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 percent).
(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 percent 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 percent but much nearer 100 percent.
(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 $n = \sqrt{N}$.

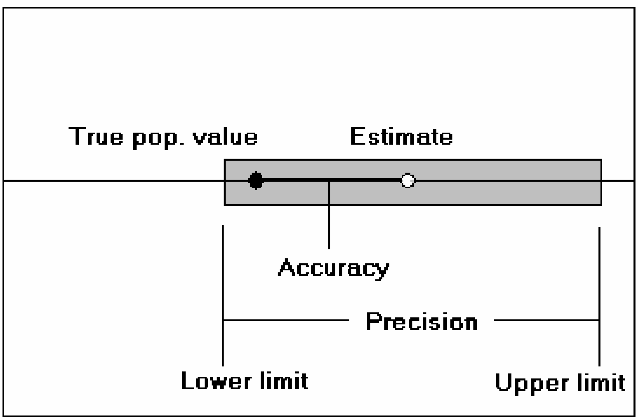

Figure 6.1 An illustrated example of accuracy and precision

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 percent.

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 percent 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 percent, 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 percent 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 percent, 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

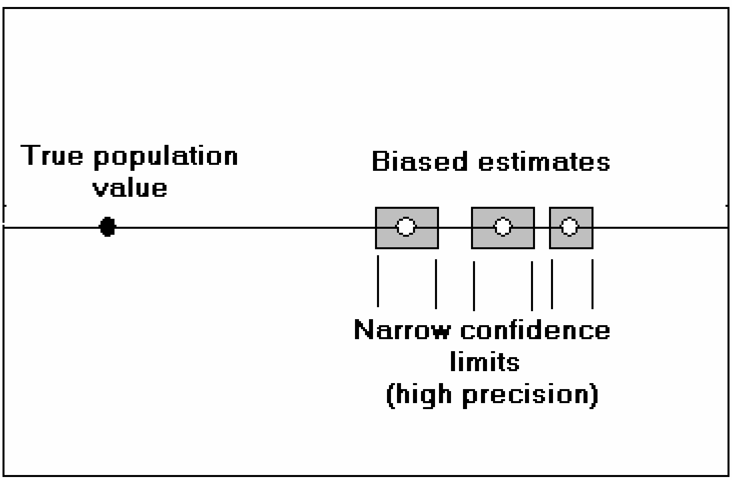

Figure 6.2. An illustrated example of biased estimates

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 percent 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:

<u>Effort data</u>

(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.

<u>Landing data</u>

(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.

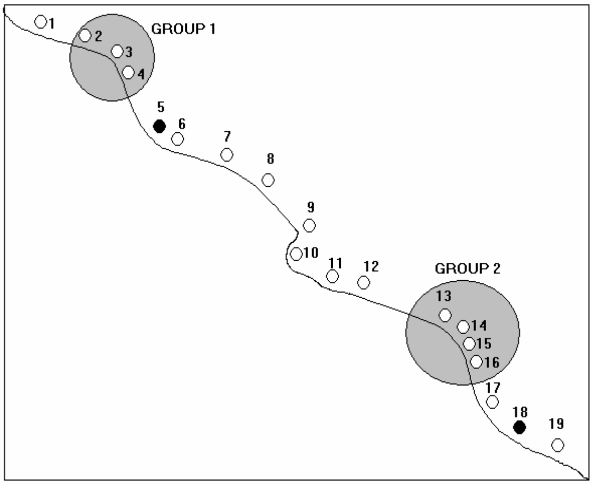

Figure 6.3. Grouping of sampling sites

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.