (5.3) and (5.4) it is easy to verify that $g$ lies between 0 and 1. Its exact position depends on whether the target population is assumed to be concave or convex.

(b) For $x$=1 function $G(x)$ also becomes 1.

To prove that $G(1) = a_1 + a_2 N^{-k} = 1$ will involve some calculations, starting with the observation that any variable $C$ can be written as:

$$C = N^{\frac{\log C}{\log N}}$$

Based on the observation above and the fact that (5.9):

$$k = -\frac{2}{\log N} \log \frac{S}{1 - S - g}$$

we can write:

$$G(1) = a_1 + a_2 N^{-k} =$$

$$= g - \frac{(1 - S - g)^2}{2S + g - 1} + \frac{(1 - S - g)^2}{2S + g - 1} \frac{S^2}{(1 - S - g)^2} = 1$$

(c) As with probabilistic boundaries, also in algebraic boundaries there exists a breakpoint at sample size $n = \sqrt{N}$, at which accuracy becomes stable and starts a slow convergence towards 1.

By considering the function:

$$B(x) = g + 1 - a_1 - a_2 N^{-k(1-x)}$$

and recalling the earlier property $a_1 + a_2 N^{-k} = 1$, we first notice that:

$$B(0) = g + 1 - a_1 - a_2 N^{-k} = g + 1 - 1 = g$$

That is at x=0 $B(x)$ has the same intercept g.

On the other hand at x=1 $B(x)$ also becomes 1 because of the relationship:

$$B(1) = g + 1 - a_1 - a_2 N^{0} = g + 1 - g = 1$$

In other words functions $G(x)$ and $B(x)$ are both exponential, have the same intercept $g$ and meet at the same final point 1.

However, their growth patterns are contrasting. Function $G(x)$ shows a rapid growth up to a certain value of $x$ and from then on it grows steadily until it becomes 1. Function $B(x)$ shows a slow and steady growth for small values of $x$ and beyond a certain point it starts a rapid growth until it also becomes 1.

Evidently the critical value of $x$ is at a point where the difference $G(x)$-$B(x)$ becomes maximum since it is from that point on that the growth of $G(x)$ becomes slower and steadier and that of $B(x)$ faster.

In terms of differential calculus we are seeking a critical point $x$ at which the difference $G(x)$-$B(x)$ becomes maximum, which occurs when:

$$\frac{d}{dx}\left[G(x) - B(x)\right] = 0 \qquad \text{or} \qquad \frac{dG}{dx} = \frac{dB}{dx} \qquad \text{or:}$$

$$-a_2 N^{-kx} \log N = -a_2 N^{-k(1-x)} \log N$$

It is easy to verify that x=0.5 is a solution to the above equation, which leads to the conclusion that algebraic accuracy boundaries also have a breakpoint at sample size $n = \sqrt{N}$ .

## 5.4 Criteria for applying algebraic boundaries

Figure 5.3 illustrates the application of both algebraic and probabilistic accuracy boundaries in two concave populations with 0-1 elements and sizes N=30 and N=900.

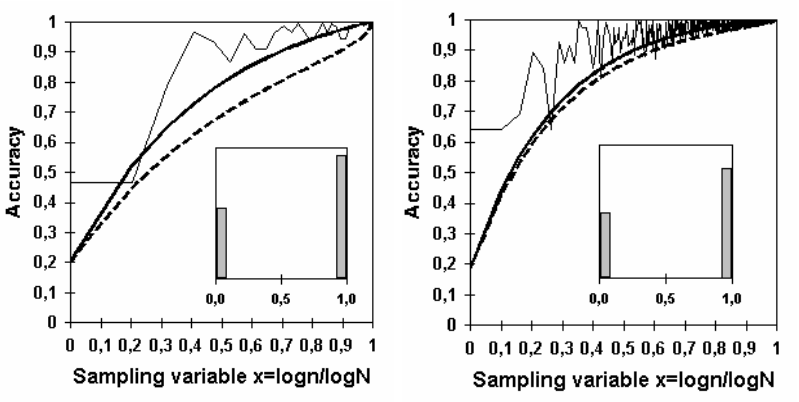


*Figure 5.3. Algebraic and probabilistic boundaries (dotted line) in very small (N=30) and small/medium size (N=900) populations.*

In the first case the probabilistic boundary (dotted line) is found much below the accuracy fluctuation and the algebraic boundary seems to provide more realistic lower limits. In the second case the two lines almost coincide.

As $N$ increases the situation is reversed. Algebraic boundaries become excessively pessimistic and probabilistic lower limits are more realistic. It would thus seem that an empirical criterion for choosing between the two approaches is the following:

*Algebraic boundaries are more effective in very small populations with size not exceeding 900. Beyond that size the probabilistic boundaries should apply.*

## SUMMARY

In this section readers were presented with an approach for setting-up accuracy boundaries using algebraic, rather than probabilistic concepts. The following points have been discussed.

(a) Probabilistic boundaries tend to be excessively “pessimistic” when applied to very small populations. In practical terms this would mean that a desired accuracy level would be achieved with smaller sample size.
(b) It is possible to set-up algebraic (i.e. non probabilistic) boundaries that have the same properties as the probabilistic ones.
(c) Algebraic boundaries perform better with population sizes not exceeding 900 elements. Experience shows that beyond that size probabilistic boundaries should to be used.