4. Global accuracy boundaries

In this section readers will be presented with a step-by-step approach aiming at the following propositions and conclusions:

(a) At equal sample size accuracy in concave populations is lower than in flat or convex populations.
(b) Sampling accuracy in concave and binary populations with 0-1 elements at equal proportions, is a global minimum for all population types and can therefore be used to formulate lower accuracy boundaries for concave populations. Such boundaries will only depend on population size.
(c) Sampling accuracy in flat populations is a global minimum for convex populations and can thus be used to formulate lower accuracy boundaries that will only depend on population size.
(d) Global accuracy boundaries offer the major advantage that safe sampling schemes can be planned in advance (i.e. *a priori*). No prior knowledge about the population parameters is required, except some idea on its size.

4.1 Impact of population density to accuracy

In Section 2.2 a population was described as “convex” when its density is higher near the mean, “flat” when its density is more or less uniform, and “concave” when its density is higher near the boundaries. It was also stated that this categorization would have a direct impact to the sampling accuracy. In this first of a series of propositions it will be shown that by making a normalized population more concave, its variance increases with the result that sampling accuracy decreases (refer also to Section 3.6).

Proposition 1

The variance of a normalized population increases when one of its elements is shifted away from the population mean.

Proof:

Let us consider a normalized population with N elements $u_1, u_2, ..., u_N$ and population mean μ. We also select arbitrarily an element u such that $u<\mu$. By considering all the other N-1 elements $u_k \neq u$, the population mean and variance will be:

$$\mu = \frac{1}{N}\sum u_k + \frac{1}{N}u \qquad (4.1)$$

$$\sigma^2 = \frac{1}{N}\sum (u_k - \mu)^2 + \frac{1}{N}(u-\mu)^2 \qquad (4.2)$$

Element u is then shifted away from μ and towards 0, by applying a negative increment $du<0$ (Figure 4.1). The impact of du on the population mean and the variance is found by differentiating the above two expressions with respect to u. We find that:

$$d\mu = \frac{1}{N}du \qquad (4.3)$$

$$d\sigma^2 = \frac{1}{N}\sum 2(u_k - \mu)(-\frac{du}{N}) + \frac{1}{N}2(u-\mu)(du - \frac{du}{N}), \quad \text{or}$$

$$d\sigma^2 = -\frac{2du}{N^2}\sum (u_k - \mu) + \frac{2du}{N^2}(u-\mu)(N-1)$$

Since $\sum (u_k - \mu) + (u-\mu) = 0$ the last expression is reduced to:

$$d\sigma^2 = \frac{2du}{N^2}(u-\mu) + \frac{2du}{N^2}(u-\mu)(N-1)$$

$$d\sigma^2 = \frac{2du}{N}(u-\mu) \qquad (4.4)$$

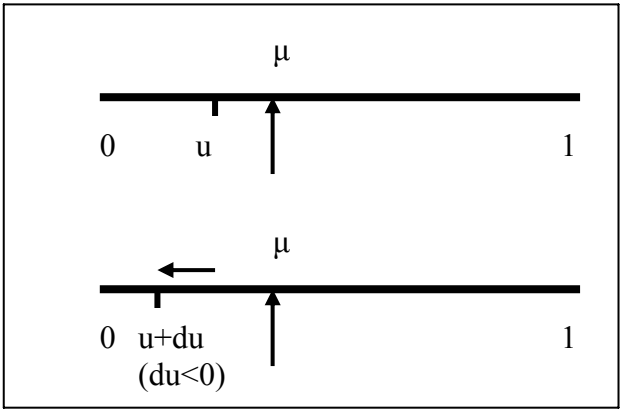

Figure 4.1. Moving an element toward the lower limit and away from the population mean will make the normalized population more concave.

Expression (4.4) indicates that the impact of du to the population variance is positive since we had selected $u<\mu$ and $du<0$ (the same conclusion would have been derived by assuming $u>\mu$ and $du>0$).

The new population resulting from the elementary transformation of the arbitrary element u has the following two properties:

a) it is more concave than the original population since its density has decreased near the mean and increased near one of the two boundaries;
b) its variance is higher than that of the original population.

Proposition 1 is proved. To be noted that in the transformed population, and because of (4.3), the new element $u+du$ will still remain to the left ($du<0$) or to the right ($du>0$) of the population mean, which means that if the above process is repeatedly applied on the same element u, it will finally make it equal to 0 or 1.

4.2 Upper limits for variance

Proposition 2

Any normalized population can be transformed to a concave population with only 0-1 elements and with a higher variance.

Proof:

According to proposition 1 any set of normalized elements can be transformed to a population with higher variance through an elementary increase or decrease of the value of one of its elements. Repeated transformations of the same element will finally make it become 0 or 1. By expanding this process to include all the other elements, the original population will finally become a population with values 0 and 1 only, and its variance will be higher than that of the original population.

Proposition 3

The maximum variance in normalized populations is 0.25.

Proof:

According to propositions (1) and (2), the variance of any normalized population will always have an upper limit determined by a concave population with 0- elements. The question then is which proportion of the 0-1 elements will result in the highest (=global) variance.

If p is the unknown proportion of the zero elements, the population variance will be $p\,(1\text{-}p)$. It can be seen that its maximum value is 0.25, occurring when $p=0.5$, that is when the 0 and 1 elements appear at equal proportions.

Proposition 4

The variance of a normalized and flat population is closely approximated by:

$$\sigma_f^2 = \frac{2N-1}{6(N-1)} - \frac{1}{4} \qquad (4.5)$$

Proof:

A normalized flat population can be approximated by a normalized population with mean equal to 0.5 and N elements $u_1, u_2, ..., u_N$ defined as:

$$u_i = \frac{i-1}{N-1}, \quad i=1,2,...N$$

The population variance will thus be:

$$\sigma_f^2 = \frac{1}{N}\sum (u_i - 0.5)^2 = \frac{1}{N(N-1)^2}\sum (i-1)^2 - \frac{1}{N(N-1)}\sum (i-1) + \frac{1}{4}$$

Expression (4.5) is derived by recalling the algebraic properties:

$$\sum (i-1)^2 = \frac{(N-1)\ N\ (2N-1)}{6} \quad \text{and} \quad \sum (i-1) = \frac{(N-1)\ N}{2}$$

4.3 Accuracy boundaries for concave populations

The results and conclusions of the previous propositions will be the basis for the formulation of accuracy boundaries that will depend only on population size.

The first task will be the formulation of accuracy boundaries for concave populations. This will have immediate application in populations of boat activities which, as discussed in Section 2.2, consist of 0-1 elements at varying proportions.

Setting up of a global accuracy boundary should be feasible if a fixed normalized population with maximum population variance could be identified. According to Proposition 3 in the previous section, such a population does exist and consists of 0 and 1 values at equal proportions. The variance of this population constitutes a global upper limit for all population categories of the same size N (whether convex, flat or concave), and this limit is given by:

$$\sigma_g^2 = 0.25 \tag{4.6}$$

By recalling (3.8), and the fact that the standard deviation σ will always be smaller than 0.5, the general expression for a global lower boundary for accuracy will take the form:

$$G(n) = 1 - z\frac{0.5}{\sqrt{n}}\sqrt{1 - \frac{n}{N}} \tag{4.7}$$

It is reminded that z is usually set to 1.96.

Figure 4.2 illustrates the application of expression (4.7) in the case of a 0-1 population with size N=1000. Also plotted is the population-

specific boundary defined by expression (3.8). It is recalled that the sample variable used to clarify the plot is $\log n / \log N$.

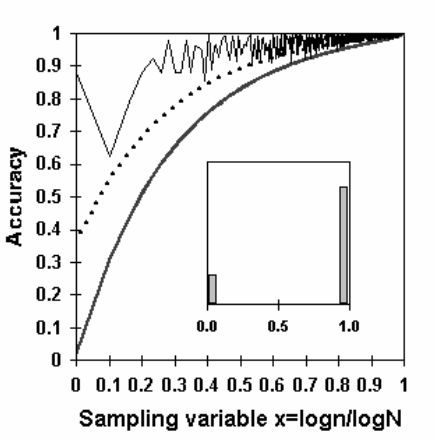

Figure 4.2. Global accuracy boundary $G(n)$ and population specific boundary (dotted line) for a 0-1 population with size $N=1000$.

The practical result of this approach is that for 0-1 populations of equal size the global boundary $G(n)$ is standard and remains unchanged, while the population-specific curve is variable and depends on the population variance.

4.4 Accuracy boundaries for convex populations

A second task is the formulation of accuracy boundaries for convex populations. This will have immediate application in populations of landings which, as discussed in Section 2.2, are frequently skewed and, at times, normal or flat (with uniform density).

In theory the global boundaries already formulated for 0-1 populations would also apply to convex populations, as it has been shown that in the latter category accuracy will always be higher. However this would lead to a rather “over-pessimistic” selection of sampling approach that would use far larger samples than actually required.

Therefore the objective here is to identify a fixed normalized population with maximum population variance among all flat or convex populations.

Proposition 4 states that the variance of a flat population is closely approximated by:

$$\sigma_f^2 = \frac{2N-1}{6(N-1)} - \frac{1}{4} \qquad (4.8)$$

On the other hand, Proposition 1 states that all normalized populations can be transformed to a “more concave” population with higher variance. Since a flat population will always be “more concave” than any convex population, it follows that its variance (see above formula) will be an upper limit for all convex populations. From which it is concluded that its accuracy boundary will be a global boundary for all convex populations.

By recalling (3.8), and the fact that the standard deviation σ will always be smaller than the value given in (4.8), the general expression for a global lower boundary for accuracy will take the form:

$$C(n) = 1 - z\frac{\sigma_f}{\sqrt{n}}\sqrt{1-\frac{n}{N}} \qquad (4.9)$$

with σ_f defined as in (4.8) and z usually set to 1.96.

Figure 4.3 illustrates the application of expression (4.9) in the case of a convex and skewed population with size N=1000. Also plotted is the population-specific boundary defined by expression (3.8). It is recalled that the sample variable used in the plot is $\log n / \log N$.

The practical result of this approach is that for convex populations of equal size the global boundary $C(n)$ is standard and remains unchanged, while the population-specific curve is variable and depends on the population variance.

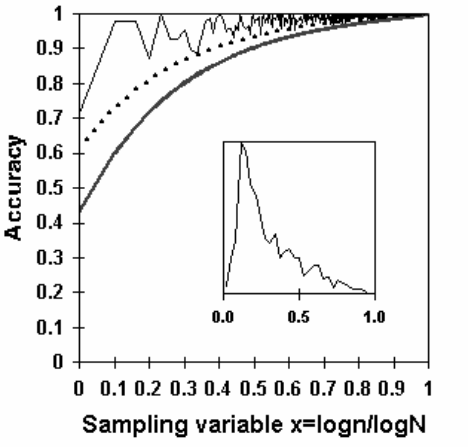

Figure 4.3. Global accuracy boundary $C(n)$ and population specific boundary (dotted line) for a convex population with size N=1000.

4.5 Exponential form of accuracy boundaries

It has already been mentioned that to facilitate reading of accuracy plots, the variable $x = \log n / \log N$ is used to denote sample size, rather than the proportion n/N. In this manner sample size n is written as:

$$n = N^x \qquad \text{with} \qquad x = \frac{\log n}{\log N}$$

and expressions (4.7) and (4.9) for concave and non-concave populations take the exponential form:

$$G(x) = 1 - z\frac{0.5}{\sqrt{N^x}}\sqrt{1 - N^{x-1}} = 1 - 0.5z\sqrt{N^{-x} - \frac{1}{N}} \tag{4.10}$$

$$C(x) = 1 - \sigma_f z\sqrt{N^{-x} - \frac{1}{N}} \tag{4.11}$$

All plots illustrating accuracy boundaries have, in fact, made use of expressions (4.10) and (4.11).

4.6 Critical sample size

We will now prove that critical sample size is reached when $x = \frac{\log n}{\log N} = 0.5$ (equivalent to $n = \sqrt{N}$), and it is at that value that the exponential boundaries reach a breakpoint and start a steady and slow growth versus 1.

We start with the observation that for each $A(x)$ defined in either (4.10) or (4.11), there exists an associate curve $B(x)$ of the form:

$$B(x) = 1 - \sigma\, z\sqrt{1 - \frac{1}{N}} + \sigma\, z\sqrt{N^{x-1} - \frac{1}{N}} \qquad (4.12)$$

The following relations apply:

$B(0)=A(0)$, which means that both A and B start from the same point.

$B(1) = A(1) = 1$, which means that both A and B end at the same point.

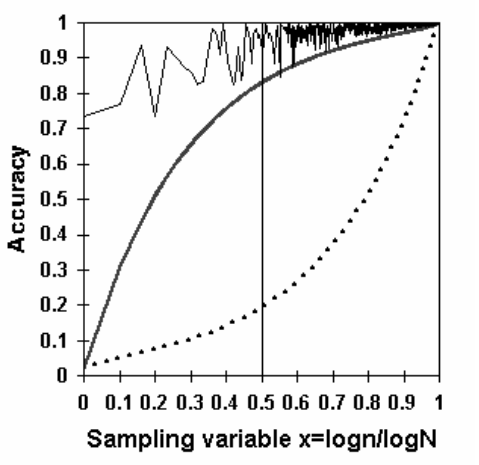

Figure 4.4. Graphical representation of an exponential boundary A(x) and its associated B(x) function (dotted line)

Figure 4.4 shows two contrary patterns of growth. Curve A shows a rapid growth up to a certain value of x and from then on it grows steadily until it becomes 1. Curve B 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. Curve B is the exact inverse of curve A.

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

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

$$\frac{d}{dx}\left[A(x)-B(x)\right]=0 \quad \text{or} \quad \frac{dA}{dx}=\frac{dB}{dx} \quad \text{or:}$$

$$\frac{N^{-x}\log N}{\sqrt{N^{-x}-\frac{1}{N}}}=\frac{N^{x-1}\log N}{\sqrt{N^{x-1}-\frac{1}{N}}} \tag{4.13}$$

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

The practical meaning of accuracy breakpoints is that by just knowing the population size, users may immediately get an idea about the minimum sample size at which the accuracy will be expected to become stable and growing. However, to obtain expected accuracy levels at variable sample sizes special tables have to be used, and this aspect will be covered in some detail in the coming sections.

4.7 Accuracy boundaries in infinite populations

As $N \to +\infty$ (cases of large or effectively infinite populations), expression (4.6) for variance remains the same while (4.8) takes its limit form: $\sigma_f^2 = 1/12$. Formulae (4.7) and (4.9) for lower accuracy boundaries are thus reduced to:

For all populations: $$G(n) = 1 - z\frac{0.5}{\sqrt{n}} \qquad (4.14)$$

For all flat and convex populations: $$C(n) = 1 - z\frac{1}{\sqrt{12}}\frac{1}{\sqrt{n}} \qquad (4.15)$$

The practical conclusion here is that when a population is known to be very large (i.e. its size is 30 000 elements or more), then even the knowledge of its exact size is not a requisite for setting up accuracy boundaries.

SUMMARY

In this section readers were presented with a step-by-step approach that achieved the following propositions and conclusions:

(a) At equal sample size accuracy in concave populations is lower than in flat or convex populations.
(b) Sampling accuracy in concave and binary populations with 0-1 elements at equal proportions, is a global minimum for all population types and can therefore be used to formulate lower accuracy boundaries for concave populations. Such boundaries will only depend on population size.
(c) Sampling accuracy in flat populations is a global minimum for convex populations and can thus be used to formulate lower accuracy boundaries that will only depend on population size.
(d) Global accuracy boundaries offer the major advantage that safe sampling schemes can be planned in advance (i.e. *a priori*). No prior knowledge about the population parameters is required, except some idea on its size.
(e) The two accuracy boundaries described in (c) and (d) are better described in exponential form.
(f) Sampling in finite populations results in accuracy that is highly fluctuating up to a certain critical sample size. Beyond that size accuracy growth becomes slower and stable. This breakpoint corresponds to the square root of the population size.
(g) In large and infinite populations the accuracy boundaries (c) and (d) take a simpler limit form and desired accuracy levels are independent of the population size.