Safety in sampling

Methodological notes

Prepared by

C. Stamatopoulos, FIDI
Senior Fishery Data Officer
Fishery Information, Data and Statistics Unit
FAO Fisheries Department

Rome 2003

The designations employed and the presentation of material in this publication do not imply the expression of any opinion whatsoever on the part of the Food and Agriculture Organization of the United Nations concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

Preface

Fishery statistical programmes require a great deal of effort and funds for their development and implementation and these are major constraints for many countries with limited human and financial resources. The merit of sampling approaches lies in providing cost-effective and efficient methods for the collection of data, thus accelerating the development of statistics urgently needed by fishery managers and planners.

Collection of basic data on catches, fishing effort and prices constitute a key factor in a wide variety of applications. Sample-based fishery surveys that are conducted on a regular basis constitute an important source of fishery information of wide utility and scope.

To help meet national needs for basic fishery data FAO has been assisting countries in upgrading their data collection, processing and reporting capabilities. Technical assistance at national and regional level is a significant component of the work programme of FAO's technical units responsible for fishery statistical development and involves both normative and field programme activities. Outputs of normative activities include technical documents on statistical methodology and guidelines for data collection, while field programme activities involve project formulation and implementation, technical backstopping and organization of training courses and workshops.

While the present paper was written with the special concern of sample-based catch/effort assessment surveys for artisanal fisheries, it is envisaged that several of its methodological and utility aspects could be applicable to other types of sample-based fishery surveys, particularly in cases where large-scale data collection programmes operate under financial and personnel constraints. Emphasis is placed on "safety in sampling" and some simple approaches are presented by

means of which statistical indicators regarding sampling accuracy are formulated in advance.

Methodological aspects and statistical indicators that relate to the accuracy and reliability of estimates are presented in handbook form. They summarize experience gained over the recent years in fishery statistical development by the Fishery Information, Data and Statistics Unit (FIDI) of the FAO. The concepts and methods included in the paper apply equally to both marine and inland capture fisheries and are presented in a manner that is generic enough to make them adaptable in commonly used data collection systems.

Readers interested in a more in-depth discussion on statistical and computing approaches, may make use of the list of references that is given at the end of the handbook.

Richard Grainger

Chief, Fishery Information, Data and Statistics Unit (FIDI)

Fisheries Department, FAO

Stamatopoulos, C.

Safety in sampling. Methodological notes.

FAO Fisheries Technical Paper. No. XXX. Rome, FAO, 2003. XXp.

ABSTRACT

The presented methodological notes address the question of sampling accuracy when sample-based data collection operations are performed under operational constraints, a frequent concern of fishery administrations with limited budget and human resources. Such a question is directly related to the frequency and extent of field operations for data collection. The paper focuses on an *a priori* determination of safe sample size using classical statistical methods appropriately adjusted to respond to specific target populations. The concepts and methods included in the paper apply equally to both marine and inland capture fisheries and are presented in a manner that is generic enough to make them adaptable in commonly used data collection systems.

Contents

1. Introduction

The approaches described in the paper address the question of sampling accuracy when sample-based data collection is performed under operational constraints, a frequent concern of fishery administrations with limited budget and human resources. Such a question is usually directly related to the frequency and extent of field operations for data collection and a number of classical methods are available for determining appropriate sample size on the basis of population parameters that have been derived at an earlier stage from the same or similar populations (see Cochran, 1977; Thompson, 1992; for discussion). This study focuses on *a priori* determination of safe sample size using the same methods and adjusting them to respond to specific target populations.

During the design phase of large-scale catch/effort assessment surveys for artisanal fisheries the question often arises as to what should be the appropriate sample size guaranteeing an acceptable level of reliability for the estimated population parameters. In some circumstances, and particularly at the early stages of implementing a fishery statistical monitoring programme, very little is known about the distribution and variability of the target population. Consequently, statistical developers tend to initially operate on a large-sample basis, with the intention of scaling down data collection as soon as some guiding statistical indicators, used for improving the cost-effectiveness of the sampling schemes, become available. Usually such indicators can only be formulated and verified after a complete operational cycle of a fishery statistical programme, which means that for long periods data collection is performed at high operational capacity. Generally, lack of any *a priori* guidance on sample size tends to increase the size and complexity of field operations and this, in turn, has a direct impact on the logistical aspects of data collection and data management procedures.

Most of the discussions in this publication concern a relative index of proximity between a sample mean and the population mean that

derives from a maximum allowable difference between a true population value and its estimate. This index, referred to as *accuracy* A throughout the paper, has several statistical and geometrical properties that are only a function of the population size or the knowledge that the population under study is large or infinite. Using these properties it is possible to formulate accuracy boundaries that can be used for predicting a lower limit for accuracy at any sample size. Construction of these boundaries is fairly simple and can be quickly achieved through the use of standard computing tools (such as spreadsheets) that are available on most personal computers.

Formulation of *a priori* boundaries for sampling accuracy consists of:

a) Guessing the general shape of the distribution of the target population in catch/effort assessment surveys (or accepting that no guessing is possible); and

b) Setting-up global accuracy boundaries that are only a function of the population size.

Before introducing the underlying concepts of *a priori* accuracy indicators, this introductory section will deal with some general aspects of sample-based fishery surveys with emphasis on:

(a) Basic fishery data.
(b) Justification for regular collection of basic fishery data.
(c) Scope and utility of basic fishery data.
(d) Need for fishery surveys to be cost-effective and sustainable.
(e) The key role of survey planners and statistical supervisors in the monitoring and evaluation of the performance of a data collection system.

1.1 Utility of basic fishery data

The definition of basic fishery data is rather empirical and based on the traditional method of setting-up general-purpose datasets that are subsequently used by a variety of application-specific systems.

In this handbook basic fishery data refer to catch, fishing effort, catch by species, first-sale prices (i.e. prices at landing), values, and fish size (in weight units).

A fishery statistical programme collecting basic fishery data should not be an end in itself. People involved in such programmes are sometimes challenged to provide a justification for regularly conducted (and thus costly) fishery surveys. From a long list of potential uses of basic fishery data, the following applications may be considered as representative:

1.1.1 Food Security

Food security is an over-riding concern for policy-makers, planners and administrators of natural living resources. In many communities, particularly in developing countries, fish is the major source of animal protein and people are dependent on fish as a food source.

Food balance sheets constitute a principal source of information for studies concerned with food security. Estimated total production of fish, combined with data on catch disposition, imports and exports, constitutes the basis for calculating *per capita* consumption of fish, which is subsequently used in the formulation of food balance sheets.

Basic data involved: Estimated total catch, estimated catch by species. Estimations are usually based on sampling approaches.

1.1.2 Fishing mortality

Fishing effort is one of the variables used to estimate fishing mortality. Fishing mortality is a fundamental variable in stock assessment, representing the proportion of stock that is removed due to fishing. Effort is used in setting most fishing controls. Changes in total fishing effort may be an indication of stock status or fishing profitability.

Basic data involved: Estimated fishing effort. Estimation is usually based on sampling approaches.

1.1.3 Fishing operations

Fishing operations indicators describe the composition of fishing fleets and fishing patterns and are the basis of most management decisions. They are important for monitoring compliance and in analyses involving fishing effort.

Basic data involved: Thematic maps of homeports and landing sites, numbers of fishing units by gear category, estimated fishing effort by boat/gear category. Effort estimates are usually based on sampling approaches.

1.1.4 Gear selectivity

It is often useful to obtain data indicative of the species that are targeted by different boat/gear categories and/or fishing methods, together with other information relating to the size of the fish being caught. These datasets are used for a wide variety of in-time and in-space comparisons of gear selectivity indicators.

Basic data involved: Species composition, average weight of fish. Such indicators are usually based on sampling approaches.

1.1.5 Abundance and exploitation

Catch-Per-Unit-Effort (CPUE) or catch rate is frequently the single most useful index for long-term monitoring of a fishery. It is often used as an index of stock abundance, where some relationship is assumed between the index and the stock size. It can also be used in monitoring economic efficiency. Catch rates by boat and gear categories, often combined with data on size at capture, permit a large number of analyses relating to gear selectivity and indices of exploitation.

Basic data involved: Sample Catch Per Unit Effort (CPUE).

1.1.6 Importance to national economy

For national and local policy-making and planning it is essential to describe the contribution of fisheries to the economy. Assessment of the economic contribution of fisheries needs to take into account several important variables and indicators, among which product prices and gross value of production.

Basic data involved: Estimated total catch, estimated catch by species, sample prices, estimated values. Estimations are usually based on sampling approaches.

1.1.7 Fleet performance and profitability

Boat profitability is a vital micro-economic indicator of fishery performance since it provides a measure of economic sustainability of artisanal fleets. Prices at landing, combined with data on investment and operational costs can provide indices of fleet performance.

Basic data involved: Overall Catch-Per-Unit-Effort (CPUE), unit value (average price) of catch. Such indicators are usually based on sampling approaches.

1.1.8 Socio-economic studies

Time series of fishing effort, catch, Catch-Per-Unit-Effort (CPUE) and prices are often used in socio-economic studies. Such data are indicative of declining or increasing trends of fisheries in districts and regions.

Basic data involved: Monthly data on catch, effort, CPUEs, prices and values. Estimations are usually based on sampling approaches.

In the description of the basic data involved in each of the listed applications, sample-based approaches are used for the estimation of key variables and the formulation of fishery indicators. Thus it becomes evident that the utility of the applications and their impact on fishery planning and management depends directly on the reliability of estimates resulting from sample-based data collection operations. This reliability and its relation to size and frequency of samples, together with its measurement and control constitute the main focus of this paper.

1.2 Cost-effective fishery surveys

Regularly conducted fishery surveys are costly since they involve salaries and wages of field and office personnel, recurring field operations costs and other overhead and maintenance costs relating to office infrastructure and operations. In many developing countries these costs constitute a major constraint and it is thus important for fishery statistical programmes to be as effective as possible at lowest cost. Sample-based fishery surveys are cost-effective when:

(a) They are economical in data collection effort and yet capable of producing reliable estimates.
(b) They make good use of existing human and financial resources involved in data collection and processing.
(c) They respond to users' needs in a timely and reliable manner.

The above three criteria indicate the need for realistic survey planning, well-defined sampling programmes and regular monitoring of survey results by means of meaningful and simple statistical indicators. As it will be seen in the coming sections of this document, most of these indicators are directly related to sampling accuracy.

SUMMARY

In this introductory section readers have been introduced to:

(a) The need for *a priori* statistical indicators that will guide the use of sufficient and appropriate samples and guarantee a desired level of accuracy in the estimates.
(b) Importance and utility of basic fishery data such as catch, effort, prices and values and a list of commonly used applications that make use of such data.
(c) The need for sample-based fishery surveys to be cost-effective and sustainable and some criteria for evaluating them from these two standpoints.

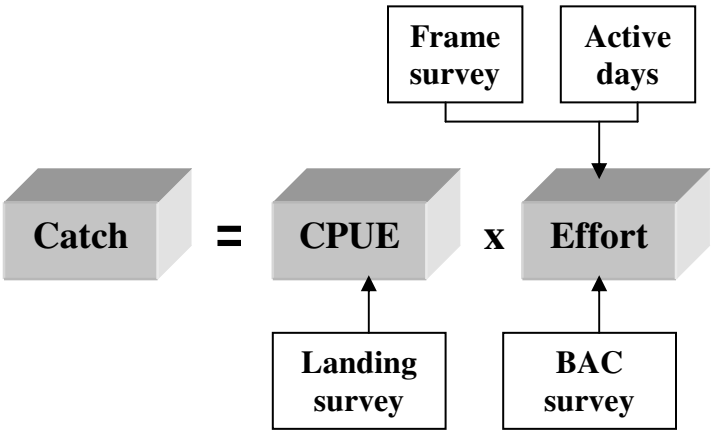
2. Concepts in estimating catch and effort

In this section readers will be presented with a generic approach for estimating total catch from basic fishery data and within an estimating context of a geographical stratum, a reference period and a specific boat/gear category. Other secondary data such as catch by species, values and average fish size, are estimated at a secondary stage and on the basis of the estimated total catch.

No complete enumeration (i.e. census) approaches for determining total catch are discussed in this handbook. In most small-scale fisheries the amount of information regarding total landings, species composition, prices, etc., is so large that the use of census approaches is impractical and sampling techniques are almost invariably employed.

2.1 A generic formula for estimating catch

The generic expression given below describes the relationship between *estimated* catch, *sample* CPUE and *estimated* effort.



A brief description of the variables involved is given below.

2.1.1 Catch (total)

Estimated total catch refers to all species taken together and is usually computed within the logical context of:

(a) A limited geographical area or stratum.
(b) A given reference period (i.e. a calendar month).
(c) A specific boat/gear category.

2.1.2 CPUE (sample, overall)

The sample Catch-Per-Unit-Effort (CPUE) is an *overall* average deriving from sampling and expressing how much fish (all species) is caught by a unit effort. Sampling context is the same as that for the estimated catch.

2.1.3 Effort

BAC (= Boat Activity Coefficient) expresses the probability that any boat (in general a fishing unit), is active (i.e. fishing) on any active day during the survey period.

2.1.4 Number of boats (from frame surveys)

This is a spatial extrapolating factor and relates to the total number of boats that are potentially operating in the geographical area of the estimation context. It is usually recorded by frame surveys conducted at relatively large time intervals. When multiplied by BAC it specifies the total number of fishing units that are expected to be active on any day of the survey period.

2.1.5 Active days

This is a time-extrapolation factor and specifies the total number of days that are assumed to be normal fishing days during the survey period. It is usually formulated by first considering the total number of calendar days and then reducing it according to empirically known factors, such as holidays, weekends, bad weather, etc.

This number accounts uniformly for days of normal activity, whereas the probability BAC accounts for individual variability of fishing activity.

2.1.6 Numerical example

The following theoretical example uses the formulae given above and illustrates a stepwise process for deriving primary estimates for catch and effort.

A. Assumptions

(a) Estimating context: *Lake Volta, Area VII, February 2001, Gillnets*
(b) Number of boats in Area VII (from frame survey): 2 000.
(c) Active days: 20.
(d) During February 2001 a total of 50 landings were sampled amounting to 500 kg and corresponding to 50 one-day fishing trips.
(e) During the same month a total of 100 canoes were examined for daily activities and 80 were found active (i.e. fishing).

B. Estimations

(f) From sample (e) we deduce that the probability BAC is 0.8, computed as the proportion of the number of canoes that were found active (80) over the total number of canoes sampled (100).
(g) This probability is then multiplied by 2000 to find the expected number of canoes that are active on any day. The result is 1600. Finally this number is multiplied by 20 in order to obtain total

number of fishing days. We thus find that fishing effort is estimated at 32 000 fishing days (or canoe days).

(h) From sample (d) we find that the mean CPUE is 500 kg divided by an associated effort of 50 days, i.e. CPUE = 10 kg day^{-1}.

(i) Finally, by applying the generic formula for estimating total catch, we multiply the estimated CPUE by the estimated effort and obtain a total of 320 000 kg.

2.2 Target populations and their distributions

Were it possible to know all landings made by a fleet during a survey reference period, then the distribution of these landings would correspond to one of the following three categories.

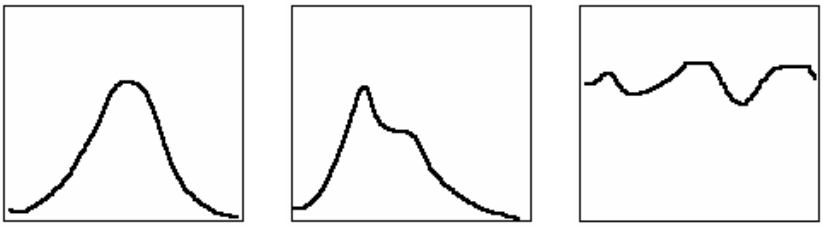


Figure 2.1. Normal, skewed and flat distributions of landings

In some cases the distribution of landings is normal (or close to normal), but this is not always possible to verify. In most cases landings follow an asymmetrical distribution (second plot), whereas flat (rectangular) shapes are at times observed with small pelagics.

However, the overall shape of the distributions of landings is “convex”, that is the frequency (or population density) increases near the mean and decreases near the boundaries or all values between and

including the population boundaries have approximately the same frequency.

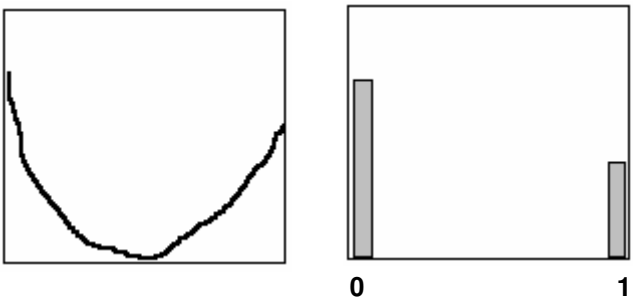


Figure 2.2. General shape of concave distributions and the specific case of 0-1 distributions describing boat activities.

In contrast to convex populations, the density of a “concave” population increases near the boundaries and decreases near the mean. The described set of boat activities is a specific case of a concave population with all of its values being equal to the boundaries 0 and 1 at varying proportions.

The observations made above on the two major categories of populations (convex and concave) will have a direct impact to the precautionary approaches in sampling operations and the manner in which *a priori* accuracy indicators will be handled.

SUMMARY

At this point readers should be familiar with the parameters involved in the estimation of CPUE, fishing effort and total catch. The following points have been emphasized:

(a) All estimations are performed within a specific logical context of a stratum, a reference period and a boat/gear category.
(b) Within each context, estimates of total catch are derived from a generic formula involving CPUE and fishing effort.
(c) CPUE estimates are derived from samples of catches (landings) combined with associated fishing effort.
(d) Estimation of effort is based on sample BACs (boat activity coefficients) that are raised to total effort by using spatial and in-time extrapolating factors.
(e) The population of landings is assumed to be “convex” that is with population density higher near the mean and thinning out near the boundaries.
(f) BACs are estimated through sampling from a population of 0-1 elements corresponding to non-fishing and fishing boats. This is a specific case of a “concave” population whose density is higher near the boundaries and thin around the mean.
(g) Categorization of populations into convex and concave will have a direct impact to the sampling approaches that will be described in the coming sections.

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.

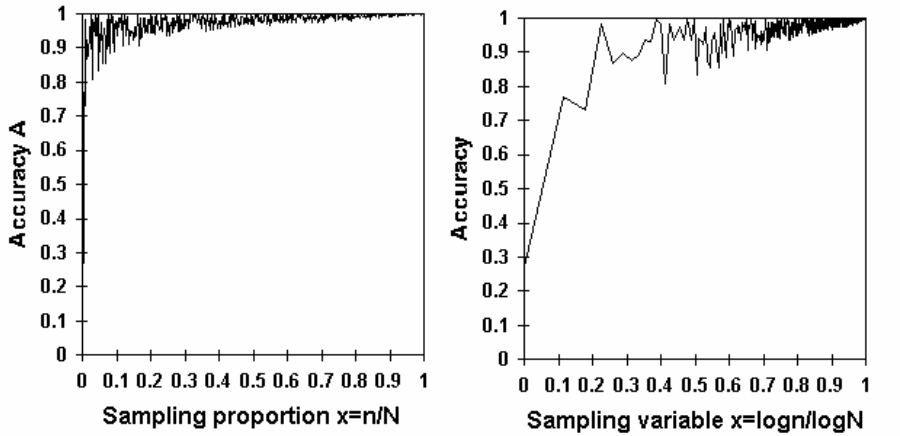


Figure 3.1. Accuracy plots from a normalized and finite population of size N. Accuracy values correspond to successive random samples with sizes 1, 2, 3, …, N. Notice the different shapes of the two plots depending on the expression used for sample proportion.

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.

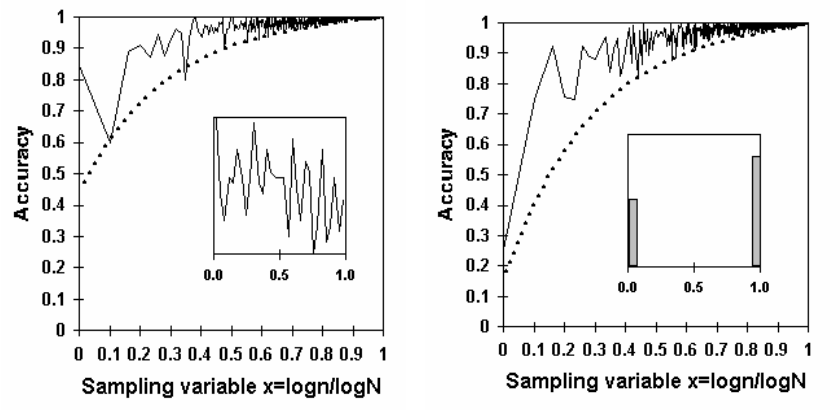


Figure 3.2. Fluctuating sampling accuracy and population-specific accuracy boundaries (dotted line) for two finite and normalized populations of size N=1000. The first population is flat, whereas the second is concave and binary.

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.

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:

$$\mu = \frac{1}{N}\sum u_k + \frac{1}{N}u \tag{4.1}$$

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

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:

$$d\mu = \frac{1}{N}du \tag{4.3}$$

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

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

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

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

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

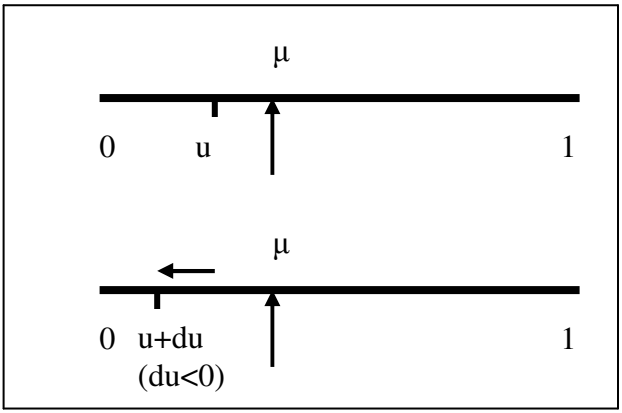


Figure 4.1. Moving an element toward the lower limit and away from the population mean will make the normalized population more concave.

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 $\log n / \log N$.

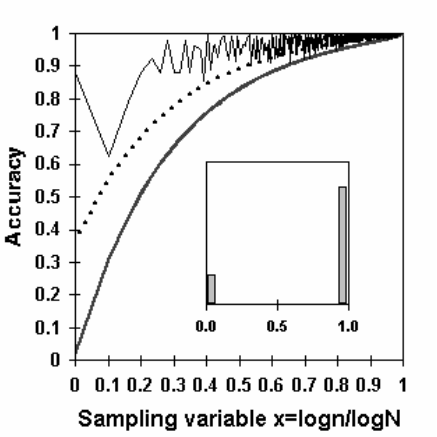


Figure 4.2. Global accuracy boundary $G(n)$ and population specific boundary (dotted line) for a 0-1 population with size $N=1000$.

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:

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

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:

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

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.

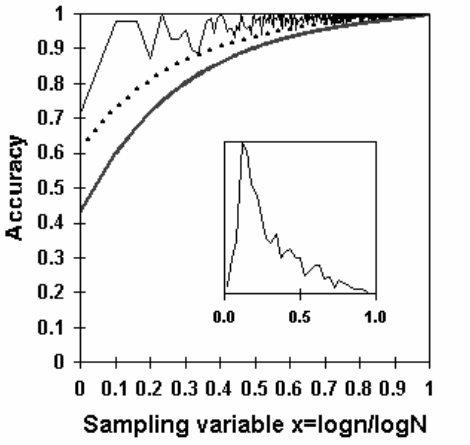


Figure 4.3. Global accuracy boundary $C(n)$ and population specific boundary (dotted line) for a convex population with size N= 1000.

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.

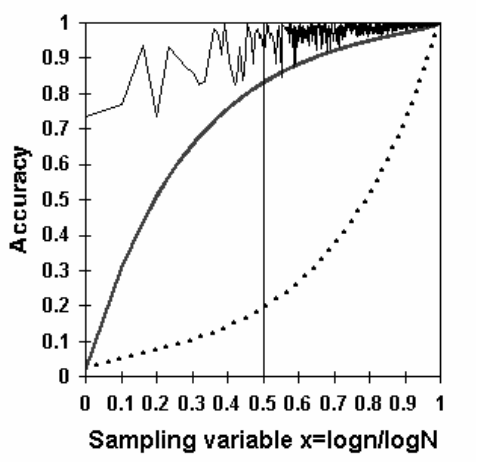


Figure 4.4. Graphical representation of an exponential boundary A(x) and its associated B(x) function (dotted line)

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.

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% (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.

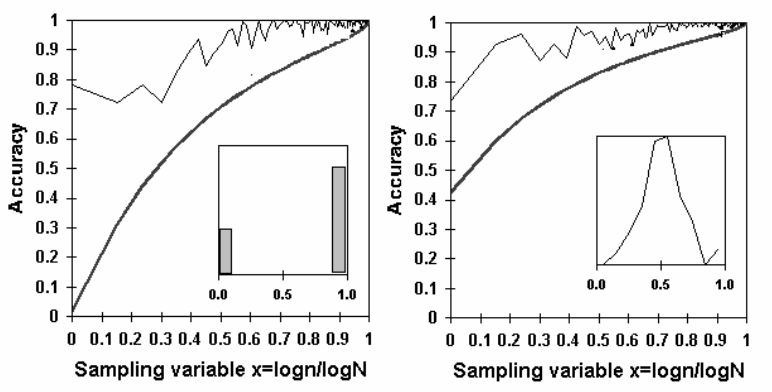


Figure 5.1. Accuracy plots and probabilistic accuracy boundaries for small concave (0-1) and convex populations. Population size is N=100. Notice the excessive safety space between global boundary curves and fluctuating sampling accuracy.

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.

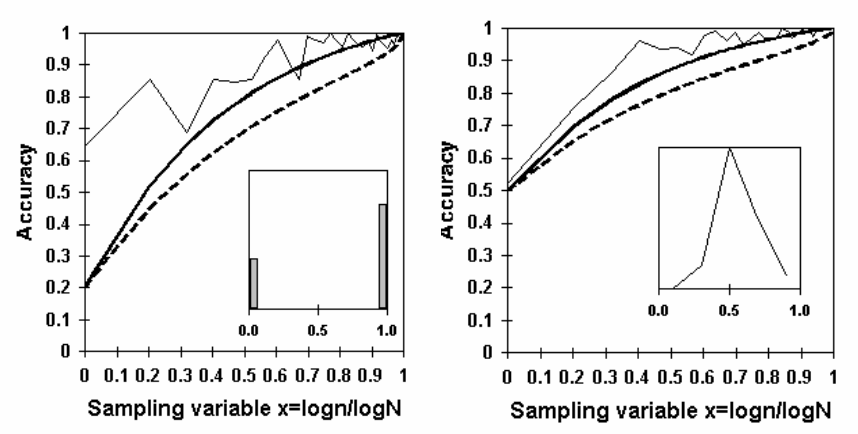


Figure 5.2. Accuracy plots and algebraic and probabilistic boundaries (dotted line) for two small populations with size N=30.

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:

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

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

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.

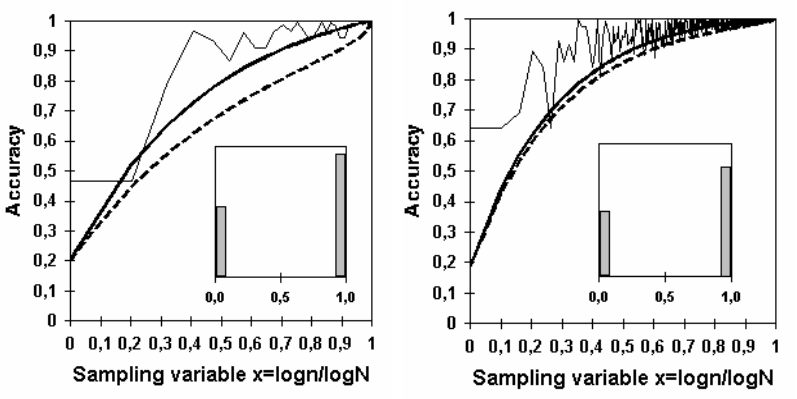


Figure 5.3. Algebraic and probabilistic boundaries (dotted line) in very small (N=30) and small/medium size (N=900) populations.

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.

6. Applicability aspects of accuracy boundaries

In this section readers will be presented with:

(a) Important questions in sample-based data collection schemes.
(b) Distinction between accuracy and precision indicators.
(c) Quick methods for *a priori* determining safe sample size.
(d) Practical guidelines to reduce risks of biased samples.

6.1 Important questions in sampling

In catch/effort assessment surveys for artisanal fisheries, size and frequency of samples for estimating fishing effort and catch, constitute critical methodological and operational decisions. The following is a list of most frequent questions concerning survey design:

1) How is data reliability measured?
2) What are the criteria for selecting sampling sites?
3) How should sampling operations be distributed over a reference period?
4) How many samples on boat activities and landings ought to be collected during each visit and how many samples should be totally available at the end of a reference period?

Questions such as those listed above become particularly pressing in the early phases of newly developed statistical monitoring programmes, when little is known about the distribution and variability of the target populations. The theoretical topics discussed so far can provide the basis for the formulation of *a priori* statistical indicators to be used for improving the cost-effectiveness of the sampling schemes.

6.2 Accuracy and precision in sampling

In sampling procedures *accuracy* and *precision* are two different statistical indicators and it is perhaps worth clarifying their meaning at this point.

6.2.1 Accuracy in sampling (theory already discussed)

(a) It is usually expressed as a relative index in percentage form (i.e. between 0 and 100%).
(b) It indicates the *closeness* of a sample-based parameter estimator to the true population value.
(c) When expressed as a relative index, it is independent of the variability of the population. In other words, population parameters of high variability can still be estimated with good accuracy which is essentially the primary issue in sampling.
(d) When sample size increases and samples are representative, sampling accuracy also increases. Its growth, very sharp in the region of small samples, becomes slower and steadier beyond sample size $n = \sqrt{N}$.
(e) Accuracy has its lowest value for $n=1$ and becomes 100% when the entire population has been examined (as in a census).
(f) The pattern of accuracy growth is not linear but follows a hyperbolic-type curve. The accuracy of a sample equal to half the population size is not 50% but much nearer 100%.
(g) Good accuracy levels can be achieved at relatively small sample sizes, provided that the samples are representative.
(h) Beyond a certain sample size the gains in accuracy are negligible, while sampling costs increase significantly.

6.2.2 Precision in sampling

(a) Precision is related to the variability of the samples used and measured as the inverse of the *coefficient of variation* (CV). CV is a relative index of variability that involves the sample variance and the sample mean.
(b) The CV index also determines the *confidence limits* of the estimates, which is the range of values that are expected to contain the true population values at a given probability.
(c) Estimates can be of high precision (that is with narrow confidence limits), but of low accuracy. This occurs when samples are not representative and the resulting estimates are systematically lower or higher than the true population value (cases of biased samples and estimates).
(d) When sample size increases the precision also increases as a result of decreasing variability. Its growth, very sharp in the region of small samples, becomes slower and steadier beyond sample size $n = \sqrt{N}$.

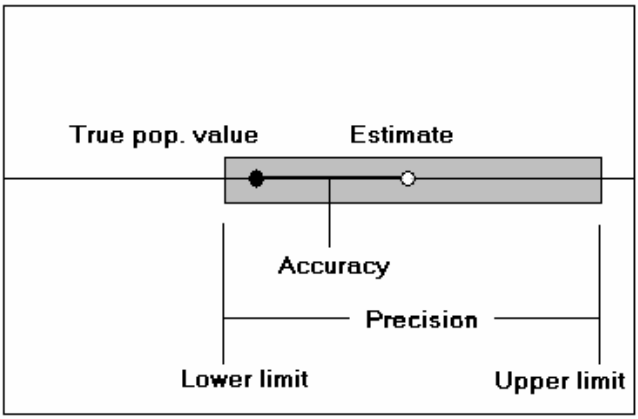


Figure 6.1 An illustrated example of accuracy and precision

Figure 6.1 illustrates the concepts of accuracy and precision. They are both important statistical indicators and regularly used for assessing the effectiveness of sampling operations. Their correct interpretation can greatly assist in identifying problem areas and applying appropriate corrective actions where and when necessary.

6.3 Design phase of a sample survey - guidelines

During this phase little is known about the distribution and variability of the target populations, and yet a number of decisions must be taken with regards to size and frequency of samples, so as to guarantee an acceptable level of reliability for the estimated population parameters.

According to the presented theory, *a priori* guidance on sample size requirements is feasible, provided that:

(a) The general shape of the distribution of the target populations is known.

(b) The population size can be determined with reasonable accuracy.

Regarding point (a) it has already been clarified that landings by all boats constitute a flat or convex population, whereas the set of 0-1 values equivalent to “boat not fishing”, “boat fishing”, is a specific case of a concave population. These two populations have different sampling requirements for achieving the same level of accuracy. The next paragraph provides more details as to how sample size is determined in each case and in accordance with the level of accuracy desired.

6.4 Safe sample size for landings and effort

The question of determining the desired accuracy level in a sampling and estimation process depends on the subsequent use of statistics and the amount of error that users are willing to tolerate. Experience

indicates that in basic fishery data the accuracy of estimates ought to be in the range 90 - 95%.

In setting-up safe sample sizes the first task is to approximately define the size of the target population.

In the case of boat activities population size is given by:

N = (No. of boats of a specific gear) x (No. of calendar days).

The above expression derives by the observation that boat activities can be considered as a two-dimensional matrix, the rows of which are represented by the boats that are potentially active and the columns by the days of a reference period. Each element of the matrix would only take the values of zero and one. Zero if a boat is not active, one if it is.

A more realistic N would exclude days for which fishing is known to be uniformly zero or negligible (such as standard non-working days, bad weather, etc.). However, in practice such considerations do not affect the sampling scheme and can be ignored.

With regards to the size of the population of landings the practical approach is to use the same N of the population of boat activities, the reason being that it is not possible to know on an *a priori* basis the actual fishing effort. In other words it is assumed that all boats were active on everyday. Again, this exaggerated assumption does not affect seriously the sampling considerations and at the same time simplifies the approach.

Populations with size between 10 and 900 are considered as “small” and for determining safe sample size the algebraic (non probabilistic) approach would provide more practical results. Tables A.1 and A.2 in Annex A illustrate sample sizes depending on the estimated population size and the desired level of accuracy.

For instance, if the desired level of accuracy is 95% and the number of boat activities is estimated to be about 600, then the suggested sample size (number of boats to be examined for activity status during the month), is n=87 (Table A.1). Likewise if the desired level of accuracy is 90%, then sample size required is only 34.

Table A.2 concerns the population of landings and it works the same way. If the desired level of accuracy is 95% and the number of landings is set to the theoretical maximum of 600, then the suggested sample size (number of landings sampled over the month), is n=47. Likewise if the desired level of accuracy is 90%, then sample size required is only 16.

Tables B.1, B.2 and B.3 in Annex B provide similar guidance with respect to large (N>900) or infinite (N>50,000) populations.

It is to be noted that the above sampling requirements refer only to a given estimating context, that is a geographical stratum, a reference period (i.e. calendar month), and a specific boat/gear category. The process of determining safe sample size at a given level of accuracy must be repeated for all estimating contexts with the view of determining overall sampling requirements.

6.5 Stratification and its impact on survey cost

Stratification is the process of partitioning a target population into a number of more homogeneous sub-sets. Stratification is normally based on the following three criteria:

(a) For statistical purposes and when there is a need to reduce the overall variability of the estimates.
(b) For non-statistical purposes and when current estimates are not meaningful to users of the statistics.
(c) At times stratification is “forced” due to administrative needs in terms of data collection and reporting functions and responsibilities.

Stratification is an expensive exercise and should always be applied with caution since all new strata would have to be covered by the sampling programme. Introducing a large number of strata may have serious cost implications for the following two reasons:

(a) Resulting strata will be more homogeneous than the original population, but the overall accuracy of the estimates will not be increased if data collection effort is kept at the original level.
(b) To fully benefit from a stratified population, safe sample sizes must be determined for each new stratum. In very large populations this would mean that a new set-up with three strata would need three times more samples for achieving the desired accuracy.

6.6 The problem of biased estimates

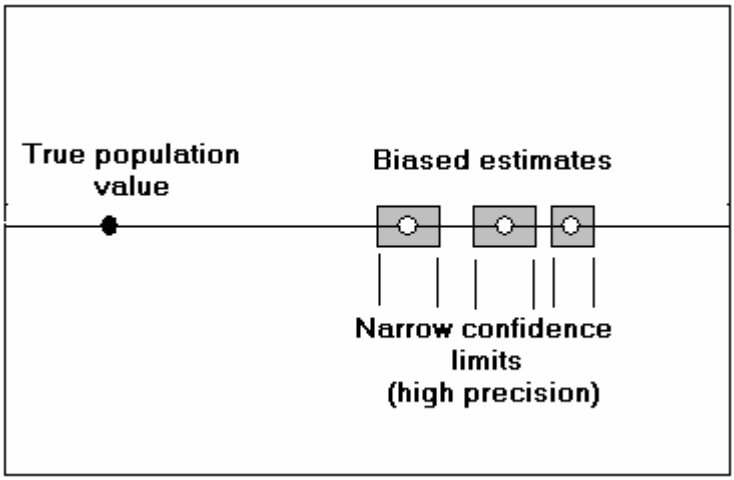


Figure 6.2. An illustrated example of biased estimates

In Section 6.2.2 dealing with variability indicators mention was made of biased estimates. Figure 6.2 illustrates in basic terms the problem of bias. To be noted that:

(a) Biased estimates are systematically lower or higher than the true population value and derive from samples that are not representative of the population. In example 6.2 estimates are systematically higher than the true value.
(b) Bias is independent of the precision (i.e. variability) of the estimates. In this example accuracy is bad but precision is misleadingly good and this is indicated by the narrow confidence limits.
(c) Users are unaware of such a situation since they do not know the true population value.
(d) Precision (or the relative variability indicator CV) cannot be used to detect bias.
(e) In general, bias is not easily detectable and at times it is not detectable at all.
(f) Bias can remain in the system even with drastic increases in sample size.
(g) Repeated cases of extremely small variability (<1% for instance), may be indications of a biased estimate.
(h) There exist no *a priori* indicators that could be used to safeguard sampling operations against systematic errors.
(i) Attempts to increase the representativeness of samples are often compromised due to operational constraints.

6.7 Need for representative samples

As already mentioned in earlier discussions the risks of biased data are considerably reduced if sampling operations collect data that are as representative as possible.

6.7.1 Data collection at sampling sites

Collection of representative samples at a sampling site is not a difficult task provided that data collectors are adequately trained and briefed. Points to be considered are:

Effort data

(a) Random selection of fishermen for information on activity status.
(b) Fishermen that are known to have been fishing should not be included in the sampling process.

Landing data

(a) When boats land within a short period, recorders at times tend to sample those with small catch in order to cover as many landings as possible. This introduces negative bias in the CPUEs and possibly in species composition.
(b) If landings occur over longer periods and recorders have to visit other sites during the day, only the first landings will be sampled. This may introduce bias in CPUEs, species composition and prices.

6.7.2 Selection of sampling sites

In the previous topic it was assumed that once a recorder has reached a sampling site he/she is capable of applying good sampling practices that were part of his/her training and brief.

In medium and large-scale fishery surveys the major problem in obtaining representative samples is associated with the first sampling stage that concerns *a priori* selection of locations at which data will be collected.

A good approach is to select sampling sites on a rotational basis. Field teams would then cover a good number of sampling locations by

visiting all of them at least once during a month. Such a sampling scheme requires sufficient and mobile human resources for data collection as well as a certain amount of survey planning work at the beginning of each operational month.

In most cases and due to operational constraints (accessibility, availability of data collectors, limited mobility, etc.), the above approach is not feasible and data collection is performed at fixed locations that for long periods constitute the sampling sites of the survey.

The risk of biased samples is thus associated with limited geographical coverage and the fact that pre-selected homeports or landing sites are not representative of the entire statistical area.

6.7.3 Criteria for selecting sampling sites

Selection of fixed sampling sites is usually done on an *a priori* basis through the use of frame surveys and existing geographical information.

(a) The geographical location of homeports and landing sites indicates requirements for in-space statistical coverage.
(b) The numbers of boats (fishing units) by site and boat/gear type indicate the relative importance of sites in potential fishing effort terms (i.e. very important, important, less important, etc.).

Thus, the criteria in selecting sampling sites are:

(a) Sampling sites ought to provide a satisfactory geographical coverage of the statistical area. This is usually the major operational constraint due to limited human resources and/or transportation means.
(b) Sampling sites ought to be representative of all boat/gear types involved in the survey.

(c) Sampling should focus on sites with larger numbers of fishing units.

6.7.4 Utility of analytical tools

Selection of suitable sampling sites is a common problem in survey programmes and the reader may find topics of interest in the references of this handbook, particularly in the FAO field documents. It may also be noted that in most cases fishery statistical programmes operate with limited human and financial resources. Due to these constraints the application of analytical techniques for sampling optimization is not always feasible.

Instead, simple and practical methods may serve as guiding, rather than optimizing, tools and an example of such an approach is discussed in the coming topic.

6.7.5 Example

Rather than examining sites on an individual basis, planners may look at *groups of sites* which, due to their mutual proximity, can offer a better statistical coverage.

Criterion for grouping several sites together is:

(a) whether a recorder can visit all grouped locations within the daily sampling schedule.
(b) whether the group of sites contains fishing units from most or all boat/gear types and in large enough numbers.

Figure 6.3 illustrates a hypothetical example of a minor stratum with 19 homeports. Table 6.1 contains the results of a frame survey for gillnets, beach seines and castnets.

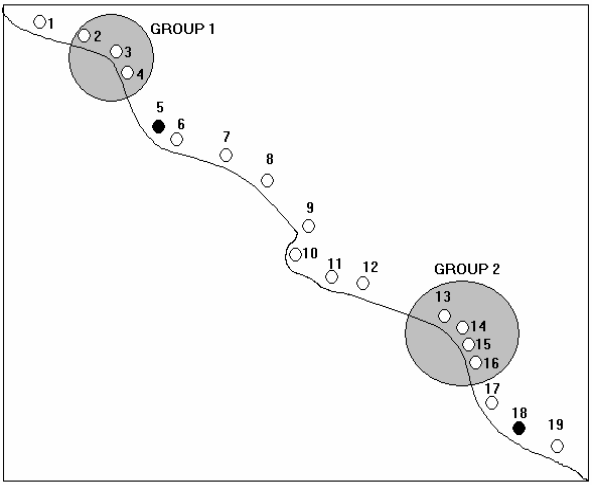


Figure 6.3. Grouping of sampling sites

On an individual basis, sites 5 and 18 are the most important sites since they contain the largest numbers of all boat/gear types. However, if secondary sites are looked as groups they can offer better statistical coverage, as indicated in Table 6.1. Thus, if planners consider the options of:

(a) Sampling from Sites 5 and 18, or
(b) Sampling from groups 1 and 2

the second option offers more statistical advantages for both in-space coverage and boat/gear representativeness.

Table 6.1. Frame survey data

Site	Gillnets	Beach seines	Castnets
1	4	0	7
2	11	0	0
3	1	8	2
4	5	0	9
Group 2,3,4	**17**	**8**	**11**
5	**12**	**4**	**5**
6	3	0	0
7	2	1	3
8	2	2	0
9	4	1	0
10	5	3	6
11	4	3	0
12	3	2	4
13	1	0	9
14	0	0	7
15	8	3	6
16	7	4	3
Group 13,14,15, 16	**16**	**7**	**25**
17	6	0	0
18	**14**	**5**	**9**
19	5	0	7

SUMMARY

In this section readers were presented with guidelines related to safe sample size and methods for reducing the risks of bias. The following points have been discussed.

(a) Distinction between accuracy and precision indicators.
(b) Practical ways for determining population sizes in landings and boat activities.
(c) Use of special tables providing safe sample size depending on population size and desired level of accuracy.
(d) Practical guidelines for *a priori* selection of sampling sites.
(e) Impact of stratification on survey cost.

7. A case study

A regional workshop was held in November 1998 in Tunisia with the purpose of presenting and discussing methodological and operational aspects in the design and implementation of sample-based catch/effort assessment surveys. The discussions focused on sampling and estimation involving mean daily catch and fishing effort, these two variables being the basis for estimating total catch. A case study was worked out concerning an artisanal fishery of 165 non-motorized gillnetters, operating from the ports of Bizerte and Ghar el Melh on a one-trip-per-day basis. Objective of the case study was to evaluate alternative sampling schemes for estimating two important catch/effort parameters:

(a) The Boat Activity Coefficient (BAC) expressing the probability that any boat would be active (i.e. fishing) on any day;
(b) The mean overall CPUE expressing the average catch per day of any boat (including all species).

In order to verify the applicability of the presented theory on safe sample size, use was made of census data relating to daily catches and boat activities.

7.1 Estimation of Boat Activity Coefficient (BAC)

A dataset of boat activities was made available showing recordings for all 165 boats during a reference period of 31 calendar days. Boat activities consisted of values of 1 and 0, 1 if a boat was active during a day or 0 if the boat was not active. Thus, the population of boat activities consisted of $\mathbf{N}$ = 31 x 165 =5,115 elements with values 0 or 1. The population mean (found to be 0.691) indicates the average activity level of a boat (BAC); that is the probability that any boat would be expected to be active on any day.

7.2 Estimation of CPUE through landings

A second dataset was also made available showing 3,555 recordings of daily catch for all 165 boats during the same reference period of 31 calendar days. Total catch was found to be 17,704 kg corresponding to a total effort of 3,555 fishing days, thus resulting in a population CPUE of 4.98 kg per boat per day. The maximum daily catch found in the dataset was 9.93 kg and the minimum zero; thus the population range of the daily catches is 9.93 kg.

7.3 Emulating "safe" sampling operations

7.3.1 Boat activities

In order to evaluate the operational requirements for future sampling operations and with the objective of deriving estimates of mean daily catch and mean boat activity, it was suggested that sampling from both datasets should target at a minimum sampling accuracy level of 95%.

Using Table B.1 it was found that for the concave population of 5,115 elements of boat activities a sample size of about n=358 would be needed in order to achieve the desired minimum accuracy level of 0.95.

In actual surveys this sample is composed of:

(a) The total number of boats of the sampling sites, as reported by a frame survey, multiplied by the number of sampling days. For instance if sites A and B are sampling sites and they contain 10 and 20 boats respectively, then the sample size over 8 days of sampling operations will be n=(10 + 20) x 8 = 240.
(b) If only sub-sets of the total number of boats are examined for state of activity (case of large ports), then the sum of all these numbers over the sampling period will provide the sample size.

7.3.2 Sample landings and sample CPUE

Using Table B.2 for the flat or convex population of daily landings, it was found that a sample size n=125 landings would guarantee the same minimum accuracy of 95%.

Sample CPUE is formulated by adding up all landings sampled and dividing by the associated fishing effort.

7.3.3 Remarks

In determining safe sample size the following points were considered:

(a) The population of boat activities was determined as the total number of boats (=165) multiplied by the number of calendar days in April (=30).
(b) The exact total number of landings cannot be predicted in a real situation. Consequently the size of the population of daily catches (although known to be 3,555) was set at the theoretical maximum N=5,115, assuming that all 165 boats made landings on every day;
(c) No other information regarding population means and ranges was used and sample sizes were determined only on the basis of population size and desired level for minimum accuracy. Census data were used for verification purposes only.

Table 7.1 illustrates 10 trial sampling operations applied to both datasets, each trial using the appropriate sample size determined above. For each trial operation the table shows the resulting mean daily catch and mean boat activity and resulting accuracy levels, as well as the estimated total effort and total catch. It is worth noticing that most of the resulting catch estimates compare well with the known total catch figure of 17,704 kg.

Table 7.1. Trial sampling and resulting estimates using census data of landings and boat activities (FAO Regional Workshop, Tunisia, 1998).

Trial	Sample CPUE (2)	Sampling accuracy	Sample BAC (4)	Sampling accuracy	Estim. effort (6)	Estim. catch (2)x(6)
1	5.123	0.985	0.685	0.994	3503	17945
2	5.128	0.985	0.702	0.989	3590	18409
3	4.905	0.993	0.674	0.983	3447	16907
4	4.769	0.979	0.721	0.970	3687	17583
5	4.867	0.989	0.680	0.989	3478	16927
6	4.962	0.999	0.710	0.981	3631	18017
7	4.848	0.987	0.649	0.958	3319	16090
8	5.182	0.979	0.669	0.978	3421	17727
9	5.068	0.991	0.699	0.992	3575	18118
10	4.978	1.000	0.685	0.994	3503	17437

SUMMARY

In this section readers were presented with a case study dealing with census data on landings and boat activities. The following points have been emphasized.

(a) Setting-up population sizes for boat activities.
(b) Considering the population size of boat activities as a theoretical maximum for determining the population size of landings.
(c) Using Tables B.1 and B.2 from Annex B in order to determine safe sample sizes achieving desired accuracy levels.
(d) Verifying that trial samples taken as per theory result in estimates that have a sampling accuracy that is higher or at least equal to the level selected.

8. Diagnostics on accuracy

In this section readers will be presented with some general points concerning numerical treatment of primary data (i.e. samples of boat activities and landings), including the formulation of statistical indicators relating to the reliability of resulting catch/effort estimates.

8.1 Estimation process

An estimation process usually involves the following computational steps:

8.1.1 Estimation of fishing effort

(a) Boat activity samples, active days and frame survey data are directed to the appropriate estimation context of a minor stratum, a month and a boat/gear type.
(b) BACs (Boat Activity Coefficients) are formulated in each context.
(c) The accuracy of BAC estimates is computed.
(d) The overall BAC variability and its confidence limits are computed.
(e) BAC variability is explained in space and time.
(f) BACs are combined with frame survey data and active days to produce estimates of fishing effort.
(g) Effort variability and confidence limits are computed.

8.1.2 Estimation of catch

(a) Sample landings are directed to the appropriate estimation context of a minor stratum, a month and a boat/gear type.
(b) Overall CPUEs are formulated in each context.
(c) The accuracy of CPUE estimates is computed.
(d) The overall CPUE variability and its confidence limits are computed.
(e) CPUE variability is explained in space and time.
(f) Sample species proportions are formulated.

(g) Sample prices are formulated.
(h) Estimates of average fish size (in weight units) are produced.
(i) Estimated CPUEs are combined with estimated effort to produce estimates of total catch.
(j) Variability of catch estimates and related confidence limits are computed. This compound parameter is based on the computed variances of effort and CPUE.
(k) Sample species proportions are combined with estimated total catch to produce estimated catch by species.
(l) Sample prices are combined with catch by species to produce estimated values by species.
(m) Values by species are added up to produce total values for landings.

8.1.3 Data grouping

The computational steps given above are repeated for each estimation context of a minor stratum, a month and a boat/gear type. At the end of this process the following data grouping procedures are performed:

(a) Catch, effort and values are grouped at major stratum and grand total levels.
(b) Average CPUEs and prices are formulated at major stratum and grand total levels.

8.2 Basic reporting

There are several ways to prepare basic reports on estimated data and this topic only provides some examples. Generally, monthly catch/effort estimates constitute “first generation” statistics that do not require much sophistication in its reporting functions. The following points may be considered:

(a) First reporting level must be that of the estimation context where all computations and related statistical indicators and diagnostics are produced.
(b) Before analyzing the results, users should check the system messages and diagnostics to ascertain the level of completion of each estimating context.
(c) All data involved in the estimation process must be reported so as to allow manual verification of the results if needed.
(d) Reporting sequence usually follows the computational steps discussed earlier.

8.3 System diagnostics

8.3.1 Messages issued during an estimation process

The example given below illustrates system messages that were produced during an estimation process. For each estimation context messages are displayed describing the outcome of the estimations.

```
KETA                          Beach Seine                   Estimated

Accuracy for CPUE below 90%

.....................................................................................
KETA                          Hook & Line                   Not estimated

No active days
No frame data
.....................................................................................
KETA                          Set Net                       Not estimated

No landings
No effort data
.....................................................................................
KETA                          Drifting Gillnet              Estimated

Only one site for landings
Only one site for effort
Accuracy for BAC below 90%
Accuracy for CPUE below 90%
No variance computed for CPUE
```

Figure 8.1. Messages issued during an estimation process

Messages displayed for different estimation contexts inform users that:

(a) Accuracy for CPUE is below 90%. Estimation continued.
(b) No extrapolating factors. Estimation failed.
(c) No landings and no effort data. Estimation failed.
(d) Limited geographical coverage. Accuracy levels for BAC and CPUE are below 90%.

8.3.2 Messages relating to estimated effort

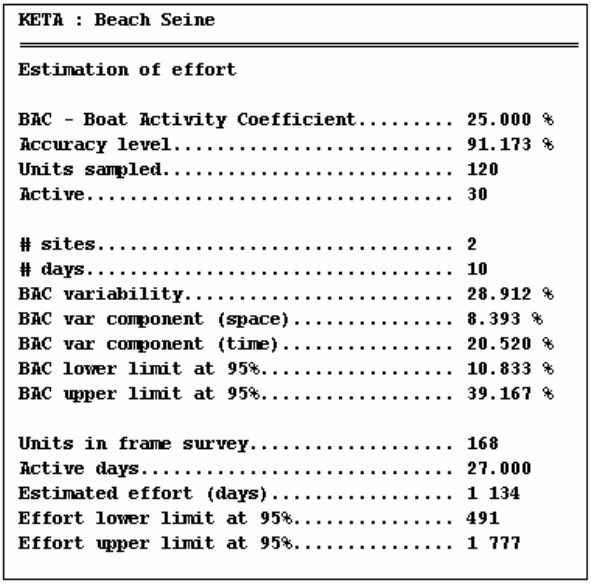


Figure 8.2 Messages relating to estimated effort

In the example given by Figure 8.2, estimated effort is described in three sections.

(a) The estimation of BAC and resulting accuracy can be verified with the sampling information displayed.
(b) The variability of BAC is high (29%) and is explained in space and time. Note that variability in time (20.5%) is the primary cause.
(c) The estimation of fishing effort can be verified using the estimated BAC and the data on active days and frame survey extrapolating factors. Confidence limits for estimated effort are also displayed.

8.3.3 Messages relating to estimated CPUE and catch

```
Estimation of catch

CPUE..................................... 402.967
Accuracy level........................... 89.798 %
Smp. size required for accuracy  90%.... 31
Landings sampled......................... 30
Sample catch............................. 12 089
Sample effort............................ 30

# sites................................. 2
# days.................................. 10
CPUE variability......................... 31.993 %
CPUE var component (space)............... 4.421 %
CPUE var component (time)................ 27.572 %
CPUE lower limit at 95%.................. 150.284
CPUE upper limit at 95%.................. 655.650

Estimated catch (Kg) .................... 456 964
Catch variability........................ 43.121 %
Catch lower limit at 95% (Kg) ........... 70 747
Catch upper limit at 95% (Kg) ........... 843 182
```

Figure 8.3. Messages relating to estimated CPUE and catch

In this example given by Figure 8.3, estimated CPUE and catch are described in three sections.

(a) The estimation of overall CPUE and resulting accuracy can be verified with the sampling information displayed. It should be noted that the resulting accuracy is slightly below the acceptable level of 90% because 30 samples, instead of the 31 required, were used.
(b) The variability of CPUE is high (32%) and is explained in space and time. Note that variability in time (27.5%) is the primary cause.
(c) The estimation of total catch was verified using the estimated CPUE and the estimated fishing effort described earlier. The compound variability of catch is very high (43%) because of the high variability in CPUE and fishing effort. Confidence limits for estimated total catch are also displayed.

8.3.4 Catch by species

Total value (1000 C) 221 571
Average price (1000 C/Kg) 0.485

Catch by species	Quant. Effort	CPUE Aver.W	Value Price
Anchovy	362 899 (79.4%) 1 134	320.017 0.000	152 244 (68.7%) 0.420
Burrito	26 366 (5.8%) 1 134	23.250 0.000	8 490 (3.8%) 0.322
Round Sardinella	29 030 (6.4%) 1 134	25.600 0.000	28 341 (12.8%) 0.976
Scad Mackerel	38 669 (8.5%) 1 134	34.100 0.000	32 496 (14.7%) 0.840

Figure 8.4. Example of a report showing catch by species

In the example given by Figure 8.4, results by species are displayed in three columns describing:

(a) Estimated catch by species and related effort.

(b) CPUE by species.
(c) Average weight per species.
(d) Sample price and estimated value by species.

Total value of all landings and their unit-value are displayed on the top.

8.3.5 Grand totals

GRAND TOTALS : Drifting Gillnet

Units in frame survey..... 4
Estimated effort (days)..... 27
CPUE..... 35.000
Estimated catch (Kg) 945
Total value (1000 C) 851
Average price (1000 C/Kg) 0.900

Catch by species	Quant. Effort	CPUE Aver.W	Value Price
Atlantic Little Tuna	203 (21.4%) 27	7.500 0.000	162 (19.0%) 0.800
Sharks	473 (50.0%) 27	17.500 0.000	473 (55.6%) 1.000
Skipjack Tuna	270 (28.6%) 27	10.000 0.000	216 (25.4%) 0.800

Figure 8.5. Example of a report on grand totals

In the example given by Figure 8.5, grand totals are computed for a specific boat/gear type (drifting gillnet). These figures have resulted from grouping all statistics for this boat/gear type that were produced in different minor strata.

SUMMARY

In this section readers have reviewed general aspects concerning automatic processing of basic fishery data. The following topics were presented:

(a) Processing of primary data. Boat Activities and Landings.
(b) Data checking and monitoring.
(c) Estimation process. Data involved. Statistical indicators and diagnostics.
(d) Basic reporting functions.

.

9. Discussion

9.1 General applicability aspects

As mentioned in the introduction, the presented approach attempts to assist statistical developers with some practical guidance prior to the implementation of data collection operations and when very little is known about the populations under study. The author is of the opinion that such *a priori* guidance would offer a methodological supplement compatible with conventional statistical techniques and tools that are commonly applied in survey planning and design. Some questions, however, are likely to arise relating to the applicability of the presented approaches in the following instances:

(a) Size of the population is not known.
(b) The general shape of the distribution of the target population (i.e. whether convex, flat or concave) is not known.
(c) Aspects relating to data collection tactics, data representativeness and operational constraints.
(d) Criteria for setting-up minimum accuracy levels.

Question (a) can be answered in two ways.

If the population size is difficult to determine or guess but the population is considered sufficiently large, then Table B.3 for infinite populations can be used. This will be a rather pessimistic approach and the sample size will be, in fact, slightly larger than necessary. If, on the other hand, the target population is not very large then a maximum must be assumed. In catch/effort assessment surveys this is usually feasible (see case study in Section 7).

For question (b) the answer is simple. Boat activities always constitute a concave population with 0-1 elements, whereas landings are in most cases convex or occasionally flat.

Regarding question (c) the presented approaches do not go beyond the formulation of indicators relating to sample size and guaranteeing a minimum level of accuracy. What the survey planner knows is that if at the end of the month, n representative samples are available, then the expected accuracy of the estimated population parameter (CPUE, activity level, unit value, etc.), will be higher than the minimum accuracy limit used for determining sample size. The question of "how" to collect these data is the responsibility of the user.

As for question (d) the author's view is that it is up to the researcher to assess whether a certain level of proximity of the sample to the population mean is satisfactory or not. In general, any statement of accuracy desired is equivalent to expressing the amount of error that the user is willing to tolerate in the sample estimates and it is determined in the light of the uses to which the sample results are to be put (see Cochran, 1977; for discussion on this topic).

9.2 Stratification and its impact on sample size

Another factor is the impact of stratification on the sampling requirements in a fishery statistical programme. A more refined stratification applied to an existing data collection scheme would certainly improve the homogeneity aspects of the target population but would also have an impact on sampling effort. This factor is at times overlooked by statistical developers who continue to apply old sampling schemes proportionally to the size of the newly created target populations while maintaining the same total number of samples collected over the reference period. According to the observations and conclusions of the presented study this approach is not appropriate and safe sample sizes ought to be reviewed and adjusted after stratification.

9.3 Concluding remark

The presented approaches stress the point that in assessing sampling requirements, the target population of an estimation context (such as

daily catches of a specific boat-gear category over a month), ought to be viewed as a unique case and handled with criteria and sampling practices specific to its size and properties. This means that adapting criteria and practices applicable to other populations, however effective these are known to be, would not always constitute an appropriate approach. Experience shows that statistical developers, including the author, tend at times to think in terms of proportionality and assume that if a *sample proportion* has proved adequate for a population of given size and distribution, it would be expected to operate well also for a population of different size and/or distribution. The presented study indicates that if a sample proportion were good for a large population it would definitely not be as good for a much smaller population. Conversely, if a sample proportion has known to be effective for a small population, a much larger population would certainly require a lower sample proportion to achieve the same level of accuracy. It is the author's opinion that proportionality aspects (i.e. in what proportions samples should be collected from sampling sites), are relevant to the representativeness of samples, and ought to be considered only after the required number of samples has been determined.

10. Further reading

Banerji, S. K. (1980): The collection of catch and effort statistics. *FAO Fisheries Circular*, 730.

Bazigos, G. P. (1974). The design of fisheries statistical surveys. Inland waters. *FAO Fisheries Technical Paper*, 133.

Bazigos, G. P. (1975). Applied fishery statistics: vectors and matrices. *FAO Fisheries Technical Paper,* 135.

Bazigos, G. P (1976). Guidelines for the production of fisheries statistics. *FAO Training Courses in Fishery Statistics*.

Bazigos, G. P. (1983). Applied Fishery Statistics. *FAO Fisheries Technical Paper*, 135.

Bonzon, A. and Horemans, B. (1988). Socio-economic data base on African fisheries. *. FAO Fisheries Circular,* 810.

Brander, K. (1975). Guidelines for collection and compilation of fishery statistics*. FAO Fisheries Technical Paper,* 148.

Caddy, J. F. and Bazigos, G. P. (1985). Practical guidelines for statistical monitoring of fisheries in manpower limited situations. *FAO Fisheries Technical Paper*, 257.

Cochran, W. G. (1973). Sampling Techniques. John Wiley & Sons, New York. p. 9-17.

Deming, W. E., (1960). Sample Design in Business Research. John Wiley & Sons, New York.

FAO, (1993). Report of the Working Group on Artisanal Fisheries Statistics for the Western Gulf of Guinea, Nigeria and Cameroon, 49.

FAO, (1999). Guidelines for the routine collection of capture fishery data. *FAO Fisheries Technical Paper*, 382.

Hansen, M. H., Hurwitz, W. N., and Madow, W. G. (1953). Sample survey methods and theory. John Wiley & Sons, New York. p.58.

Sparre P. (2000). Manual on sample-based data collection for fisheries assessment. Examples from VietNam. *FAO Fisheries Technical Paper,* 398.

Stamatopoulos, C. (1996). Report on the use of a fisheries statistical software (ARTFISH). *FAO-IDAF Project Technical Report*, 83.

Stamatopoulos, C. (1999). Observations on the geometrical properties of accuracy growth in sampling with finite populations. *FAO Fisheries Technical Paper*, 388.

Stamatopoulos, C. (2002). Sample-based fishery surveys. A technical handbook. *FAO Fisheries Technical Paper*, 425.

Sukhatme, P.V. and Sukhatme B.V. (1970). Sampling Theory of Surveys with Applications. *FAO*, p.7-25.

Thompson, S. K. (1992). Sampling. John Wiley & Sons, New York. p. 11-34.

Annex A

Indicators of sampling accuracy in small populations

Table A.1. Sampling requirements at varying accuracy level and population size. Small concave populations.

Accuracy (%) Population size	90	91	92	93	94	95	96	97	98	99
10	6	6	6	7	7	7	8	8	9	9
20	9	9	10	11	11	12	13	15	16	18
30	11	11	12	14	15	16	18	20	23	26
40	12	13	15	16	18	20	22	25	29	34
50	14	15	16	18	20	23	26	30	35	41
60	15	16	18	20	23	26	30	34	41	49
70	16	17	19	22	25	28	33	38	46	56
80	17	19	21	23	27	31	36	42	51	63
90	17	19	22	25	28	33	39	46	56	70
100	18	20	23	26	30	35	41	49	61	76
200	24	27	31	36	43	51	62	78	102	138
300	27	31	36	43	51	63	78	101	136	193
400	30	35	41	48	58	72	91	120	166	244
500	32	38	44	53	64	80	103	137	193	292
600	34	40	47	57	70	87	113	153	218	338
700	36	42	50	60	74	94	122	167	241	382
800	38	44	52	64	78	100	131	180	263	424
900	39	46	55	66	82	105	139	192	284	464

Table A.2. Sampling requirements at varying accuracy level and population size. Small flat or convex populations.

Accuracy (%) Population size	90	91	92	93	94	95	96	97	98	99
10	4	5	5	5	6	6	7	7	8	9
20	6	7	7	8	9	10	11	13	15	17
30	7	8	9	10	11	13	14	17	20	24
40	8	9	10	11	13	15	17	20	25	31
50	8	9	11	12	14	16	19	24	29	37
60	9	10	11	13	15	18	22	26	33	44
70	9	11	12	14	16	19	23	29	37	49
80	10	11	13	15	17	21	25	31	41	55
90	10	12	13	15	18	22	27	34	44	61
100	10	12	14	16	19	23	28	36	47	66
200	13	15	17	20	25	31	39	52	74	114
300	14	16	19	23	28	36	47	65	95	154
400	15	17	21	25	31	40	53	74	112	190
500	16	18	22	27	34	44	58	83	128	223
600	16	19	23	29	36	47	63	90	141	254
700	17	20	24	30	38	49	67	97	154	283
800	17	21	25	31	39	52	71	103	166	311
900	18	21	26	32	41	54	74	109	177	337

Annex B

Indicators of sampling accuracy in medium, large and infinite populations

Table B.1. Sampling requirements at varying accuracy level and population size. Medium, large or infinite concave populations.

Accuracy (%) Population size	90	91	92	93	94	95	96	97	98	99
1000	88	106	130	164	211	278	375	516	706	906
2000	92	112	140	179	235	322	462	696	1091	1655
3000	93	114	143	184	245	341	500	787	1334	2286
4000	94	115	145	187	250	350	522	842	1500	2824
5000	94	116	146	189	253	357	536	879	1622	3288
6000	95	116	146	190	255	361	546	906	1715	3693
7000	95	117	147	191	257	364	553	926	1788	4049
8000	95	117	147	191	258	367	558	942	1847	4364
9000	95	117	148	192	259	368	563	954	1895	4646
10000	95	117	148	192	260	370	566	964	1936	4899
15000	95	118	149	193	262	375	577	996	2070	5855
20000	96	118	149	194	263	377	583	1013	2144	6488
25000	96	118	149	194	264	378	586	1023	2191	6939
30000	96	118	149	195	264	379	588	1030	2223	7275
35000	96	118	149	195	265	380	590	1036	2247	7536
40000	96	118	150	195	265	381	591	1039	2265	7745
45000	96	118	150	195	265	381	592	1042	2279	7915
50000	96	118	150	195	265	381	593	1045	2291	8057
> 50000	96	119	150	196	267	384	600	1067	2401	9602

Table B.2. Sampling requirements at varying accuracy level and population size. Medium, large or infinite flat/convex populations.

Accuracy (%) Population size	90	91	92	93	94	95	96	97	98	99
1000	31	38	48	61	82	114	167	262	445	762
2000	32	39	49	63	85	120	182	302	572	1231
3000	32	39	49	64	86	123	188	318	632	1549
4000	32	39	49	64	87	124	191	327	667	1778
5000	32	39	50	64	87	125	192	332	690	1952
6000	32	39	50	65	88	125	194	336	706	2088
7000	32	39	50	65	88	126	195	339	718	2197
8000	32	39	50	65	88	126	195	341	728	2286
9000	32	39	50	65	88	126	196	342	735	2361
10000	32	39	50	65	88	126	196	343	741	2425
15000	32	39	50	65	88	127	197	347	760	2638
20000	32	39	50	65	89	127	198	349	770	2760
25000	32	39	50	65	89	127	198	351	776	2838
30000	32	39	50	65	89	128	199	352	780	2893
35000	32	39	50	65	89	128	199	352	782	2933
40000	32	39	50	65	89	128	199	353	785	2964
45000	32	39	50	65	89	128	199	353	786	2989
50000	32	39	50	65	89	128	199	353	788	3009
> 50000	32	40	50	65	89	128	200	356	800	3201

Table B.3. Summary of sampling requirements at varying accuracy level for large to infinite populations.

Accuracy in %	Sample size for boat activities (boats sampled)	Sample size for Landing surveys (landings sampled)
90	96	32
91	119	40
92	150	50
93	196	65
94	267	89
95	384	128
96	600	200
97	1067	356
98	2401	800
99	9602	3201