ICES WKPICS2 REPORT 2012

ICES ADVISORY COMMITTEE

ICES CM 2012 / ACOM: 52

REF. PGCCDBS, RCM'S, STECF/SGRN

Report of the second Workshop on Practical Implementation of Statistical Sound Catch Sampling Programmes

6 – 9 November 2012

ICES, Copenhagen

ICES International Council for the Exploration of the Sea
CIEM Conseil International pour l'Exploration de la Mer

International Council for the Exploration of the Sea
Conseil International pour l'Exploration de la Mer

H. C. Andersens Boulevard 44–46
DK-1553 Copenhagen V
Denmark
Telephone (+45) 33 38 67 00
Telefax (+45) 33 93 42 15
www.ices.dk
info@ices.dk

Recommended format for purposes of citation:

ICES. 2013. Report of the second Workshop on Practical Implementation of Statistical Sound Catch Sampling Programmes, 6 - 9 November 2012, ICES Copenhagen. ICES CM 2012 / ACOM:52 71 pp.

Contents

Executive summary

This workshop, chaired by Jon Helge Vølstad (Norway) and Mike Armstrong (UK) was held at ICES headquarters, Copenhagen, from 6–9 November 2012.

Prior to the workshop, participants from each country were provided with a questionnaire to summarize the survey sampling schemes and estimation procedures employed in national sampling programmes. These were collated and developed further at the workshop. WKPICS2 outlines four principal classes of probability-based sampling schemes, and discusses how sampling frames, primary sampling units and strata can be developed and optimised to deliver the required estimates for species, fleet metiers, fishing grounds or other variables of interest. Methods for design-based estimation procedures are described. Stratified probability-based sampling has the advantage that sample sizes per stratum can be controlled, thus minimizing the need for imputations to fill in data-gaps. The design-based estimators allow samples to be easily extrapolated to the target population using weighting factors based on inclusion-probabilities. Detailed description of design-based estimation is provided for an at-sea sampling programme where vessels are primary sampling units and for an on-shore catch sampling programme where site-days are primary sampling units. In the latter, vessel-trips are sampled for a random selection of ports and days. These two design classes result in a clustered sample of trips, and in general it is not reasonable to assume that a simple random sample of trips is obtained from the fleet. Detailed advice on estimation procedures for all principal design classes will be finalized in WKPICS3.

WKPICS2 has developed guidelines for “best practice” that covers the design, implementation and analysis stages of catch sampling schemes, assuming that regional objectives and data needs are clearly defined. Ideally, all national surveys should clearly document the sampling frame, sample selection procedures, response rates (e.g. refusals to take observers), imputation methods for missing data and weighting procedures employed to derive national estimates. Best practice can be defined as sampling designs, implementation and data analysis that minimize bias and maximize precision, and which makes the most efficient use of sampling resources. For example, probability-based sampling with accurate control of the inclusion probabilities would be considered an example of best practice. However, if logistical, legal, and economic constraints dictate the use of a non-probability based scheme to select primary sampling units (for example legal requirements in the selection of a reference fleet), it is good practice if the selection is done in a way that ensures representative coverage of the target population and minimises bias, and if this can be demonstrated with suitable diagnostics. Bad practice would be an ad-hoc, non-probability based sampling scheme, particularly where there are no census data to show how representative the samples are of the population or to re-weight the samples during analysis.

WKPICS2 also proposes revised data quality indicators, including a simple one-page form that can be used to evaluate quality of data used for stock assessments. It is recommended that the quality indicators be further refined through practical testing by Regional Coordination Groups and stock assessment working groups, based on several case studies.

WKPICS2 advises on future development of regional databases (RDB) and analysis software (FishFrame, Intercatch, and COST) to accommodate analysis of data that are collected according to best-practices survey sampling methods advocated by ICES WKPICS and SGPIDS. In particular, it is recommended that the RDB and analysis framework be further developed so that catch sampling schemes that result in clustered samples of trips can be accommodated.

1 Introduction

1.1 Terms of reference

WKPICS2 is the second of three workshops aimed at providing guidance on the design of fishery sampling programmes. The Terms of Reference for WKPICS2 are given below, and the scientific justification is given in Section 1.3. The proposed ToRs for WKPICS3 are in Annex 3.

2011/2/ACOM53. The Second Workshop on practical implementation of statistical sound catch sampling programmes (WKPICS2), chaired by Jon Helge Vølstad, Norway, and Mike Armstrong, UK, will meet in ICES HQ, Copenhagen, in 6–9 November 2012, to:

a) On the basis of case studies, examine how national catch sampling programs can be designed and coordinated between countries to meet DCF or other objectives at a regional scale in the most cost-effective way. Develop operational quality assurance indicators for evaluating sampling surveys that can be incorporated in and enhance the WKACCU bias scorecard.

b) Develop guidelines for design-based and model-based data rising and precision estimation, taking account of multi-stage survey design and cluster sampling effects and the need to combine estimates over different sampling programmes within and between countries at a regional or stock level. Consider how national and regional sampling databases could be designed to raise data following best practice.

c) Develop and define quality indicators and levels for onshore and offshore sampling schemes and advise on revisions to the WKACCU score cards to accommodate them.

WKPICS2 will report by 7 December 2012 for the attention of PGCCDBS, RCMs, STECF/SGRN, and ACOM.

1.2 WKPICS2 participants and meeting agenda

The list of participants and the adopted agenda are in Annex 1 and 2, respectively. All the working documents, presentations and national sampling scheme reports are located on the meeting SharePoint site.

1.3 Background to WKPICS2

The data collected from fisheries have a primary function of supporting stock assessments and informing fleet-based management decisions. To this end, the overall aim for a design-based sampling strategy is to:

1. Collect data in a way that accuracy (bias and precision) can be reliably assessed at national and regional level
2. Ensure that sampling intensity is allocated in a way that would maximize precision at the level where it matters most in the context of assessment of stocks and fisheries

The European Commission has advised that the work of WKPICS2 is relevant for the development of the new multi-annual programme of data collection (DC-MAP) in terms of providing definitions/ best-practice for implementing statistically sound catch sampling. In a communication to WKPICS2, DG Mare stated they would be particularly interested in the following outcomes of WKPICS2:

- Specific proposals on a repository with best-practice for implementing statistically sound catch sampling in the DC-MAP, based on design-related indicators and data quality vs. cost indicators.
- Given that the objective of the catch sampling schemes under the DC-MAP will remain, as for the DCF, to provide catch/ (discard data) structured by metier and that the DC-MAP will aim to strengthen regional coordination, what degree of harmonisation of national catch sampling will be required in order to allow regional sampling?

WKPICS2 understood the European Commission's second point concerning objectives of catch sampling schemes to mean that estimates should in principle (and with sufficient sampling resources) be able to provide unbiased estimates by fleet metier, and not necessarily that sampling schemes should adopt metiers as prior sampling strata.

Section 2 of WKPICS2 addresses its Term of Reference (a) and the specific European Commission request by i) considering how these issues could be addressed within a regional sampling programme; ii) developing guidelines for best practice in catch sampling schemes (Annex 3); iii) identifying if and where harmonisation of national sampling schemes would be advantageous; and iv) how sampling effort could be optimised between and within national sampling schemes. Section 3 addresses Term of Reference (b) and gives detailed advice on estimation procedures for different classes of catch sampling schemes. Given the importance of developing the European Regional Data Bases in a way that can accommodate a range of probability-based sampling designs and associated estimation procedures, Sections 3.6 and 3.7 provide a summary of WKPICS2 advice on changes that would need to be made to the data base structure. Section 4 addresses Term of Reference (c) and draws upon a range of proposals for data quality indicators given by previous ICES meetings (WKPRECISE, WKACCU, SGPIDS, WGRFS and PGCCDBS) to develop a system of recording data quality at the national and regional / stock scale that could be of the most practical use for stock assessment and regional coordination.

2 Design and coordination between countries to meet DCF / DC-MAP or other objectives at a regional scale in the most cost-effective way (ToR a)

2.1 Proposed DC-MAP framework for regional sampling programmes

A major change associated with the DC-MAP will be a revision of the roles and work programmes of the current Regional Coordination Meetings (to be re-designated as Regional Coordination Groups; RCGs) as proposed by STECF 12-07 in 2012 (Ebeling et al., 2012). The STECF report proposes that the RCGs would develop regional work plans in which end-user priorities are ranked to ensure work plans operate within (limited) capital and human resources. For example, it would be for the RCGs in close liaison with the end-users to determine whether for a given resource base it was preferable to take fewer samples from more species or vice versa. Assuming that Member States develop statistically-sound schemes for sampling commercial fisheries, regional coordination would revolve around the stock/species-orientated sampling priorities based on regional assessment and advisory needs. A national catch-sampling scheme could be seen as comprising sampling frames and strata within the overall regional sampling activity, but with priorities and sampling levels coordinated at the regional level. STECF 12-07 also considered the possibility of defining appropriate sampling frames and strata that could cross national borders, and also of accommodating nationally important issues that may have a lesser priority in regional terms.

The STECF 12-07 report emphasizes that it is essential that the quality of data is known when it is used for analysis by end-users, because management actions based on poor data should be avoided. However in its report, STECF no longer advocates pre-defined quality targets (e.g. precision levels) and instead proposes a minimum sampling effort, with an interim minimum standard being to maintain the sampling effort by region as specified in the current NP proposals for 2011-2013. If it appears that this would lead to unacceptable quality, there should be provisions to adjust the minimum sampling level in consultation with the end-user.

These proposals by STECF identify a need for: clear documentation and prioritising by end-users of the estimates needed to support regional assessment and advisory needs; implementation of best practice in designing and running statistically-sound sampling schemes; and a need for some degree of optimisation of sampling across countries to achieve the most cost-effective data collection supporting assessments and advice.

The challenges of establishing a coherent regional sampling programme within a stronger system of regional coordination by RCGs was recognised by the ICES Planning Group on Commercial Catches, Discards and Biological Sampling (PGCCDBS), which in 2012 proposed a Studies Contract to "support design-based regional data collection programmes" (ICES, 2012c). The objective of the proposed study is to develop an operational framework for establishing and coordinating design-based sampling programmes at a regional scale for the most cost-effective delivery of fishery and biological data required by the revised DCF and any specific additional needs to support assessment and fishery management. The material provided by WKPICS2 provides essential background for establishing a project of this nature.

The different steps in designing and implementing a regional data collection scheme to meet end-user needs are illustrated in Figure. 2.1. The most critical first stages are for the end users to clearly define the objectives and estimates required at a regional level to support fisheries management or conservation objectives, and an indication of minimum precision needed (steps 2 & 3). For example, for a data-rich stock of high economic or conservation importance, this could be estimates of stock status (biomass and fishing mortality) with sufficient precision to allow detailed estimation of MSY reference points, stock status and TAC forecast options. For a relatively minor, data-limited stock, simpler metrics may be required (e.g. a series of survey abundance indices and fishery landings). The subsequent steps 4 -6 (type of data required; data collection methods and design, sampling intensity and allocation of sampling effort across countries and strata) cannot take place in an effective way without the information specified in step 3. The relationship between the cost of data collection and the precision achieved should be considered when specifying a regional catch sampling programme, although it may take some time to develop the data required for such optimisation and until then a judgment-based approach may be required.

The subsequent steps are the data collection, the handling of the data (e.g. archiving in the regional data bases), evaluation of data quality (quality indicators) and finally the analysis of the data to provide the required estimates and associated measures of uncertainty. The evaluation of sampling schemes against benchmarks for good practice, and the monitoring of data quality using suitable indicators, should be given considerable attention. The data analysis (step 9) may provide additional evidence of problems with data quality that may be traced back to sampling schemes, leading to improvements in sampling design.

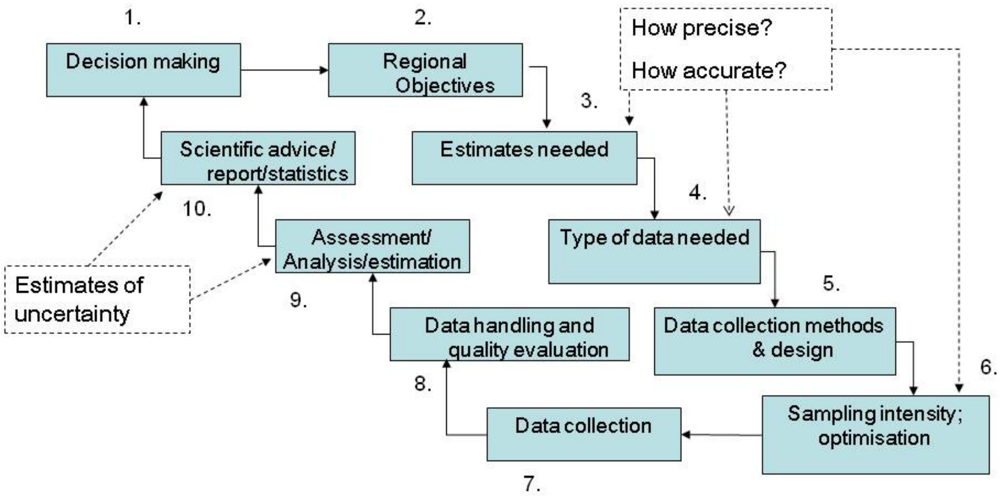

Figure 2.1 Stages in design and implementation of a regional data collection scheme providing data supporting assessments and management advice (adapted from schema provided by Mika Kurkilahti).

2.2 Ensuring best practice in design and implementation of catch sampling schemes

The STECF EWG 12-07 report (Ebeling et al. 2012) advises that, together with the National Programme, it is important that Member States provide a protocol which describes how the proposed sampling programme is designed. This allows for an evaluation of whether the programme is designed respecting guidelines for good practice to avoid bias, and is cost effective. In the national Annual Reports, deviation from the best practice protocol should be described to allow the identification of possible bias in the final estimates. Ideally, all national surveys should clearly document the sampling frame, sample selection procedures, response rates (e.g. refusals to take observers), imputation methods for missing data and weighting procedures employed to derive national estimates.

WKPICS2 has developed guidelines for "best practice" that apply to steps 5 to 8 in the schema shown in Figure 2.1. This covers the design, implementation and analysis stages of catch sampling schemes, assuming steps 3 & 4 are clearly defined.

Best practice can be defined as sampling designs, implementation and data analysis that lead to minimum bias and an accurate estimate of precision, and which make the most efficient use of sampling resources. For example, probability-based sampling with accurate control of the inclusion probabilities would be considered an example of best practice. However, if logistical, legal, and economic constraints dictate the use of a non-probability based scheme to select primary sampling units (for example legal requirements in the selection of a reference fleet), it is good practice if the selection is done in a way that ensures representative coverage of the target population and minimises bias, and if this can be demonstrated with suitable diagnostics. Bad practice would be an ad-hoc, non-probability based sampling scheme, particularly where there are no census data to show how representative the samples are of the population or to re-weight the samples during analysis.

Where bias is unavoidable, best practice requires collection of information that allows the form and level of bias to be investigated, and to develop mitigating measures where possible. For example, by recording all refusals (and the reasons) in an on-board sampling scheme, and the characteristics of those vessels and their activities, there is a better possibility to evaluate the potential for bias.

In the fullest sense, best practice for national catch sampling schemes on shore or at sea encompasses survey design, documentation of objectives, design and protocols, staff training, data collection and archiving, systems for monitoring sampling performance, and data analysis. Some of these aspects would require lengthy documentation, so WKPICS2 has restricted the guidelines mainly to aspects of design in the expectation that good practice for the other aspects of sampling schemes would be demonstrated by the availability of detailed national sampling protocols.

Annex 3 provides draft "best practice" guidelines developed by WKPICS2, which could be included in a repository with best practice for the new DC-MAP.

2.3 Harmonisation of national sampling schemes

Regional sampling schemes will typically involve national sampling schemes with separate sampling frames of vessels or shore-based access points for sampling. The national sampling frames can then be defined as "super-strata" in a regional program. This approach is only sensible if the frequency and volume of foreign landings is a sufficiently large component of the area landings of the country of origin of the vessels to warrant separate foreign-vessel strata that can be sampled with sufficient frequency. The inclusion or exclusion of these trips from the sampling frame for the foreign country should be evaluated through consideration of potential bias in raised length or age compositions for the country". As shown in Section 3, the national sampling frames can be established and stratified in ways that are appropriate for that country. However there does not have to be harmonisation of frames and strata definitions between countries provided the sampling follows best practice as identified in Annex 3, the same data are being provided, and the estimates and variances can be combined across countries to give regional estimates. The main exception is in the definition of domains of interest, such as metier-groups, which must be identical between countries to allow estimation for these domains at the regional scale (see Section 3).A particular issue for harmonisation is when vessels from one country land into the ports of another country. This is currently handled within the DCF by establishing bilateral agreements for sampling. The bilateral agreement requires the country of landing to carry out sampling and transmit the data to the country the vessels originate from. In practice, this has not always happened, or the data are collected in a way that is not compatible with the sampling schemes or data analysis methods in the native country and are therefore not used. Methods to address this complex issue should be further discussed in WKPICS3.

2.4 Optimisation of sampling across and within national schemes

2.4.1 Stratification and sampling effort allocation

It is recommended that each national catch sampling scheme limits the number of strata to ensure sufficient sample sizes, taking into account the total sample sizes that likely will be achieved at the primary sampling unit level. For multiple objectives, it is desirable to allocate sampling sizes to strata in a manner that approximately is proportional to expected effort or total catch, using historic data to guide the allocation. Over-stratified sampling schemes should be avoided, as this can lead to zero observations for some strata due to inadequate resources for sampling, resulting in a need for imputation at the analysis stage and potential bias. Cochran (1977, page 133) suggests that there will usually be little reduction in variance by employing more than 6 strata although this will depend on individual circumstances and available sampling effort. If the between-PSU variability by stratum can be estimated from previous sampling years, and the differences between strata is fairly stable over time, it is relatively straightforward to determine how a given amount of sampling effort regionally and nationally should be distributed across strata to minimise the overall variance of estimates

In multipurpose surveys it is recommended that the stratification scheme represents a compromise (Kish and Anderson, 1978). Miller et al. (2007) present analytical results for the sampling fractions and sample sizes for primary units within each stratum of a stratified sampling design employed for the North Pacific Groundfish Observer Programme, that are optimal with respect to a weighted sum of relative variances for multiple estimation objectives.

Once a practical and efficient stratification of the primary sampling units for estimating the characteristics of a fishery (or fisheries) is developed, then the next step is to determine the appropriate allocation of sampling effort to each stratum. In general, the optimum sampling allocation that minimizes the stratified estimate of the mean, $\overline{y}_{st}$, for a fixed total sample size, n, is given by (Cochran, 1977, page 98):

$$n_h^{opt} = n \frac{N_h S_h}{\sum N_h S_h},$$

where N_h is the number of primary sampling units in stratum h and S_h is the expected (i.e. "true") standard deviation for stratum h. That is, more effort is allocated to those strata that are larger (N_h) and/or are more variable (S_h). In addition, when cost is taken into account then more samples should usually be collected in strata where sampling is cheaper (Cochran, 1977).

There are two drawbacks for using this optimization procedure. The first is that this is the optimum allocation for one particular parameter of interest (e.g., average discards per fishing trip). For surveys with multiple objectives, an allocation that is optimum for one objective may be far from optimum for another. The other shortcoming is that the true value of the standard deviation for each stratum is rarely, if ever for fisheries data, known and must be estimated.

A more practical sampling allocation that often performs well for surveys with multiple objectives is to allocate effort proportional to stratum size, or:

$$n_h(prop) = n\frac{N_h}{\sum N_h}.$$

Another advantage of proportional sampling is that it is self-weighting, which is often a useful feature when analysing survey data.

In terms of relative sampling precision, it can be shown (Cochran, 1977, page 99) that:

$$Var_{opt}(\overline{x}_{st}) \leq Var_{prop}(\overline{x}_{st}) \leq Var_{ran}(\overline{x}).$$

That is, proportional sampling is always more, or just as precise as, random sampling (i.e. random sampling with no stratification), and that the stratified estimate based on the optimum allocation is always more, or just as precise as, proportional sampling.

When information is available on sampling variability, then a sampling scheme can be evolved to improve precision and survey efficiency. For example, the sources of variability for surveying the length of fish in the commercial catch of some pelagic species based on a Norwegian self-sampling reference fleet were isolated using a variance component analysis technique (Box et.al, 1978). The results are shown in Table 2.1 (from Pennington and Helle, 2011). Based on these estimates, different sampling scenarios were assessed (Table 2.2).

Table 2.1 (reproduced from Pennington and Helle, 2011).

Table 1. The estimated variance components for fish length in the commercial catch based on self-sampling for three pelagic stocks in 2008, and the percentage of the total variance associated with each component.

Stock	Boat component	Sample-day component	Within-sample component
Mackerel	0.24 (1.8%)	5.34 (39.2%)	8.03 (59.0%)
North Sea herring	0.53 (22.2%)	0.72 (30.5%)	1.13 (47.3%)
Norwegian herring	0.24 (5.2%)	0.91 (19.6%)	3.48 (75.2%)

Table 2.2 (reproduced from Pennington and Helle, 2011).

Norwegian self-sampling fleet 1767

Table 3. Summary statistics for various hypothetical sampling schemes and associated sources of variance for estimating the mean length of the commercial catch of Norwegian spring-spawning herring, with the first row approximating the prescribed (i.e. balanced) sampling protocol for 2008.

Sampling scheme				Sources of variance			Estimation precision		
Number of boats	Number of sample days	Fish per sample	Total measured	Boat component	Day component	Within component	Var($\hat{\mu}$)	s.e.($\hat{\mu}$)	Effective sample size
8	19	114	17 328	0.030	0.006	0.000	0.036	0.190	127
8	18	60	8 640	0.030	0.006	0.000	0.037	0.192	125
8	17	40	5 440	0.030	0.007	0.001	0.037	0.193	123
8	16	20	2 560	0.030	0.007	0.001	0.039	0.196	119
12	15	10	1 800	0.020	0.005	0.002	0.027	0.164	170
15	10	10	1 500	0.016	0.006	0.002	0.024	0.156	188
20	10	5	1 000	0.012	0.005	0.003	0.020	0.142	229

This analysis demonstrates that when there is variability among boats, or among sampling days (Table 2.1), little is gained by measuring more fish per vessel or sampling day. In particular, the first line in Table 2.2, which was the actual sampling scheme and intensity in 2008, shows that 17 328 herring were measured, the standard error of the estimated mean length was 0.190 and the effect sample size was 127 fish. While, if 20 boats measured a total of 1000 herring (last line in Table 2.2), then the standard error would only have been 0.142 with effective sample size equal to 229. Again, this demonstrates that it is generally better to sample a few fish from as many primary sampling units as possible. More details on using a variance component analyses for choosing a sampling strategy are in Helle and Pennington (2004) and Pennington and Helle (2011).

2.4.2 Effective sampling strategy for estimating biological characteristics

Significant cost-savings can often be achieved by sampling fewer fish from each PSU, with marginal loss in precision. Fish sampled during a survey are not a random sample of individual fish from the entire commercial catch but a sample of *n* clusters, one cluster from each, e.g., fishing trip, a single catch, a port sample, etc. Since fish caught together are usually more similar than those in the general population, a total of *M* fish collected from *n* clusters will contain less information about the distribution of the variable of interest for the entire population than if *M* fish were randomly sampled from the population – which is impossible to do in practice. One way to measure the information contained in a sample of, for example length or age measurements, is to estimate the number of fish that one would need to sample at random (the effective sample size) to obtain the same information about the variable contained in the cluster samples (Kish, 1965; Skinner et al., 1989). It should be noted that if the effective sample size is small, then this implies that the estimate of the entire distribution is rather imprecise. For details on calculating the effective sample size for marine cluster sampling see, e.g., Pennington and Vølstad (1994), Folmer and Pennington (2000) and Pennington et al. (2002); Chih, (2010).

The effective sample size is a much more informative number about the amount of information contained in a sample than is the number of fish that are measured or aged. Table 2.3 (Pennington et al, 2002) shows the sampling efficiency for estimating the mean length of cod based on data from a survey in the Barents Sea. For example, 46 593 fish were measured in 1999, while the effective sample size was 211. This relatively small effective size is reflected in the estimated variance of the mean, which is rather large given the number of fish that were measured.

Table 2.3 (reproduced from Pennington et al. 2002)

76 Fishery Bulletin 100(1)

Table 1

Summary statistics for assessing the precision of the estimated length distributions of Northeast Arctic cod based on the winter and summer bottom trawl surveys in the Barents Sea. The estimated effective sample size is denoted by $\hat{m}_{eff}$, n is the number of stations at which cod were caught, M is the total number of cod caught, m is the number measured, $\hat{R}$ is the estimate of mean length, and var($\hat{R}$) is its variance.

Year	n	M	m	$\hat{R}$(cm)	var($\hat{R}$)	$\hat{m}_{eff}$	$\hat{m}_{eff}/n$	$(\hat{m}_{eff}/m) \times 100\%$
Winter								
1995	296	175006	47286	20.0	0.7	313	1.1	0.7
1996	314	209114	44021	18.0	0.3	511	1.6	1.1
1997	177	71418	25689	19.0	2.1	119	0.7	0.7
1998	197	60746	32536	22.1	0.7	394	2.0	1.2
1999	223	50192	21760	25.0	1.9	107	0.5	0.5
	Avg.	113295	34258			289	1.2	0.8
Summer								
1995	329	66643	46161	31.2	1.4	252	0.8	0.6
1996	341	115834	45286	24.4	0.6	478	1.4	1.1
1997	266	72093	26947	23.1	0.8	266	1.0	1.0
1998	218	72360	23461	25.1	1.1	184	0.8	0.8
1999	217	46593	23253	30.8	0.9	211	0.9	0.9
	Avg.	74705	33022			278	1.0	0.9

Figure 2.2 (Pennington et al 2002) shows the typical outcome of reducing the number of cod measured at each station. The 95% confidence limits for each length class based on the entire sample are rather wide (right panel), which demonstrates that a small effective sample size implies that the estimate of the entire population distribution is rather imprecise. In addition, as shown in Figure 2.2, the length of the 95% confidence intervals decrease only marginally if the number of cod measured is reduced from 21 769 to 2597.

Pennington et al.: Assessing the precision of frequency distributions from trawl-survey samples 79

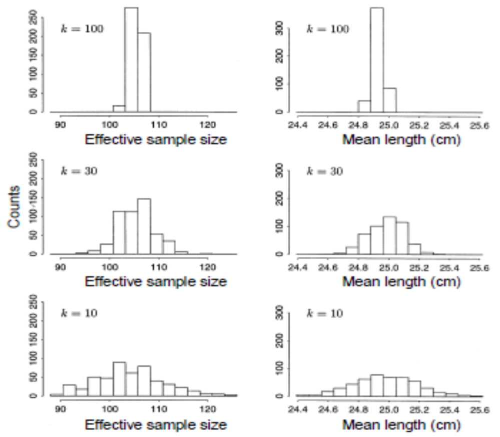

Figure 2

Simulated estimates of the distribution of the effective sample size, $\hat{m}_{eff}$, and of the mean length, $\hat{R}$, when the total number of fish measured is reduced for the 1999 winter survey in the Barents Sea. The top panel is when a maximum of k = 100 fish are selected per subsample for a total of m = 17,615 fish in each run; the middle panel, k = 30, m = 7240; and the bottom panel, k = 10, m = 2597. The estimate of the population mean based on the entire sample (m=21,769) is 24.96 and its 95% confidence interval is (22.26, 27.66).

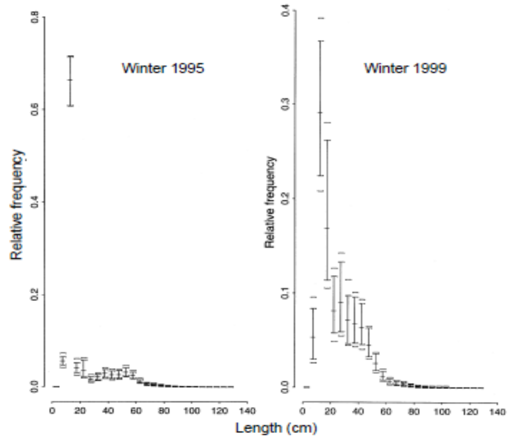

Figure 3

Bootstrapped estimates of the 95% confidence intervals for the proportion of cod in the Barents Sea in each 5-cm length bin, for winter 1995 and for winter 1999. The inner brackets denote the confidence intervals if the estimates are based on all the cod measured during the surveys and the outer brackets denote the confidence intervals if 10 fish are measured for each subsample.

Figure 2.2. (reproduced from Pennington et al, 2002).

The next example (Figure 2.3) shows the effect of reducing the number of Northeast Arctic cod that were aged from each sampled commercial fishing trip (from Aanes and Pennington, 2003).

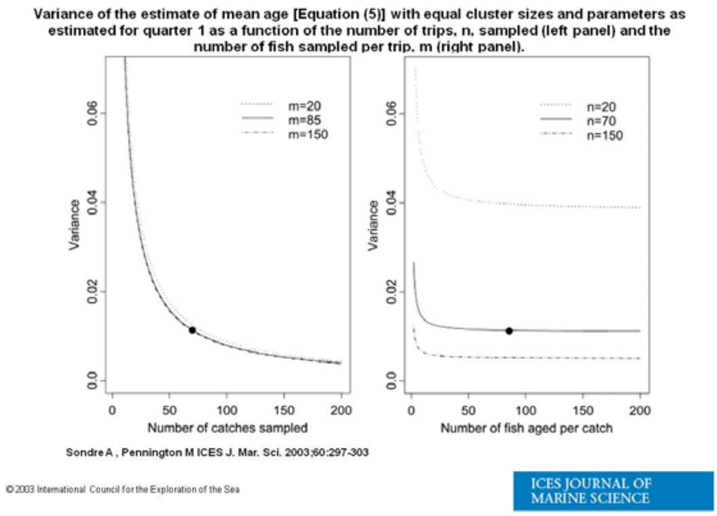

Figure 2.3. (reproduced from Aanes and Pennington, 2003). The dots correspond to the sampling intensity in quarter 1 (n=70; m=85).

The data are from the Norwegian port-sampling scheme ("Amigo") for estimating the age distribution of the commercial catch of Northeast Arctic cod. As can be seen little is gained from aging more than about 20 cod from each catch, and the only way to increase precision is to increase the number of catches sampled. As shown, such a reduction in the number of fish sampled within each cluster would only marginally decrease the precision of the estimates. The time saved by collecting and reading fewer otoliths is significant. For example, a reduction from about 85 fish to 20 from each catch sampled in 2000 would have reduced the number of fish sampled from 11 000 to 2500, which would decrease reading and handling costs by 150 work-days (Aanes and Pennington, 2003). Part of the time saved could be used to collect samples from more landings, which is often the only practical way to improve the precision of surveys of commercial catches.

The next example from the Greenland trawl survey of the shrimp stock off west Greenland (Folmer and Pennington, 2000) again shows the effect of positive intra-cluster correlation on estimates of length-composition of shrimp. In a small part of the survey area, 7 stations were sampled and a total of 5341shrimp were measured (approximately 763 per station). The 7 stations were randomly divided into two groups, and separate length-distributions were computed for the full dataset, and for the random half-splits of stations. . The three estimates of the length distribution very different even through each distribution were based on measurement of several thousand shrimps. The reason for the apparently different estimates is a low effective sample size: even though 5341 shrimp were measured for all 7 stations, the estimated effective sample size is only 24 shrimp (Folmer and Pennington, 2000).

The last example is from Iceland. The Icelandic sampling protocol for sampling commercial catches is to choose a sample of 150 fish, of which 50 are aged at random. As always, the preparation and ageing of otoliths is time consuming and expensive so it is of interest to explore the effects of reducing number of aged otoliths on the precision of estimates of catch in numbers, mean weight and mean length.

To explore this, haddock samples from the longline fishery were used since they are an actual cell in the calculation of catch in numbers for haddock in Va. In total 48 haddock samples were collected from longlines in 2011. Three scenarios were explored, using 50, 35 or 20 otoliths from a sample to construct age length keys for estimating catch in numbers. All the 150 length measurements were used from each sample. In the three scenarios the same total catch and length-weight relationship was used.

To measure the precision of the estimated catch at age, the Relative Standard Error [RSE $= s.e.(\bar{x}) / \bar{x}$] was estimated via a bootstrapping approach, which conserved the effect of clustering sampling. In short, 100 bootstrap datasets were constructed by sampling with replacement from a pool of all the 48 samples. For all the 100 bootstrap datasets, catch in numbers was calculated in the same way as described above and based on bootstrapping estimates of RSE were generated.

The results are presented in Figure 2.4. As expected, the effect of reducing the number of aged otoliths in each sample from 50 to 20 does not affect the estimates of mean catch in weight or age. What is important with regards to sampling strategy is that there was only a marginal increase in the bootstrap estimates of the RSE for the most important age groups in the catches: the RSE is low in all three scenarios, around 0.1 (age 4 to 8). For the youngest and oldest fish, which form a relative small proportion of the total catch, the difference in RSE is somewhat greater among the different scenarios, though it should be noted that for these two age groups, all the RSEs are relatively large for all sampling intensities.

Given these results, it would not have a marked effect on the precision of estimates of catch in numbers, mean weight and length if the number of otoliths in each sample is reduced from 50 to 20. The resources saved would be better used to increase survey efficiency and effectiveness by collecting smaller samples from more catches.

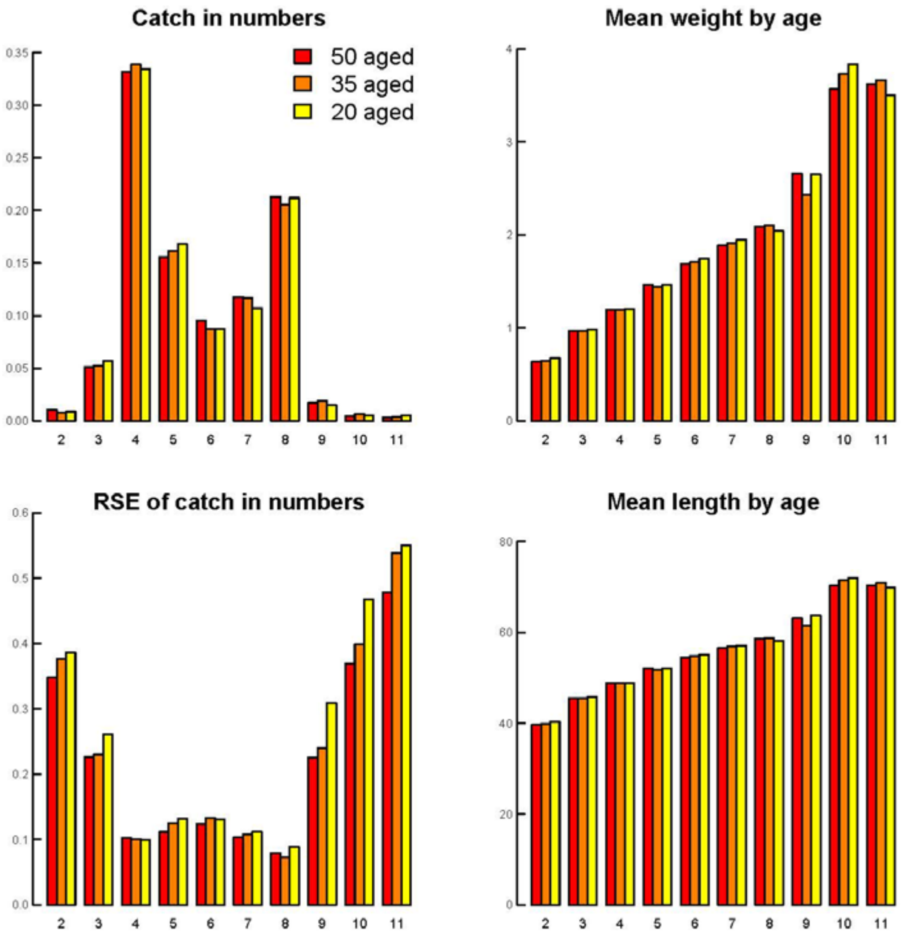

Figure 2.4. Haddock in ICES Va. Estimates of catch in numbers, mean weight and length and bootstrap estimates of RSE of catch in numbers using age-length keys based on 20, 35 or 50 otoliths in each sample. The estimates were based on 48 samples collected from the Icelandic longline fishery in 2011.

In summary, it appears to be generally true that for marine fishery sampling it is best to sample a few fish from as many locations as possible. Conversely, it is usually a waste of time to sample too many fish at one location or from one catch, etc. This is also demonstrated for at-sea biological sampling of length and otoliths of groundfish in the very extensive North Pacific Groundfish Observer Program (NPGOP). In 1999, the NPGOP implemented new sampling protocols for the selection of fish for biological samples following recommendations in Vølstad et al. (1997). By reducing the number of fish sampled per fishing unit (trawl haul-back, long-line, or pot), and increasing the number of fishing units sampled, the NPGOP achieved the same or increased precision in estimates of length- and age-distributions, while reducing the overall number of fish measured (Barbeaux et al. 2005).

2.5 Monitoring of how national sampling schemes evolve over time

A questionnaire on the national sampling design and data analysis methods used was circulated prior to the meeting. The responses are collated in Appendix 4 and the main findings are summarised here:

- The majority of sampling programmes in ICES are still opportunistic although some nations have implemented probability-based sampling.
- Sampling effort is generally allocated in proportion to the previous year's fishing effort or landings.
- Target precision levels are generally based on DCF targets.
- Records of refusal rates are the most common quality indicator for both at-sea and onshore sampling.
- The main purpose of the sampling programmes is to provide data for assessment working groups.
- In some sampling programmes, the weighting procedures do not follow the sampling design correctly.
- Nearly all sampling programmes provide precision estimates, most commonly estimated analytically or using bootstrap.
- Most respondents stated that post-stratification takes place if sample sizes are low. (Note that the 'correct' definition of post stratification is to apply weights based on an auxiliary variable; if by chance you sample too many beam trawl trips, you can down-weight these samples based on the total number of beam trawl trips in the logbooks).
- COST tools are commonly used but other R-code, SAS, SQL, Access and Excel are also in wide use.

A number of new questions were added during the workshop and participants were invited to fill these in as well. It is hoped that the questionnaire can be used as a tool for RCGs to provide overviews of sampling schemes and to monitor improvements in schemes over time.

3 Guidelines for design-based and model-based data raising and precision estimation (ToR b)

3.1 Principal classes of survey designs for catch-sampling programs

Fisheries catch sampling schemes considered here can broadly be categorized into four principal classes based on the number of stages in the sample selection. For at-sea sampling programs, the sampling frame is ideally constructed so that vessels, trips, and fishing operations can be selected with known probability over time. The effective sample size can be maximized by spreading out the collection of data across all vessels, trips, and fishing operations in each stratum. For at-sea sampling the two principal design classes are:

A) Trips as primary sampling units. When trips can be selected randomly from a fleet of vessels, at least approximately, it is often reasonable to treat vessel-trips as the primary sampling units. For a fleet with day-trips this can easily be achieved by randomizing the selection of days and vessels. In such cases, it is reasonable in the analysis phase to treat the list of all trips (obtained at the end of the year) as the sampling frame. This is a virtual frame that cannot be used in stage 1 to select the trips. The actual selection is typically based on a frame with a vessel list crossed with time. For fleets with varying trip-length it is more difficult to selected vessels and trips with approximately equal inclusion probabilities. It can be helpful to create strata where vessels with a similar trip length are grouped.

B) Vessels as primary sampling units. When it is not possible to approximately achieve a random sample of trips for a fleet, then another design option for at-sea sampling is to select vessels randomly in stage 1, and then select a sub-sample of trips throughout the year for each vessel. In this case, the vessel is the PSU, with trips as second stage sampling units and fishing operations as third stage sampling units. This design introduce an extra level of clustering, since trips and fishing operations to be sampled now are nested with a fixed number of vessels selected in stage 1. Clearly, these trips may not be considered a simple random sample from the entire fleet.

For at-shore sampling, a common approach is to conduct the sampling of catches from vessels and trips that can be accessed in ports where they land their catches. In these cases, the sampling frame is based on a list of access-sites crossed with time (for example port-days). The two principal design options for on-shore sampling are:

C) Site-days as primary sampling units. Where the primary sampling units can be defined as site-days which can be randomly selected, there is one extra level of clustering, where site-days are selected in stage 1, trips in stage 2, boxes in stage 3 (for sorted catches), and fish in stage 4.

D) Sites/ports as primary sampling units. Another design option is to select a sample of sites/ports (PSUs) in the first stage, and then conduct catch sampling for a subsample of days (stage 2) days within each site/port selected i stage 1. In stage 3, catch sampling is conducted for a sample of trips on a selected day and port. If landings are sorted by market categories and packed into boxes, then a stratified random sample of boxes (stage 4) may be taken for each trip. This option may be cost-effective if ports are scattered over large areas, and field samplers near a selected number of ports can be recruited.

We summarize the four primary classes of catch sampling schemes in Table 3.1.

Table 3.1. Design options within a nation, with examples of sampling units and stratification (STR) for multiple stages. The level of clustering increases from scheme A to D.

Design	Stage 1	STR 1	Stage 2	STR 2	Stage 3	Stage 4	Sampling frame
A	Trip	Fleet, vessel characteristics, quarter	Fishing operations	Statistical area	Fish		Vessels*time
B	Vessel*Time	Fleet, vessel characteristics, quarter	Trip or fishing operation	Statistical area	Fishing operation	Fish	Vessels*time
C	Site*Time	Geographic, quarter	Trip	Species and size category	Box	Fish	Sites * time
D	Site	Geographic	Time	Week	Day/trips	Box	Sites * time

It should be noted that stratification can be employed for several stages of the selection. When sampling port-days (stage 1) in options C and D, it is common that the catches for many completed trips (stage 2) are sorted by market category before they are landed. The market categories will then form strata. Sub-sampling of trip-catches may then be conducted by selecting a random sample of boxes (stage 3) from each market category (stratum), and then measure all fish from each box, or a sub-sample of fish from each box in stage 4.

In Figure 3.1 we show an example of sampling frames based on list of vessels (a) and ports (b). It is apparent in example (a) that it is not possible to select the trips directly since they are only know completely at the end of the year. However, it may be possible to select vessels and trips randomly throughout the year so that the probability of sampling trips is approximately equal within strata. If so, it can be assumed that the PSU is a trip, and the data can be analysed according to sampling scheme A. In example (b) the scheduling of port visits is conducted by stratified sampling of site-days. This is an example of design class C.

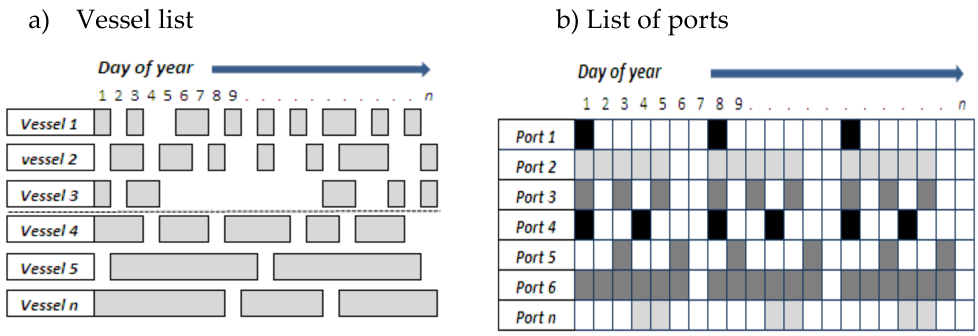

Figure 3.1. Structuring of the target population (catches) in space and time for (a) a list of vessels with trips of varying duration, and (b) a list of ports with varying frequency of days when catches are landed and can be accessed (degree of shading in boxes indicates "size" of port – e.g. number of vessel trips or size of catch landed per day.

3.2 Examples of catch sampling schemes by principal design class

3.2.1 At-sea sampling, design class A

The Observer programme in Kattegat

This case study presented in WKPICS1 tested a new sampling scheme involving stratified random sampling of vessels/trips. The sampling frame is a list of vessels, and the individual vessel-trips are treated as primary sampling units. The vessels/trips are stratified by gear group. Random selection of vessels and trips are done using a "lucky wheel", where probability of selecting a vessel is weighted by number of trips in previous year. The refusals to take observers and reasons for refusals are documented. The goal of the new catch sampling scheme is to reduce bias through representative sampling of trips from the fleet, increase precision by sampling more vessels/trips, and improve cost effectiveness.

The Observer program in Sweden

Observers collect data on discard from Swedish demersal fisheries in the Skagerrak (WKPICS1). The fleet can be grouped into three main fisheries based on the mesh size of the nets: small-mesh (35mm+) Pandalus fishery; Nephrops fishery using 70 mm mesh with sorting grid; other demersal trawlers using 90mm+ mesh. The fleet consists of 140 vessels, with typical trip length of 1-5 days. Vessels participate is several fisheries. The Swedish programme to estimate discards involves 50 observer trips per year, covering 5 fisheries. A separate vessel list frame is defined for each fishery, with the PSUs being the individual fishing trips in the fishery. The number of trips to be sampled for each fishery (frame) is based on the fishing effort in the previous year (days at sea for the contributing vessels). A randomised vessel list is produced for each frame, with probability proportional to size (days at sea in previous year), and vessels contacted in order as they appear on the ranked list. Refusals to take observers and reasons for refusals are tracked.

3.2.2 At-sea sampling, design class B

The Norwegian reference fleet

The off-shore and coastal reference fleets in Norway (WKPICS1) is an example of at-sea sampling where vessels are the primary sampling units. The sampling frame is based on a list of vessels in the off-shore and coastal fleets. The selection of the reference fleet is conducted through a tender process with the aim to approximate stratified random sampling. When multiple vessels satisfy all criteria asked for in the tender process, the selection of vessels is based on a random draw. The PSUs are the individual vessels, and individual trips are the secondary sampling units (SSUs).For the coastal fleet, the sampling frame is restricted to vessels from 9.5 to 15 m length that fish with gillnets. The vessel in the coastal fleet are stratified by nine statistical areas based on home ports, and is restricted to the predominant gear types. For the high-seas Reference Fleet, vessels are stratified into demersal and pelagic vessels. The demersal RF is also stratified by gear and fishing pattern (e.g. North Sea versus Barents Sea).

For 2013 the reference fleet will consist of 20 off-shore vessels and 20 coastal vessels. For each vessel in the reference fleet, trained fishermen collects data on catch composition and age-length samples for fishing operations systematically throughout each trip. Crew members onboard the RF vessels are trained to conduct self-sampling following IMR's protocols and are required to record their catch logbooks electronically. On trawlers, Danish seiners and purse seiners they are required to make a complete record (including discards) of each catch. The crew of longliners and gillnetters provide such data for one representative fishing operation per day and in addition record the total catch per day.

3.2.3 On-shore sampling, design class C

The Scottish on-shore catch sampling scheme for demersal species

The Scottish on-shore catch sampling scheme for demersal species (WKPICS1) covers five ports that accounted for 76% of trips and 78% of annual demersal landings (in weight) into Scotland in 2010. The sampling frame is a matrix of 5 ports (sites) crossed with days of the year, with site-days being the primary sampling units. The decision of which ports to include was based on practical and budgetary considerations. The frame excludes remote, minor ports and the major port that presented sampling difficulties, but allocation of resources to five major ports accounting for almost 80% of landings is considered an appropriate trade-off between under-coverage bias and improved precision by concentrating efforts on major ports rather than wasting resources travelling to minor, very remote ports. The survey employed stratified random sampling, where the 5 ports are grouped into four geographic strata which are sampled with different sampling rates (port-visits per month). A "lattice" sampling design is adopted for the three mainland strata, with each stratum being sampled at least once within a standard sampling week. The selection of sampling day within a week is dictated by days of the week when there are markets, taking into account the time of market and other practical considerations. In particular, to avoid a fruitless 300 mile round trip, the West Coast stratum is sampled on either the Monday or the Thursday when the main weekly markets occur. The Shetland stratum is covered by a dedicated team based in the Shetland Islands sampling in 36 weeks of the year.

Within a site-day visit, the sampling is conducted by a team leader and a clerk. The vessels are randomly selected based on alphabetized list of all vessels that are landing on the day. This ensures that trips within metiers will get sampled in proportion to effort by metier in each stratum. Five refusals were documented out of 388 sampled vessels in 2011. Selection of fish species to sample during a port visit is largely left to the discretion of the sampling team leader, with emphasis placed on the main commercial species. Biological sampling includes length-stratified sampling for age.

The Icelandic on-shore catch sampling program

The Icelandic catch sampling program (WKPICS1) is a version of design class C, with a sampling frame based on a finite list of landing sites, stratified into regions. The sampling program is based on real-time monitoring of commercial catches at all landings sites, and employs the daily landings statistics available from the Directorate of Fisheries for scheduling site-visits in each stratum. For each species, fleet and regional strata a landings interval target is pre-specified (usually in tonnes). Once the cumulative daily landings value pass the target value an automatic request is made to the sampling team for a sample to be taken for the species/fleet/region in question. The system as such should thus take into account seasonal variability in the landings of any species.

The sampling design is not per se linked to the geographical distribution of the fisheries. However the fishing location of the fish measured at harbour is known with reasonably accuracy, because fishing date is registered for each fish boxes and can hence be linked to geographic location of the fishing at that date, based on the captain's log-book record.

Once a sample request is received by the sampling teams they will contact fish markets or fish processors in their respective area. In some cases boats that land their catches in the region are contacted and asked to set aside fish for the sampling program. Normally the fish is bought by the sampling team and re-sold at a reduced price after measuring.

3.3 Use of sample weights in design-based estimation

The statistical principles for the use of sample weights for estimation of population characteristics from sample survey data is well established (Horvitz and Thompson, 1952). However, the issue of weighting, and the importance of accounting for survey design more generally when computing estimates of catch composition, is not yet fully adopted by the ICES research community. It is important to be aware that because fisheries-dependent survey data never arise from an 'equal probability of selection method' (referred to as EPSEM in the statistical literature; e.g. Lehtonen and Pahkinen 1996) for the individual fish in the sample, then the sample of fish will provide a biased representation of the total catch unless adequate correction is made in subsequent analysis. This is usually done by weighting to correct for the non-equal probability of selection of fish. Since some types of vessels and trips are more likely to be in the achieved sample than others, it follows that the achieved sample of fish will only be representative of the target population of the survey (for example national catch for a fleet) once the data are weighted.

Multi-stage stratified cluster designs are used in catch sampling, resulting in differential probabilities of selection for the individual fish measured for age and length. Consequently, each fish that is measured does not necessarily represent the same number of fish in the target population, as would be the case if a simple random sampling approach could be employed. The sample weight for an individual fish in the catch sample can be interpreted as the number of fish in the target population (e.g., total landed catch for a species and area) that is represented by the fish measured. For probability-based catch sampling schemes, the inverse of the sampling inclusion probabilities will define the base weights. The base weights are computed from the inclusion probabilities, accounting for the multi-stage random selection that led to the sample of individual specimens of fish measured for length and age.

In practice, model-assisted estimation (Særndal et al. 1992) is often used to estimate biological characteristics of total landings. Auxiliary data such as end-of-year census information on landings by gear type, fleet-segments, and ports, is typically used to adjust the base sample weights. However, no universally protocol exists for computing weights to apply for computing regional estimates of catch characteristics (e.g., number at age). The reason is that the computation of sample weights depends on variation in survey designs, the quality of documentation of the sample selection, and the type of ancillary information that is available to adjust for imbalance in the sample due to differential frame coverage (for example as a result of quota sampling to achieve target precision levels for metiers), missing data, and non-response.

3.3.1 Estimation of population totals using sample weights

We start with some general results from standard sampling theory, using notation similar to Thompson (1992). The Horvitz-Thompson estimator introduced by Horvitz & Thompson (1952) is given in many sampling text books (*e.g.* Thompson (1992), p49) and can be described as follows. A general estimator $\hat{\tau}_\pi$ of a population total τ for the variable of interest y for any design with probability π_i of including unit i in a sample of size n distinct units is

$$\hat{\tau}_\pi = \sum_{i=1}^{n} \frac{y_i}{\pi_i} \qquad (1)$$

where y_i is the measured value from unit i for $i = 1 \ldots n$.

The variance of $\hat{\tau}_\pi$ is given by

$$\mathrm{V}(\hat{\tau}_\pi) = \sum_{i=1}^{n} \left(\frac{1-\pi_i}{\pi_i} \right) y_i^2 + \sum_{i=1}^{n} \sum_{j \neq i} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) y_i y_j$$

and an unbiased estimator of this variance is

$$\hat{\mathrm{V}}(\hat{\tau}_\pi) = \sum_{i=1}^{n} \left(\frac{1-\pi_i}{\pi_i^2} \right) y_i^2 + 2 \sum_{i=1}^{n} \sum_{j > i} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) \frac{y_i y_j}{\pi_{ij}}$$

if all joint inclusion probabilities π_{ij} are greater than zero. However, this estimator can give negative estimates and so Thompson suggests:

$$\hat{\mathrm{V}}'(\hat{\tau}_\pi) = \left(\frac{N-n}{N} \right) \frac{s_t^2}{n},$$

where N are the total units in the population, n the units in the sample and s_t^2 the estimate of the sample standard deviation:

$$s_t^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{n y_i}{\pi_i} - \hat{\tau}_\pi \right)^2$$

Now suppose that the variable of interest y has a linear relationship through the origin with an auxiliary variable x so that the use of a ratio estimator would be appropriate. Using the notation above, a generalised ratio estimator $\hat{\tau}_G$ for the population total (*e.g.* Thompson, 1992, p67) is:

$$\hat{\tau}_G = \frac{\hat{\tau}_y}{\hat{\tau}_x}\tau_x \text{ where } \hat{\tau}_y = \sum_{i=1}^{n}\frac{y_i}{\pi_i} \text{ and } \hat{\tau}_x = \sum_{i=1}^{n}\frac{x_i}{\pi_i}\text{, so that}$$

$$\hat{\tau}_G = \frac{\sum_{i=1}^{n}\frac{y_i}{\pi_i}}{\sum_{i=1}^{n}\frac{x_i}{\pi_i}}\tau_x \qquad (2)$$

It is common (*e.g.* Korn and Graubard, 1999) to introduce sampling weights, ω_i, where

$$\omega_i = \frac{1}{\pi_i}$$

so that equation (1) becomes

$$\hat{\tau}_\pi = \sum_{i=1}^{n}\omega_i y_i \qquad (3)$$

and equation (2) becomes

$$\hat{\tau}_G = \tau_x\frac{\sum_{i=1}^{n}\omega_i y_i}{\sum_{i=1}^{n}\omega_i x_i}. \qquad (4)$$

These equations can form the basis of estimation of population totals for almost any sampling scheme. They are particularly elegant because the sampling design is taken into account in the calculation of sampling weights by means of the inclusion probabilities and the summation is then over all sampling units.

However, frequently we estimate for a group of y-variables, namely numbers-at-age, and these are correlated, so although point estimates can be calculated by applying equations (3) and (4) to each age in turn, **the variance estimators given above will not be appropriate**, instead a covariance matrix must be estimated simultaneously for all ages. The analytical equations for such covariance matrices are often intractable, and estimation is often carried out by methods such as bootstrapping. It is thought that the estimation methods provided by the R package "survey" (Lumley, 2012) may correctly calculate covariance matrices for numbers-at-age but this has not yet been confirmed.

A useful overview of sampling weights is presented in the GATS Sample Weights Manual (Global Adult Tobacco Survey Collaborative Group, 2010).

3.3.2 Estimation for sub-groups or "domains" of interest

"For many surveys, estimates of some quantity are needed for different subgroups of the population that was sampled. These subgroups are often referred to as "domains of study" or "domains of interest." For example, in the EU Data Collection Framework, there is a specific requirement to collect data to allow estimation of effort, landings, discard volumes and size compositions for fleet métiers defined by fishing ground, gear type, mesh band and target species. A simpler example might be to make estimates for all vessels for a specific fishing ground, or for all the vessels of a country in a specified fleet sector such as beam trawlers. These are all examples of domains of interest.

In some cases where it is possible to control the sampling probabilities for a domain, pre-stratification of the sampling frame by domain may be optimal. More commonly, as in the case of EU fleet métiers, sampling strata contain vessels and trips that cover two or more domains of interest. In some cases, end-users may request estimates for domains that were not even considered when designing the sampling survey, and only known afterwards. These cases all require the set of samples by domain be weighted to correctly account for inclusion probabilities in the original strata. The re-weighting may be based on post-stratification, or other estimation methods for domains, where census data or other suitable auxiliary data are used to adjust the sample weights. The main issues to be addressed are minimizing bias by ensuring correct re-weighting, and having a sufficiently large sample size to ensure that all domains of interest are adequately sampled.

In the simplest case where a simple random sampling scheme is applied across all vessels or trips without any stratification of the PSUs, the domains will be represented in the samples in direct proportion to the fleet as a whole. The procedure for generating the estimate for that domain is rather straightforward (Cochran, 1977, page 248), involving calculating the estimates only over sampled trips in that domain and giving all other samples a weighting of zero.

If the domain of interest crosses stratum boundaries, for example if long-line vessels fishing in a particular fishing ground of interest have come from two or more port strata that may have different sampling rates, the estimation procedure is more complicated but manageable (see Cochran, 1977, page142). In this case the sampling (inclusion) probabilities for vessels in each port stratum have to be accounted for to avoid bias. Two situations may be encountered: firstly, census data are available allowing the inclusion probabilities in each stratum to be determined after the event; and secondly, there are no census data and inclusion probabilities are only known from the sampling design (e.g. intended fraction of PSUs to be sampled) and cannot be re-weighted to reflect the true probabilities. Where the PSUs can be assigned to the proper strata after the event, and since the sizes of the strata are known, estimates can be produced based on a sound statistical basis (Cochran, 1977, 134). As discussed in Section XX, the differences in inclusion probabilities between the original sampling design and the true values, caused by variability in fleet behavior, require an increased sampling intensity to ensure that individual domain estimates achieve the required precision. This continues to be a major concern for the EU fleet métier based sampling requirements where many individual métiers are specified.

If the sampling design is ad-hoc and the inclusion probabilities are not known either before or after the sampling, any estimates obtained will be biased. In the situation where domains can be defined as a group of strata which span the whole population, and where domain size is known, domain estimation is equivalent to post-stratification (section 3.3.3).

The equations for estimates and variances for these different situations are given below.

First we consider domain estimation for a simple random sample of size *n*. If the number of sample units N_D in a domain *D* is known, *e.g.* from census information, then the domain total can be estimated using the unbiased estimator:

$$\hat{\tau}_D = N_D \bar{y}_D = \frac{N_D}{n_D} \sum_{i=1}^{n_D} y_i$$

(e.g., Thompson (1992), p41). This can be rewritten using the indicator variable *I*, to indicate whether sample *i* is in domain *D* or not, and sum over all the samples collected:

$$\hat{\tau}_D = \frac{N_D}{n_D} \sum_{i=1}^{n} y_i I_i \text{, where } I_i = \begin{cases} 1\, i \in D \\ 0 \text{ otherwise} \end{cases} \tag{5}$$

An estimator for the variance of $\hat{\tau}_D$, conditional on n_D is:

$$\hat{V}(\hat{\tau}_D) = \frac{N_D - n_D}{N_D n_D} s_D^2 \text{, where } s_D^2 = \frac{1}{n_D - 1} \sum_{i=1}^{n_D} (y_i - \bar{y}_D)^2 \text{, for } i \in D$$

When the size N_D of domain *D* is not known, we can use the unbiased estimator:

$$\hat{\tau}'_D = \sum_{i=1}^{n} \omega_i y_i I_i \text{, where } I_i = \begin{cases} 1\, i \in D \\ 0 \text{ otherwise} \end{cases}.$$

This is effectively the same as summing only over samples in *D*, but the equation for the variance takes account of the true sample size:

$$V(\hat{\tau}'_D) = N^2 \left(\frac{N-n}{Nn} \right) \sigma'^2 = \left(\frac{N-n}{Nn} \right) \frac{N^2}{N-1} \sum_{i=1}^{n} \left(y_i I_i - \frac{\tau_i}{N} \right)^2$$

which can be estimated by

$$\hat{V}\left(\hat{\tau}'_D\right) = N^2\left(\frac{N-n}{Nn}\right)s'^2 = \left(\frac{N-n}{Nn}\right)\frac{N^2}{n-1}\sum_{i=1}^{n}\left(y_i I_i - \frac{\hat{\tau}'_D}{N}\right)^2.$$

When the domain size N_D is known, $\hat{\tau}_D$ is preferable to $\hat{\tau}'_D$ as the variance is smaller.

Equation (5) is effectively the same as reweighting samples with a new weight ψ_i, so that

$$\hat{\tau}_D = \sum_{i=1}^{n}\psi_i y_i I_i \text{, where } I_i = \begin{cases} 1\; i \in D \\ 0 \text{ otherwise} \end{cases} \text{ and } \psi_i = \begin{cases} \dfrac{N_D}{n_D}\; i \in D \\ 1 \text{ otherwise} \end{cases} \tag{6}$$

3.3.3 Post-stratification

Post-stratification is equivalent to reweighting the sampling weights to account for the proportion of samples found in the new strata. For post-stratification of a simple random sample, the post-stratified estimator of the population total, $\hat{\tau}_S$ is given by:

$$\hat{\tau}_S = \sum_{i=1}^{n}\varphi_i y_i \text{, where } \varphi_i = \frac{N_k}{n_k} \text{ for } i \in \text{group } k. \tag{7}$$

If the sample is reasonably large, this technique is almost as precise as proportional stratified sampling. To use post-stratification, the relative size of each stratum must be known. If the relative stratum sizes are not known, they may be estimated using double sampling. Further discussion of post-stratification may be found in many classic sampling texts. Variance estimation for post-stratification varies among sampling texts. See for example, Lumley (2012) p 136 or Korn & Graubard (1999) p166 or Thompson (1992), p 109.

Comparison of equations (6) and (7) show that, in the case where the domains and the post-stratification strata are the same, and the domain sizes are known, $\hat{\tau}_S$ can be thought as of the sum of estimates of the domain totals $\hat{\tau}_D$.

The current estimation methods in the European Regional Data Bases under development and hosted at ICES can result in bias for on-shore sampling schemes where site-days are primary sampling units. Suppose the sampling design is set up with sampling of trips in ports on a particular day with the intention to sample trips by gear and area. Suppose also that gear/area are treated as strata in the analysis phase, and that the sampling weights are calculated using the sample and population sizes in the gear/area strata, and do not take into account the number of samples taken at the market on that day and the number available at the market to sample on that day, which give the true sampling probabilities for the samples. This has often been the case in fisheries sampling of landings, and will lead to incorrect estimates as the true sampling probabilities are not used in the estimation. Instead the sampling probabilities of trips on that day at the market should be included in the sample and domain estimation. Post-stratification can be used to correct for relative sample sizes in different gear/area combinations.

3.3.4 A Norwegian case study of design-based estimation for design class B

The Norwegian reference fleet represents an example of at-sea sampling where vessels are primary sampling units (principal design class B, Table 3.1). Here vessels are selected as the primary units. Within vessel, catch operations are sampled systematically through time at a constant rate. Within catch operation fish are sampled at random for ages, lengths and weights. It should be noticed that the hierarchy within boat also include the trip level, but since the sampling design dictates sampling catch systematically through time (regardless of the trip unit) we omit this level in this example, and thus leaves this as an example of a three stage sampling design. In this sampling program vessels are stratified according to gear and area.

To illustrate the principles of implementing a design based estimator we show weighted estimators for means, totals and ratios within a stratum. The stratified estimators to obtain estimates across strata may be outlined following the same principles. Let Y and Z be variables for which data has been collected and y_{ijk} and z_{ijk} values observed for the kth fish, in the jth haul at the ith boat corresponding to the sampling levels

1) Samples b of total B boats
2) Samples n_i hauls of total N_i within boat i, $i = 1, \dots, b$ at an approximately constant rate τ.
3) Samples m_{ij} fish of total M_{ij} within n_i, $j = 1, \dots, n_i$

In this example three estimators from classical sampling theory are considered; the total, mean and ratio which can be written generally as:

$\hat{\theta}_{tot} = \sum_{ijk} \omega_{ijk} y_{ijk}$, $\hat{\theta}_m = \frac{\sum_{ijk} \omega_{ijk} y_{ijk}}{\sum_{ijk} \omega_{ijk}}$, and $\hat{\theta}_r = \frac{\sum_{ijk} \omega_{ijk} y_{ijk}}{\sum_{ijk} \omega_{ijk} z_{ijk}}$,

where ω_{ijk} is the sampling weight for fish ijk. If the sampling weights are the inverse of the sample inclusion probability $\omega_{ijk} = {}^{1}\!/_{\pi_{ijk}}$, the first of these estimators estimates the population total of the variable Y. This is a Horwitz-Thompson estimator (cf Lumley 2010) which is a generalized estimator accounting for inclusion probabilities. Estimates of the mean and ratio are derived from the population total, and the estimator for the mean is also a ratio estimator if the population total size is unknown.

The joint inclusion probability can be written as a product of the conditional probabilities at each level in the sampling hierarchy: Define the observation variable I_i as a binary variable indexing the data that are sampled at stage i. Then the joint inclusion probability is

$$\begin{aligned}\pi_{ijk} &= P(I_i = 1, I_j = 1, I_k = 1) = P(I_i = 1)P(I_j = 1|I_i = 1)P(I_k = 1\,|I_i = 1, I_j = 1)\\ &= \pi_i \pi_{j|i} \pi_{k|ij}\end{aligned}$$

In this example it is assumed that the units are sampled by simple random sampling at each stage. Consequently the level inclusion probabilities are

$\pi_i = \frac{b}{B}$, $\pi_{j|i} = \frac{n_i}{N_i} = \tau$, and $\pi_{k|ij} = \frac{m_{ij}}{M_{ij}}$, respectively, such that $\pi_{ijk} = \frac{b}{B}\frac{n_i}{N_i}\frac{m_{ij}}{M_{ij}}$, or $\omega_{ijk} = \frac{B}{b}\frac{N_i}{n_i}\frac{M_{ij}}{m_{ij}}$.

First notice that for this design an estimator of the total number of fish M is given by $\widehat{M}_\omega = \sum_{ijk} \omega_{ijk}$. An estimator for the total numbers at age is then given by $\widehat{M}_\omega(a) = \sum_{ijk} \omega_{ijk} y_{ijk}^{(a)} = \sum_{ijk} \frac{B}{b}\frac{N_i}{n_i}\frac{M_{ij}}{m_{ij}} y_{ijk}^{(a)}$, where $y_{ijk}^{(a)}$ is a binary variable taking values 1 if the age is a and 0 otherwise.

Furthermore, the mean estimate of $Y^{(a)}$ is the proportion $p(a)$ of the population with $Y^{(a)} = 1$ such that $\hat{p}(a) = \frac{\sum_{ijk} \omega_{ijk} y_{ijk}^{(a)}}{\sum_{ijk} \omega_{ijk}} = \frac{\widehat{M}_\omega(a)}{\widehat{M}_\omega}$.

To increase the precision in the estimate of catch at age we used the auxiliary information of reported catch weights made available by the official landing statistics in retrospect of the sampling and scale the estimates accordingly. First realize that the numbers at age is the proportion at age times the numbers $M(a) = p(a)M$. In the preceding example and estimator of M was given utilizing the sampling weights. However an obvious estimator of M frequently used is the ratio of total reported catch weight W to mean fish weight, $\widehat{M}_w = \frac{W}{\bar{w}}$.

This suggests:

$$\widehat{M}_w(a) = \hat{p}(a)\widehat{M}_w = \hat{p}(a)\frac{W}{\bar{\bar{w}}},$$

i.e. the proportion at age by mean weight scaled to the total catch weight. The proportion at age by mean weight is a ratio estimator, and substituting for the sampling weights we obtain:

$$\widehat{M}_w(a) = \frac{\sum_{ijk}\frac{BN_iM_{ij}}{bn_im_{ij}}y_{ijk}^{(a)}}{\sum_{ijk}\frac{BN_iM_{ij}}{bn_im_{ij}}w_{ijk}}W = \frac{\sum_{ijk}\frac{M_{ij}}{m_{ij}}y_{ijk}^{(a)}}{\sum_{ijk}\frac{M_{ij}}{m_{ij}}w_{ijk}}W,$$ i.e. knowledge about number of vessels and number of catch operations are not necessary. This is the estimator used for the comparison of design based and model based estimation of catch at age in the Norwegian case study presented at this workshop (Aanes and Hirst, 2012).

Comparisons of $\widehat{M}_\omega(a)$ and $\widehat{M}_w(a)$ are of interest, but has not been done for the Norwegian data because data on total number of vessels B and total number of catch operations within each vessel i, N_i, has not yet been compiled.

The above estimators are implemented using R survey package (CRAN, Lumley 2010). This includes standard approaches to estimate means, ratios and totals for a probability based design. It also offers standard methods for estimating precision and correlation structures by various methods including analytical estimators for variance where they exist, approximation by linearization (e.g. for ratio estimators), and re-sampling methods such as bootstrapping. Of particular interest in this setting is that a ratio estimator is approximately unbiased if the sampling size is sufficiently large. The variance of the ratio estimator is based on linearization (Taylor expansion) or re-sampling methods. Both approaches depend on sufficient sample sizes and generally suffer if sample sizes (#PSU's within stratum) are small. This is illustrated in Aanes and Hirst (2012) who showed by simulations that the estimates of standard error and the coverage of 80% confidence level both decreased with increasing stratification (<40% coverage for the full stratification for the Norwegian data), keeping the number PSU's constant but varying the stratification. Increased level of stratification thus means a reduced number of samples in each stratum.

For the Norwegian data, domain estimation has not been considered but the principles are outlined as follows: Estimates of a subpopulation domain $\mathcal{D}$ of interest (e.g. a specific metier) is obtained by assigning sample weights of zero for observations outside the domain, whereas sampled individuals within $\mathcal{D}$ retain their original sample weigths. This is achieved by including the indicator function $I[(ijk) \in \mathcal{D}]$ that equals 1 if the kth sampled individual in haul j in PSU i is in $\mathcal{D}$, and 0 otherwise. Effectively this reweights the data according to its actual design to achieve appropriate weights for the domain. This will also affect the variance estimate see for example Korn and Graubard (1999) pp. 207-211 for details. The obvious restriction by domain estimation is that samples for the domain of interest must be present.

3.3.5 A Scottish case study of design-based estimation for design class C

Here we consider how the estimators described in section 3.3.1 would apply to estimation of total Scottish landed numbers-at-age for key demersal species by sampling landed fish at markets. As this is for illustrative purposes, we gloss over some of the practicalities of sampling which complicate the issue, such as: markets which are not sampled; the fact that, for smaller markets, we cannot predict whether a market will take place or not on a particular day; and that the selection of fish sampled for age is not random over sales categories. The sampling frame is Scottish markets. Stratification levels are market and sales category within a trip. Sampling units are: primary sampling unit (SU1) – market-day within the stratum market, secondary sampling unit (SU2) – trip within market-day, SU3 – box within stratum sales category, SU4 – fish measured for length (and age). Fish are measured for age within a stratified sample of fish measured for length, however the lengths are currently stratified over the whole trip rather than each sales category and so we first estimate numbers-at-age at the trip level.

First we define some notation. Let N represent the total number in the population, and n represent the number sampled, with subscripts to denote the sampling unit or stratum of interest and suffices to denote the variable of interest. For example $N^{(a)}$ is the number-at-age, and N_M is number of markets. The sampling procedure is as follows:

1. Sample n_M of a total of N_M markets. (Here we set $n_M = N_M$.)
2. Sample n_m market-days of a total N_m market-days within market m.
3. Sample n_{mk} trips of a total N_{mk} within market-day k of market m.
4. Sample all N_{mki} sales categories of trip i within market-day k of market m (to ensure a complete length distribution).
5. Sample n_{mkih} boxes of a total N_{mkih} within sales category h of trip i within market-day k of market m. (Usually, but not always, n_{mkih}=1.)
6. Measure for length n_{mkihb} fish of a total N_{mkihb} fish in box b of sales category h of trip i within market-day k of market m. (Here $n_{mkihb} = N_{mkihb}$.)
7. Age n_{mkil} fish of a total $n_{mki}^{(l)}$ of length l in trip i within market-day k of market m, within a total of n'_{mki} fish sampled in trip mki. (Note that $n'_{mki} = \sum_{h=1}^{N_{mki}} \sum_{b=1}^{n_{mkih}} n_{mkihb}$).

To perform the estimation, we first estimate the landed length distribution for that trip, $\hat{N}_{mki}^{(l)}$, where $N_{mki}^{(l)}$ is the total landed number-at-length l for trip i within market-day k of market m. Historically, the usual practice is to first raise sampled numbers-at-length in a category to total numbers-at-length in the category $\hat{N}_{mkih}^{(l)}$, using the inverse proportion of boxes sampled in that category, N_{mkih} / n_{mkih}, then sum over all categories to aggregate to the trip level.

Define $y_{mkihbj}^{(l)}$ such that $y_{mkihbj}^{(l)} = 1$ if fish j in box b of sales category h of trip i within market-day k of market m is of length l and 0 otherwise:

$$y_{mkihbj}^{(l)} = \begin{cases} 1 \text{ if fish } mkihbj \text{ is of length } l \\ 0 \text{ otherwise} \end{cases}$$

Then the above procedure can be written as two equations:

$$\hat{N}_{mkih}^{(l)} = \frac{N_{mkih}}{n_{mkih}} \sum_{b=1}^{n_{mkih}} y_{mkihbj}^{(l)} \text{ and } \hat{N}_{mki}^{(l)} = \sum_{h=1}^{n_{mki}} \hat{N}_{mkih}^{(l)}$$

which can be condensed into:

$$\hat{N}_{mki}^{(l)} = \sum_{h=1}^{n_{mki}} \hat{N}_{mkih}^{(l)} = \sum_{h=1}^{n_{mki}} \sum_{b=1}^{n_{mkih}} \frac{N_{mkih}}{n_{mkih}} y_{mkihbj}^{(l)}$$

This historical raising process is actually equivalent to the application of equation (3) from section 3.3.1 at the trip level, with the sampling weight for fish *j* in box *b* of sales category *h* (of trip *i* within market-day *k* of market *m*), ω_{mkihbj}, given by $\omega_{mkihbj} = \frac{N_{mkih}}{n_{mkih}}$.

Next we estimate a proportional age-length key, $\hat{p}_{mki}^{(a,l)}$, for trip *i*, apply it to the length distribution for the trip to get an age-length distribution for the trip, and sum over lengths to get an age-distribution for the trip. The age-length distribution for the total catch in a stratum is then estimated as a weighted average of age-length distributions across the trips within the stratum. This is an alternative approach to estimating catch-at-age, where age-length keys are used only within primary sampling units (see also Hirst et al. 2012). The age data within a trip (e.g. 1 or 2 otoliths per length group) is applied to the length frequency for that trip. This is an alternative approach to the common practice of applying a pooled ALK for many trips (for example by strata) to the raised telngth-frequency distributions (LFDs). The proportional age-length key is simply based on the number at age *a* for a given length *l* of the stratified sample of length-measured fish over the trip:

$$\hat{p}_{mki}^{(a|l)} = \frac{n_{mki}^{(a,l)}}{n_{mki}^{(l)}}$$

and we apply this to the length distribution to give an age-length distribution for the trip:

$$\hat{N}_{mki}^{(a,l)} = \frac{n_{mki}^{(a,l)}}{n_{mki}^{(l)}} \hat{N}_{mki}^{(l)} = \frac{1}{n_{mki}^{(l)}} \sum_{u=1}^{n_{mki}^{(l)}} y_{mkiu}^{(a,l)}$$

which we then sum over length to estimate an age-distribution, i.e. numbers-at-age *a* for each trip *mki*, $\hat{N}_{mki}^{(a)}$.

$$\hat{N}_{mki}^{(a)} = \sum_{l} \hat{N}_{mki}^{(a)} = \sum_{l} \frac{n_{mki}^{(a|l)}}{n_{mki}^{(l)}} \hat{N}_{mki}^{(l)} = \sum_{l} \frac{\hat{N}_{mki}^{(l)}}{n_{mki}^{(l)}} n_{mki}^{(a,l)}$$

Return now to the market, and consider the number-at-age of each trip, $\hat{N}_{mki}^{(a)}$. The total landed number-at-age *a*, $N^{(a)}$, sold at the markets sampled, can be estimated by the weighted estimator given by equation (3) of section 3.3.1:

$$\hat{N}^{(a)} = \sum_{mki} \omega_{mki} \hat{N}_{mki}^{(a)} = \sum_{m=1}^{n_M} \sum_{k=1}^{n_m} \sum_{i=1}^{n_{mk}} \omega_{mki} \hat{N}_{mki}^{(a)}$$

where

ω_{mki} is the sampling weight for trip *mki*.

Now $\omega_{mki} = \dfrac{1}{\pi_{mki}}$, where π_{mki} is the inclusion probability of trip *i* on market-day *k* at market *m*, and this is given by the probability of selecting trip *i* on market-day *k*, multiplied by the probability of selecting market-day *k* for market *m*, multiplied by the probability of selecting market *m*, *i.e.*

$$\pi_{mki} = \pi_m \pi_{k|m} \pi_{i|mk}$$

All markets are selected, then market-days are selected and then trips at the market, so $\pi_m = 1$, $\pi_{k|m} = \dfrac{n_m}{N_m}$ and $\pi_{i|mk} = \dfrac{n_{mk}}{N_{mk}}$ and hence $\omega_{mkihblj} = \dfrac{1}{\pi_{kihblj}} = \dfrac{N_m}{n_m}\dfrac{N_{mk}}{n_{mk}}$.

Thus $\hat{N}^{(a)} = \sum_{m=1}^{n_M} \sum_{k=1}^{n_m} \sum_{i=1}^{n_{mk}} \omega_{mki} \hat{N}_{mki}^{(a)} = \sum_{m=1}^{n_M} \sum_{k=1}^{n_m} \sum_{i=1}^{n_{mk}} \dfrac{N_m}{n_m} \dfrac{N_{mk}}{n_{mk}} \hat{N}_{mki}^{(a)} = \sum_{m=1}^{n_M} \dfrac{N_m}{n_m} \sum_{k=1}^{n_m} \dfrac{N_{mk}}{n_{mk}} \sum_{i=1}^{n_{mk}} \hat{N}_{mki}^{(a)}$

Now compare this to the historical practice of "raising". Here the number-at-age at trip level $\hat{N}_{mki}^{(a)}$ is raised to the number-at-age on market day *mk*, $\hat{N}_{mk}^{(a)}$, by summing over trips and then raising by the inverse proportion of trips sampled on that market day, N_{mk}/n_{mk} :

$$\hat{N}^{(a)} = \sum_{m=1}^{n_M} \frac{N_m}{n_m} \sum_{k=1}^{n_m} \hat{N}_{mk}^{(a)}, \text{ where } \hat{N}_{mk}^{(a)} = \frac{N_{mk}}{n_{mk}} \sum_{i=1}^{n_{mk}} \hat{N}_{mki}^{(a)}.$$

Next, the numbers-at-age for the sampled market-days, $\hat{N}_{mk}^{(a)}$, are raised to numbers-at-age for the market, $\hat{N}_m^{(a)}$, by summing over sampled market-days, and raising by the inverse proportion of market-days sampled from each market *m*, N_m/n_m :

$$\hat{N}^{(a)} = \sum_{m=1}^{n_M} \hat{N}_m^{(a)}, \text{ where } \hat{N}_m^{(a)} = \frac{N_m}{n_m} \sum_{i=1}^{n_m} \hat{N}_{mk}^{(a)}$$

and finally we sum over markets.

So estimation by means of sample weights is a formalisation of the historical raising process, with two main differences. First, historically, samples have been pooled together into groups of samples by, for example, gear and quarter, rather than acknowledging each stage of the actual sampling process, such as sampling at markets on a particular day. Secondly, the inverse proportion of units sampled to total number of unit (*i.e.* the selection probability at each stage if random sampling is used) is used as the raising factor at each stage, rather than the proportion of landed weight sampled to total landed weight which has often been used historically. Proportions of landed weights can be used to estimate the proportions of units sampled if these are not available (see Norwegian case study), but for the Scottish case study we have not ascertained which approach is better when both are available.

3.4 Model-based estimation

Hirst et al. (2004, 2005, 2012) present a model-based estimation approach that can be utilised for catch sampling schemes where the following assumptions can reasonably be made:

1) The primary sampling unit (PSU) is well defined, and the survey design involves simple random sampling of PSUs from some population of units (i.e. collection of all primary sampling units). The PSU is most likely to be the fishing trip but could be individual hauls, sets etc. The fraction of the total PSUs sampled is very low, so that sampling with replacement can be assumed;
2) a random sub sample of fish is taken from each PSU;
3) there are some representative measurements of length and age, and of length and weight for the sub sample of fish; and
4) all biological data can be linked to a PSU.

In Hirst et al. (2012) a Bayesian hierarchical model is developed to estimate catch-at-age from commercial fishery data from multiple sources. Many common forms of fisheries sampling data can be utilised in the model: age and length, length-stratified ages and length only. There is no need to construct an age-length-key in this modelling approach. The model is currently used by Institute of Marine Research to combine catch sampling data from multiple sources to estimate catch-at-age for Norwegian landings of cod and haddock.

3.5 Concepts of sampling frames, domains, and country levels in view of regional databases (RDB)

Defining sampling plans and raising procedures for fisheries data has always raised difficult issues in all fisheries institutes. The recent developments in order to improve randomness and avoid over-stratification of the sampling frame, to correctly raise the information to the aggregation demanded (domain of interest) and to move toward international database, is often perceived as something impossible to achieve.

The sampling frame was defined as the complete set of non-overlapping units of the entire population (e.g. all vessel trips in a year, all site-day combinations in a year, or all fishers) in the WKPICS1 report (ICES, 2011a). The domain of interest was introduced in WKPRECISE (ICES, 2009) and defined as a major segment of the population for which separate statistics are needed (Statistics Office of the United nations, 1984; Cochran 1977). A study domain would typically be a sub-population (or sub-set of the sampling frame) identified in the overall statistical plan as one for which a certain level of precision is required. An example of study domains in fisheries catch sampling programs is the métier system for grouping catches used in the current Data Collection Framework (DCF) in the EU, where legislation specify minim um sampling levels by metiers and target precision levels are set for stocks.

In a regional database configuration, the country level will act also as a stratifier of the overall population, so that eventually a population will be perceived as stratified with nb country * nb frame * nb domains, which should not be the case. Figure 3.2 below illustrates the point and further discussion how to address the issue.

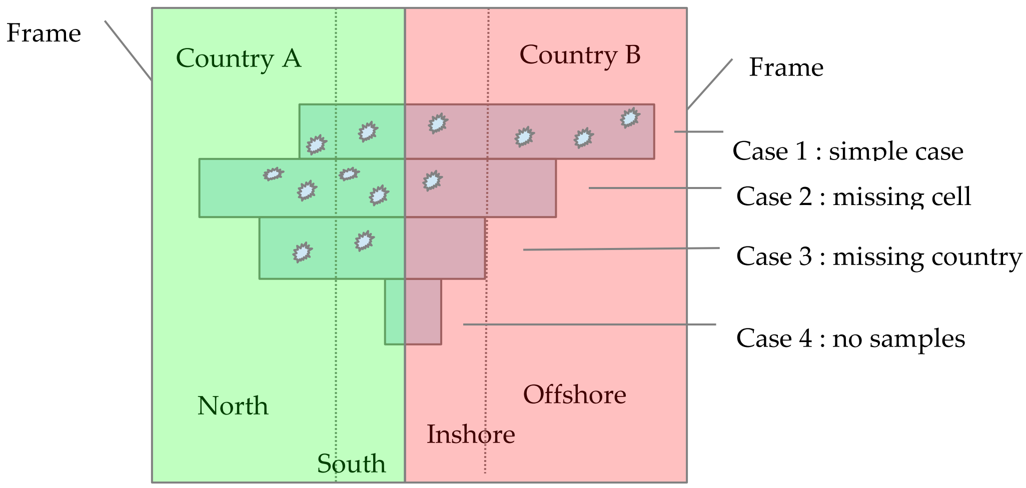

Figure 3.2. Schematic showing two national sampling frames divided into strata (north-south for country A and inshore-offshore for country B) with fishing activity and sampled trips clustered by domains of interest (cases 1 – 4).

In Figure 3.2, the two countries each have a frame composed of e.g. fishing vessels. Country A has chosen to stratify the sampling frame between North and South, and country B preferred the decomposition inshore offshore, both for internal reason. By doing so, they guarantee that the samples will be drawn randomly within each of the frame strata. The resulting samples are displayed as circles, voluntarily limited to very few for clarity. When processing the data, the client requires information at the domain level, represented in the figure by the four horizontal rectangles, within which fleet activity occurs throughout the rectangles which cover one or more strata in each country. As a result

- The upper rectangle is well covered by samples (case 1), the use of equations below will enable to proceed to any estimation for the domain;
- The second and third domains (case 2 and 3) have a missing cell, either in one frame stratum or in the country, so that imputation or reweighting is needed for enabling the estimation;
- The fourth domain (case 4) has not received any samples. It is impossible in this case to do any imputation (unless expertise knowledge) and thus it is impossible to respond to the client request. If the client insists for sampling this domain, the strat-

ification of the sampling frame must be changed, keeping in mind the necessary proprieties of the stratification and the need to guarantee the full randomness of sample draw in each of the stratum.

The first message is that countries should be free to stratify their sampling frames in manners that are suitable for their national sampling programs, taking into account their resources for sampling and fisheries characteristics. Moreover, the stratification of the sampling frames from one country can be entirely different from another country, without direct implications on the data processing. To this end, each of the samples should be accompanied by its inclusion probability within the frame stratum (e.g. 1/N PSU), or the necessary data to compute the sample weights. In a regional database perspective populated by national institutes, it recommended that all data required to calculate the sample weights for the national data be provided, or alternatively that the inclusion probabilities be given for all data at a national level. This would allow design-based estimation for each country, which then forms a stratum for regional estimates. It is recommended that an inclusion probability field be added at the sample level in the data exchange format for the regional database. In addition, it is recommended that auxiliary data that can be used to adjust sample weights be included in the regional database (see section 3.7).

The equations to be used to estimate parameters at the domain level taking into account inclusion probabilities can be found in Korn and Graubard (1999). The estimation of catch statistics and biological parameters for domains that cut across sampling strata requires careful consideration, ensuring that the weighting factors associated with the sample selection in each stratum are represented in the analysis. An example provided by Joël Vigneau to WKMERGE (ICES 2010) helps to clarify the requirements (Figure 3.3). The example is for two domains (metiers) M1 and M2 occurring in two sampling strata (North and South), and a third (M3) occurs only in the Northern stratum. The analysis problem is to obtain estimates for the metiers across strata. Table 3.2 lists some options and their merits.

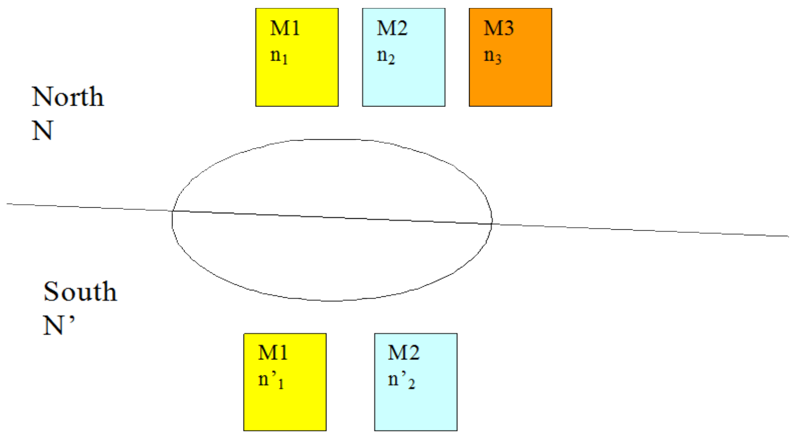

Figure 3.3. Theoretical example where two domains (metiers) M1 and M2 occur in two sampling strata (North and South), and a third (M3) occurs only in the Northern stratum. The $n_{1,2,3}$ refer to sample sizes (number of PSUs).

Table 3.2: Estimation formula for different combinations of metiers and strata, for situations with equal or unequal sampling probabilities in each stratum.

π_i = probability of sampling one unit in strata
x_i^1 = variable collected for unit i in metier 1

	Equi probability of sampling between strata	Raising method	Benefits	Drawbacks
Raising to the population				
1st solution	Yes	$\theta=\frac{N+N'}{(\sum n+\sum n')}(\sum_i x_i+\sum_i x_i')$	May be used to estimate the design effect. N by strata may be unknown	Irrespective of the stratification Likely to be biased
2nd solution	No	$\theta=\frac{N}{(n_1+n_2+n_3)}(\sum_i x_i^1+\sum_i x_i^2+\sum_i x_i^3)$ same calculation for θ' then $\theta=\frac{N}{N+N'}\theta+\frac{N'}{N+N'}\theta'$	Easy to calculate Likely avoiding gaps and poor sampling	Highly reliable on the randomness. Accurate only for providing information at the strata level
Raising to the metier (M_1)				
1st solution	Yes	$\theta=\frac{N+N'}{(n_1+n'_1)}(\sum_i x_i^1+\sum_i x_i^1{}')$	Easy to calculate	Difficult to implement a sampling scheme with equiprobability over all population
2nd solution	No	$\theta=\frac{N}{n_1}\sum_i x_i^1+\frac{N'}{n'}\sum_i x_i^1{}'$	Easy to calculate	One domain may have received very poor or no samples
3rd solution	No	$\theta=N\frac{\sum_i x_i^1}{\sum_i \pi_i}+N'\frac{\sum_i x_i^1{}'}{\sum_i \pi_i'}$ (Horvitz-Thomson)		π_i not easy to evaluate
Merging of metier (M_2+M_3)				
1st solution	Yes	$\theta=\frac{N+N'}{(n_2+n_3+n'_2)}(\sum_i x_i^2+\sum_i x_i^3+\sum_i x_i^2{}')$		
2nd solution	No	$\theta=\frac{N}{n_1+n_2}(\sum_i x_i^2+\sum_i x_i^3)+\frac{N'}{n_2'}\sum_i x_i^2{}'$	Easy to calculate	One domain may have received very poor or no samples
3rd solution	No	Horvitz-Thomson		

The system works well in the cases where at least one sample is received from each of the frames concerned by the domain of interest.

Numerical example of domain estimates

The figure below provides a simple numerical example of a stock sampled by two countries, each with their own sampling frame. Each sampling frame has two strata. The figure illustrates how a parameter estimate can be raised to the population level using the sampling proportions. If sampling is truly random across the domains of interest, the overall sampling proportions in each stratum can be used. If there is reason to believe that certain domains are over- or under- represented, the weights can be adjusted by using domain-specific weights (this is, in effect, post-stratification).

Stock	Stock X			
Country	Country A		Country B	
Sampling frame	Sampling Frame A		Sampling Frame B	
PSU	Port-day		Vessel-week	
Stratum	North	South	Small vessels	Large vessels

Stock X is sampled by two countries with different sampling frames and different primary sampling units (PSU's). Each country has two strata.

Number of PSU's in population (N)

Domain of interest				
OTB	240	550	20	55
TBB	125	120	80	96
SSC	0	10	4	5
GNS	0	50	0	0
Total	365	730	104	156

Number of PSU's sampled (n)

Domain of interest				
OTB	7	7	3	3
TBB	3	2	17	7
SSC	0	0	0	0
GNS	0	1	0	0
Total	10	10	20	10

Parameter estimate (x)

Domain of interest				
OTB	2	3	7	9
TBB	10	5	5	1
SSC				
GNS		5		
Total	12	13	12	10

This can be any parameter, for example the number of three-year old fish in the samples.

Sampling proportions (n/N)

Domain of interest				
OTB	0.029	0.013	0.150	0.055
TBB	0.024	0.017	0.213	0.073
SSC		0.000	0.000	0.000
GNS		0.020		
Overall	0.027	0.014	0.192	0.064

Sampling proportions are more or less constant down the columns. If this is not the case, you have over- or under-sampled some domains. In that case you can post-stratify by using domain-specific sampling proportions. In the ideal case this is not neccesary.

Population estimate (X*N/n)

Domain of interest					Total
OTB	68.6	235.7	46.7	165.0	516.0
TBB	416.7	300.0	23.5	13.7	753.9
SSC					
GNS		250.0			250.0
Total	485.2	785.7	70.2	178.7	1519.9

Population estimate using domain-specific sampling proportions (i.e. post-stratified - corrected for unbalanced sampling).

Population estimate (X*N/n)

Domain of interest					Total
OTB	73.0	219.0	36.4	140.4	468.8
TBB	365.0	365.0	26.0	15.6	771.6
SSC	0.0	0.0	0.0	0.0	
GNS	0.0	365.0	0.0	0.0	365.0
Total	438.0	949.0	62.4	156.0	1605.4

Population estimate using overall sampling proportions.

Figure. 3.4. Simple numerical example of a stock sampled by two countries, each with their own sampling frame comprising two strata, illustrating how a parameter estimate can be raised to the population level using the sampling proportions

3.6 Regional databases in ICES

3.6.1 Present situation

The primary aim of the national catch (landings + discards) sampling programs under the DCF is to provide total population catch numbers at age, and weights at age for input to stock assessments. In addition, the ICES Working Group on Mixed Fisheries Advice for the North Sea (MIXFISH) and STECF require catch numbers at age at a finer scale, essentially by fleet segment, for evaluating mixed fishery management options. The high cost of sampling programmes means that the data collected are based on a relatively small sample size that is insufficient for estimating parameters for fine-scale domains.

At present there are three systems developed for processing the catch sampling data ready for use at the assessment working groups in ICES – InterCatch, RDB-FishFrame and COST. Currently none of these programs process the data in accordance with best practice set out in Annex 3, as explained below.

3.6.2 InterCatch

Estimated catch in numbers-at-age or non-sampled reported landed weights on a relatively fine-scale resolution (métiers) are currently submitted to InterCatch on a national level. Figures for métiers with no samples are then "filled-in" by assigning the age-structure from a similar métier where samples have been collected. This is a highly complex procedure that is largely unnecessary since the requirements of most stock assessment working groups is numbers at age for the total catch of a stock, and not by métier.

The filling in of data gaps for fine-scale domains (métiers) for which no samples have been obtained on a large scale in InterCatch may introduce serious bias. An alternative approach for future data collections, as recommend in WKPICS, is to employ stratified catch sampling schemes, where the sample sizes in strata can be controlled. Domains of interest in the estimation phase should be of a broad enough scale to ensure that a sufficiently high proportion of the relevant landings within that domain have been sampled. Parameter estimates for the domain can then be estimated by proper weighting of the data, for example by using census data on landings by domains for post-stratification.

3.6.3 RDB-FishFrame

The raising procedures currently used in the RDB are based on stratified estimators, applied to post-strata based on fishery/gear, time and space. It is generally assumed that fishing operations or trips are primary sampling units, and that these are sampled randomly within the post-strata. The current practice of quota sampling for métiers, and the filling in of data-gaps for un-sampled fine-scale métiers on a large scale in the RDB can result in serious bias. Domains of interest should be of a broad enough scale to ensure that a high proportion of the relevant landings within that domain have been sampled. The current estimation system, where fishing operations or trips define the PSUs, cannot accommodate correctly weighted estimation for more clustered sampling schemes, such as schemes B, C and D described in this report, where primary sampling units are vessels or site-days. The data-structure for the current RDB cannot accommodate the development of appropriate estimation weights for samples collected using designs recommended by groups such as WKPICS1, WKPICS2 and SGPIDS, which are based on vessel lists (at-sea sampling) or sale/landing locations (on-shore sampling). Therefore the current raising procedures in the RDB cannot be used to estimate population totals for data collected by national programs if they move towards the best-practice sampling schemes advocated here.

3.6.4 COST

The COST package contains data structures based on the FishFrame format. However, like for the RDB, many of the "raising" tools in COST are based only on trip as a PSU (design option A, Table 3.1) which generally is not appropriate for more clustered design such as design options C and D where sampling of trips is conducted at site-days (landing/sale locations). Further developments of COST would require updating the data structures in line with the RDB, and updating the raising functions to the best-practice methods recommended by WKPICS2.

3.7 Recommendations for future improvement of Regional Data Bases

3.7.1 Data storage

In the short term, national data should be uploaded to the RDB, and estimation carried out with the RDB according to the national scheme, and submitted to InterCatch. In the longer term national and regional estimates (and their variances) also need to be stored in the RDB and formats for these estimates need to be specified. Metadata on sampling programmes need to be stored regionally, so that numbers-at-age provided by national programmes can be linked to the sampling programme.

3.7.2 Estimation

The strengths of the RDB lie in data storage and archiving at a regional level, and to a lesser extent, the ability to run standard queries or reports on the regional data. However, WKPICS suggests that the statistical estimation of catch-at-age be carried out in a statistical framework, using generic functions developed in a statistical package such as R where appropriate. The analysis could be done within the RDB, but it would be advantageous to include the option to link to R macros from the RDB. Many countries already use R to some extent in their analysis of catch sampling data. An alternative would be to develop an analysis module outside of the RDB which can use modules from the R survey package developed by Thomas Lumley[1] (Lumley 2010). Such a module could for example be developed as an extension of the COST software (Figure 3.5). The R-survey package is widely used and quality-assured by respected statisticians. This would also allow immediate use of developed code as it became available, while this new estimation framework is being developed, and would also allow continuous development of the analyses as required. If national sampling schemes are designed in accordance with best practice, this will allow generic R macros to be used within the RDB (or in a separate module) for estimation on a regional scale. Estimation of sampled population totals (e.g. catch numbers-at-age) at a regional level should be a very straightforward process of summation of national totals, then weighting these regional "sampled population" estimates to "raise" them to the total population (i.e. including non-sampled landings).

---

[1] http://faculty.washington.edu/tlumley/survey/

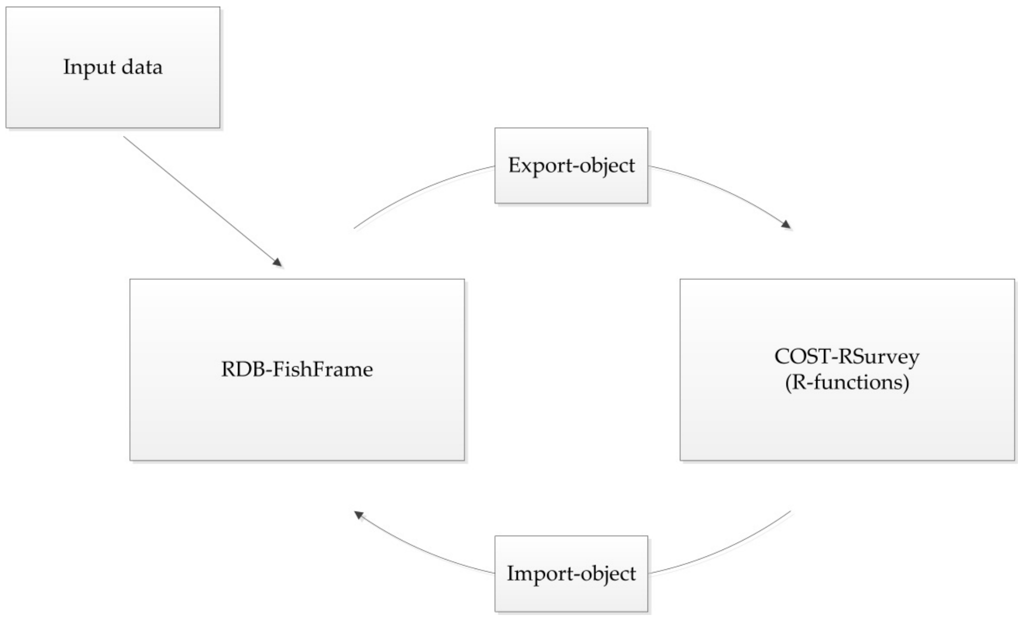

Figure. 3.5. An example of a regional database and analysis framework where the regional estimates are computed in an analysis module based on COST and R survey functions.

Regional estimates of e.g. catch numbers-at-age based on raw data stored in the RDB would most easily be facilitated if sample weights are computed and stored for all records. The actual computational steps to derive appropriate sample weights will vary among national surveys. An example of a protocol for computing survey weights are found in the sample weights manual for the Global Adult Tobacco Survey Collaborative Group (2010). It is recommended that manuals for the computation of sample weights be developed through an ICES workshop to ensure efficient and coordinated procedures.

3.7.3 Estimation for non-sampled domains on a regional level

The TORs of one the RDB workshops in 2013 should specifically address the data requirements of ICES MIXFISH, assessing the number of non-sampled domains of interest, and the sample size in sampled domains with a view to develop a stratified sampling design and the definition of fewer, more broad-scale domains that would still fulfil the needs of MIXFISH, and to suggest estimation procedures for these domains. Members of MIXFISH who understand the key requirements of the models used by MIXFISH would be required at this workshop, and this workshop should be held in the first months of the year, prior to use of the data at the assessment working groups.

3.7.4 Exchange format for RDB

The sampling frame, and the primary, secondary, tertiary and quaternary sampling units (PSU, SSU, TSU and QSU) should be easy to identify for common functions.

One of the crucial requirements for estimation is knowledge of the sampling inclusion probabilities used to specify design-based sample weights. To enhance the regional database so that it will support regional, national, stratum-specific, and domain estimates in the future, two development scenarios can be envisaged:

1. The database be modified to include all necessary data for computing sample weights for raw data from each national scheme to support estimates at all levels (regional, national, strata, domains)
2. The database be modified to include design-based sample weights (e.g., based on inclusion probabilities) for the raw element data, and adjusted model-assisted sample weights (e.g., using census data) when appropriate for each national scheme.

Option (1) has the advantage that sample weights computed in the RDB can be verified. The weights can also be re-computed, for example to balance the sample data with census data when appropriate. An additional advantage is that this option is robust to revisions in national census data (landings, effort) in subsequent years, which is common. It is easier to update the landings when appropriate than to revise all the sample weights (option 2).

Option (2) has the advantage that all necessary national data for estimating parameters at regional scales (raw data and sample weights) may be stored in the RDB without providing detailed national landings data. Census data on national landings might be used to adjust weights for each national data set before they are submitted to the RDB, without the need to provide all national landings statistics to the RDB. However, this approach would make it more complicated to revise sample weights in the RDB if national census data are revised over time.

The WKPICS2 group had a lot of discussions on how to include these options in the exchange format, and at what level landings & effort data should be included. The following points were noted.

- To allow future adjustment of the sampling weights during post stratification, or domain estimation. Estimation for domains requires knowledge of the total number of PSUs in the domains, or estimates of a proxy such as total catch, and the ability to allocate each sampled PSU to the domain.
- It should be evaluated if it is possible to do the adjustments with an exchange format on aggregated commercial effort and landing statistics by adding a new table with metadata
- How to recalculate sampling weights when the PSU is vessel, and it is not straight forward to sum up on another level?

The group could not at this stage specify detailed format changes in the RDB that would allow for design-based regional estimates based on raw data in the RDB. WKPICS therefore recommends that one of the WKRDB in 2013 will be dedicated to the practical work with the functions suggested.

The current exchange format already supports some auxiliary variables needed for domain estimations, but whether these are enough depends on which groups the RDB-FishFrame should support.

The ICES PGCCDBS and SGPIDS have suggested the establishment of a regional repository for descriptions of each national sampling schemes and also suggested that the RDB could provide that direct link to the data. WKPICS supports this suggestion. The routine production of quality assurance reports (covered in the next section) will be dependant on being able to associate stock data to sampling schemes. The RDB could more naturally provide that function and the templates for these reports. This will require metadata to be stored, and key descriptors for each national sampling scheme would need to be held for each stratum. As well as improving on the functionality of the RDB in aggregating data with reference to how it was collected, this would also provide valuable information to assist the work of the Regional Coordination Groups.

3.7.5 Suggested fields, to be discussed by the RDB Steering Committee, are:

Table of sampling overview

1. Country
2. Sampling scheme – onshore, offshore
3. Sampling frame descriptor
4. Stratum code (vessel groups, port groups)
5. Stratum descriptor
6. PSUs
7. Start date
8. End date
9. Link to document repository

(10. – n. Additional estimators of good practice.)

Additional fields will need to be held in relation to the landings (CE and CL tables?) that identify the onshore and offshore stratum for each MS for example total number of ports or vessels.

4 Quality indicators and levels for onshore and offshore sampling schemes (ToR c)

4.1 Background

These sampling programmes provide data for regional estimates of catch at age and length, catch composition, quantities discarded, maturity ogives and length weight relationships to ICES assessment EGs, STECF or other stakeholders. With multiple countries contributing to these estimates at huge expense, one of the roles of the future Regional Coordination Groups will be to ensure these combined estimates can be quality assured. This will require a framework for documenting and archiving data quality.

Expert groups and funders such as the EU Commission are interested in a relatively high–level overview of data quality, particularly the precision, the potential level and impact of bias, and how quality varies between the countries providing the data.

There are many steps leading to the total international catch estimates for a stock, as provided to a stock assessment working group. WKPICS suggests a top down approach to providing a quality assurance (QA) report that follows the levels that describe the regional sampling design. It should start with the catch estimate for a stock at a regional level, and then the sampling design for each member state at the national level and the strata within their sampling frames. Figure 4.1 describes the breakdown of an international fleet and the different levels within it.

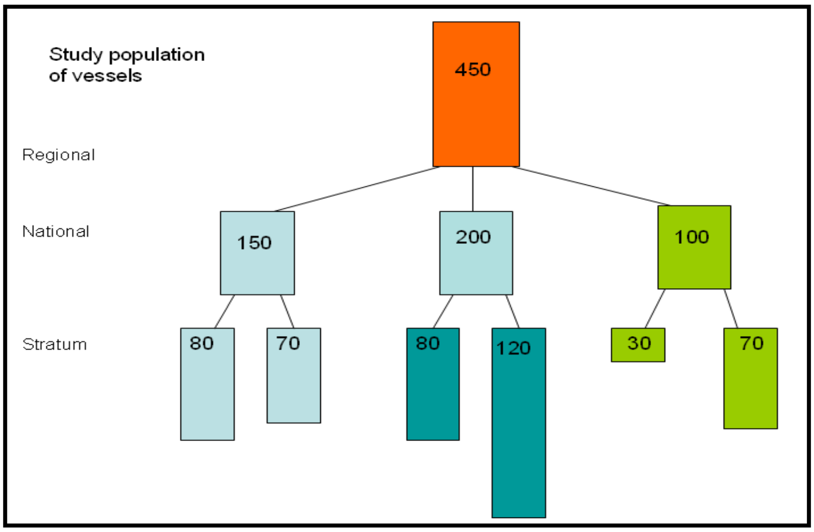

Figure. 4.1. Stratification of a regional (international) fishing fleet.

Recommendations on fishery data quality indicators from SGPIDS (ICES 2012a), WGRFS (ICES 2012b) and WKACCU (ICES 2008) were reviewed. Following the top-down approach, similar to SGPIDS and WGRFS, the most informative indicators for end users are at the highest level. Significant issues at this level and in the strata directly below, in relation to biases inherent in sampling design and implementation, need to be highlighted. The

WKACCU score card was found by WKPICS to be relatively ineffective at easily highlighting underlying biases at the highest levels. Whilst the score cards are useful as a comprehensive list for national institutes to screen their sampling schemes for a wide range of potential biases, they are qualitative and without complex weighting of each of the measures at the lower levels it is difficult to come up with an overall higher level score. The nested structure suggested by WGRFS provides a better overview of bias and a structure to indicate bias at all levels.

WKPICS has chosen to concentrate on producing a simple QA report for each stock with reference to key indicators of bias which if significant can be highlighted and explained in the overview.

4.2 Quality indicators (QI)

Two components of quality are considered: **Precision of estimates** and **Biases. Biases** can for example arise from incomplete coverage of the population, non-representative selection of the PSU, systematic errors due to non-response (e.g. refused access to vessels), or inappropriate methods of data analysis. Bias can arise at the three of the stages of the estimation process: the sampling scheme design, implementation and the analysis of the results.

Bias associated with ***sampling scheme design*** is related to the sampling coverage and the methods for selecting primary sampling units. For example: to deal with limited resources, geographically inaccessible areas, and a multitude of less 'significant' ports, a country may choose to concentrate sampling to the key ports. This has been done in Scotland for their onshore sampling scheme for demersal landings. The sampling frame has been limited to 5 key ports where approx 80% of the catches are landed - this leads to a systematic bias but it is documented and in this instance is acceptably low.

Bias associated with the ***implementation*** phase is a failure to meet an intended survey design, leading to non–representative sampling of the population. Biases may occur as a consequence of fisherman or merchants barring access to landings or trips. Determining non-response rates and refusal rates help qualify this potential bias.

Bias can also arise at the ***analysis*** stage. This can typically occur where the stratification schemes, sample probabilities and cluster sampling effects are not taken into account during the analysis by applying appropriate sample weights when raising the samples to the population.

4.3 Quality Assurance (QA) reports

4.3.1 General principles

WKPICS suggests that QA reports are produced for each stock. For each stock at the regional level it is possible to describe the contribution each country makes to the total catches (discards and landings) of that stock and the proportion caught or landed within each strata of the national sampling frame - for vessel groups for sampling at sea or port groups for sampling on shore.

For each stratum, a description of the sampling design, implementation and sampling successes with reference to best sampling practise (see Annex 3) can be provided. Simple metrics can be produced to effectively qualify these sampling schemes in relation to that particular stock. Considering the number of stocks, the number of countries and the complexity of the sampling strategies, automated reporting from the RDB is desirable (see recommendations).

These reports would be reviewed and qualified by the Regional Coordination Groups and would then be used by expert groups, stakeholders and managers (e.g. ICES stock assessment WGs, STECF, EU).

Producing these reports would be a four-stage process (Figure 4.2) – the basic framework could come from the RDB after the national data are uploaded. Some aspects of the report are going to be common across all stocks, for example the metrics relating to overall sampling effort and the summary of the sampling design. The summary of the sampling design in relation to best practice on each report is dependent on the metadata for each sampling scheme being available within or through the RDB.

1. Country uploads the data
2. QA report provided to RCG which reviews the QIs
3. With reference to the country, an assessment of sampling biases is made and the report signed off by the RCG
4. The reports are made available for EGs and review bodies.

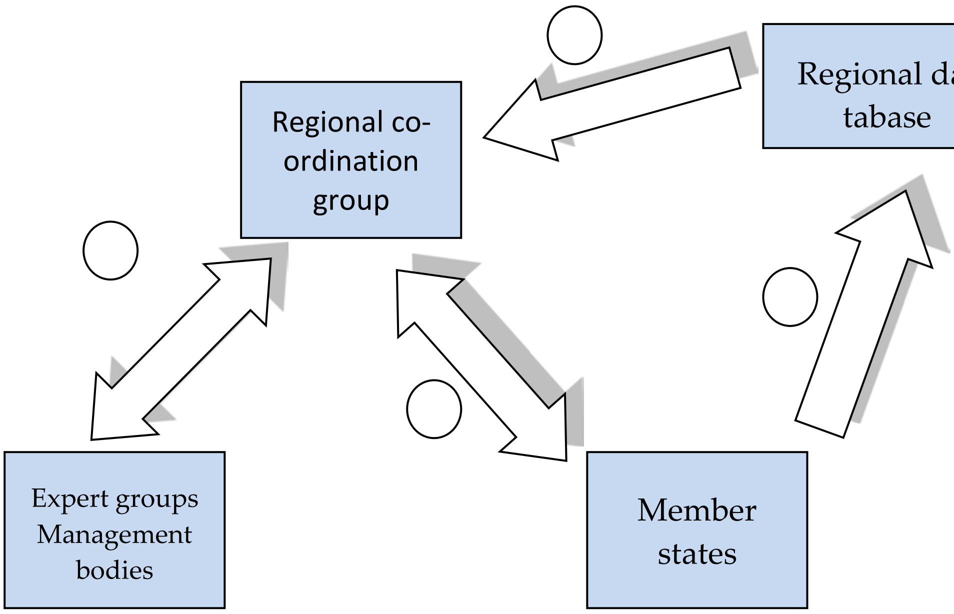

Figure 4.2. Four-stage process for quality assuring fishery sampling schemes relative to stock estimates

Table 4.1 and 4.2 are examples of QA reports proposed by WKPICS2 for at-sea and on-shore sampling schemes respectively, summarising the hierarchy and quality indicators for a given stock and highlighting potential major biases in the sampling design. The tables could be added as an annex to the stock assessment report of a given year.

The quality indicator overview tables are probably not exhaustive and unlikely to fit all stocks. They should be understood as a suggestion and backbone for further improvements. Given the particularities of each region or the stocks within a region, RCGs and/or assessment groups can and should develop the quality indicators further according to their specific needs and concerns. For instance, extra quality metrics may be added for stocks where illegal landings or consistent area or species misreporting is an issue, or metrics may be substituted by more appropriate ones or deleted if not useful. To have maximum impact, the quality indicator overview tables should be as concise as possible and focus on the major sources of bias and poor precision. Whatever metric is used, it should be an objective number or ratio that preferably can be automatically provided via the RDB.

It is clear from this exercise that work is still required in coming up with a definitive list of quality indicators. SGPIDS3 (2013) will be looking at reporting from sampling schemes and more work is suggested for resolving best-practice indicators.

WKPICS has not yet tested the real utility of this type of report. Therefore WKPICS propose that these reports be tested for some selected stocks at the RCGs. Once the utility of the reports is tested at the regional level and quality indicators defined, the next step would be to send the report to specific assessments WGs and have input from them about their utility and suggested improvements. Although WKPICS is recommending the implementation of this kind of report from the RDB we strongly believe we need to ensure that any development in reporting is useful and cost effective. Other way is risking to start a series of data collection in the RDB that may not be useful.

The following text provides a general description of each component of the suggested QA reports.

Contribution to total landings of stock

For any sampling scheme, at sea or on shore (Table 4.1 and 4.2), information for a given stock and a given year comes from several countries involved in the fisheries (country A, B, C, …). To determine the relative importance of each country for a given stock, the national contribution to stock landings is provided (this information should be available in the RDB).

At this point a simple traffic light response against each country for design and implementation can highlight any significant bias. These biases can be explained lower in the report. This flag will be set by the RCGs at the review stage after consultation with the country involved (see Stage 3, Figure 4.2 above).

Sampling scheme

As a central element, a description of the sampling scheme in relation to best practice should be provided, including a succinct overview of the sampling design, for example the type of sampling (random or non random), the primary sampling unit, the periodicity of sampling, contact protocol, etc. These should, at the very least, be a list of the key elements of a sampling scheme and should highlight how closely they follow best practice as given in Annex 3.

This information is not currently available in the RDB but if it was, the qualifiers in this table need to be simple and succinct and limited to a list of responses so that simple assessments of the schemes and comparison across the contributing countries can be made.

Quality indicators

SGPIDS (ICES 2012a) suggested a number of quality indicators for bias that could be used for at-sea sampling schemes:

1. Reporting of (i) the number of unique vessels in the total population, study population and realized samples, and (ii) the number of trips in the total population and study population relative to planned numbers of samples and the realized number of samples.

2. Non – response rates

3. "Goodness of fit" – i.e. how well the distribution and intensity of sampling matches the spatio-temporal distribution of fishing activities and catches by fleet stratum.

These QIs are equally pertinent to the on-shore sampling schemes and are represented in Tables 4.1 & 4.2.

It is important to evaluate the coverage of a sampling scheme in relation to the total population (catch), and how representative the vessels/ports in the frame are of the total population of vessels and ports, if there is a significant component of catch not accessible for sampling (e.g. at minor or inaccessible ports not included in the frame).

Within a country, the sampling frames are likely to be stratified into distinct fleet segments for sampling at sea, and ports or port groups for sampling on shore, contributing to varying extent to the total national catches for a given stock. To determine the importance of each fleet segment, the contribution to the national landing and discard estimates (if available) for the stock should be provided in the QA report (if the RDB can relate any samples and landings to a sampling frame and stratification then these figures could come from the RDB). Each fleet segment should be described, e.g. active or passive gear; or trawlers, gillnetters, purse seiners, beam trawlers, long-liners.

If the RDB is able to provide the data to calculate the precision of estimates for the stock at the required level of aggregation (domain), this could provide a 4th quality indicator.

4.3.2 QA report for at-sea sampling scheme

Table 4.1 provides suggested content for a QA report for at-sea sampling. The national at-sea sampling procedures may focus either on sampling discards only or on sampling of the entire catch (i.e. landings + discards), by an observer at sea, self reporting by fishers, or purchase of samples. Sampling on shore (Table 4.2) is exclusively dedicated to landings. This means that for a given stock, Table 4.1 can contain discards-only sampling from one country and total catch sampling (landings and discards) from another country. The landings component of the country with the discards-only sampling is provided through the on-shore sampling scheme. The country that is sampling landings and discards at sea may also have additional sampling of landings on shore, or may collect all data from sampling at sea. Potential changes in the sampling programs due to the designated discard ban were not considered by WKPICS2.

Quality indicator type 1 – Target and sampled population:

The ***Total number of vessels*** under the country flag in the given fleet is all active vessels that were in the fleet register in the given year. This metric displays the number of primary sampling units in a given fleet segment from which a country could theoretically have selected vessels to be sampled. A record could also be made of the total number of vessels falling within the sampling frame if there are segments of the fleet permanently excluded (e.g. vessels too small to be sampled), and the percentage of the catch represented by the sampling frame given.

The ***Number of trips sampled onboard*** of vessels is the number of trips that were sampled by an observer or observer team onboard of commercial vessels (2 observers on 1 trip are counted as 1 sampled trip). This metric directly shows the results of the national sampling efforts in terms of trips that were in fact sampled. A single trip is understood as an observer boarding a commercial vessel, sampling onboard and return to a harbor where the vessel is offloading the landings from this journey.

The ***Number of unique vessels sampled*** is the number of vessels that were only sampled once in the given year. When compared with total number of trips sampled, it indicates the extent to which trips are spread across different vessels or clustered within vessels. Repeat sampling of individual vessels could increase bias in estimates if vessel selection is non-random, particularly where there is a large variation between vessels in (e.g.) numbers of fish discarded or size compositions. In a sampling design where multiple trips on a vessel are not permitted, the number of unique vessels/number of trips onboard vessel = 1. A low number of unique vessels (e.g. a total of 30 trips in a year from only 5 different vessels) indicates repeat sampling of some or all vessels.

The ***Number of trips conducted by the fleet*** reflects the effort of the fleet segment. The ratio between the number of sampled trips and the total number of commercial trips gives an indication of the coverage of the sampling between the national fleets.

The ***Number of trips sampled where that stock occurred*** in the retained and in the discards provides a simple metric that relates the sampling effort to the frequency that that stock was sampled. There may be a seasonal aspect to the fishing which affects the proportion of trips in which that species occurred and the relative frequency that data is available for that stock but it may also provide some indication of how well the scheme provides data for that stock. This is could be considered an indicator of the coverage of the stock in multi-species fisheries (demersal).

The **age key quality indicator (mean number of age sampled per sampled trip)** was suggested as an indicator of the quality of the age sampling, i.e. the total number of otoliths taken from all sampled trips divided by the number of sampled trips. However a value of 100 age samples per trip from 30 trips in fleet 1 (Table 4.1) could result from anything between 3000 otoliths taken from a single trip, 500 otoliths taken from every 5th trip and zero otoliths from the others, to constant sampling of 100 otoliths from each trip. The number of trips for which otoliths were collected should be included as an indicator. The true effective sample size for age sampling will likely lie between the number of trips sampled for age and the total number of fish aged, but more likely to be closer to numbers of sampled trips due to clustering effect. If it is not feasible to calculate effective sample sizes directly, the minimum quality indicator should be numbers of trips sampled for age and number of fish aged, as proposed by ICES PGCCDBS in 2011.

Quality indicator type 2 – Non-response rates:

The **non-response rate** is the proportion of all attempted contacts that ultimately failed to provide a sample, for whatever reason, whilst the **refusal rate** is the proportion of vessel skippers who, having been successfully contacted, ultimately failed to allow the observer to go on board to obtain the sample (SGPIDS: ICES 2012a).

Both values should be calculated by the country following the procedure given in SGPIDS2 Table 2.5 (ICES 2012a). The use of an equal approach in the calculation is crucial to ensure full comparability between the national estimates. The estimates mainly highlight how easy it was to get trips following the sampling design described in the QA report (e.g. random or quota sampling) and whether or not there is a major problem in getting observer trips from a fleet. If refusal rates are considered alarmingly high by the country, explanations can be given as part of the sampling design description. For instance, refusal rates of 40% are high but tests (e.g. using spatial coverage, catch composition) might suggest that the sampled vessels still provide representative samples of the fleet.

Quality indicator type 3 – "Goodness of fit":

The goodness of fit indicators consist of metrics that illustrate the spatial and temporal coverage of the sampling relative to fleet activity and catches as a whole, and indicates whether the selection process leads to non-representative coverage of vessel size classes etc. These metrics are likely stock specific. Possible metrics to be used are given in SGPIDS2 report, 2.5 (ICES 2012a; see also WKPRECISE: ICES 2009). Stock coordinators, stock assessors or RCGs could suggest and agree about standard metrics considered to be useful for a given stock in a region.

Quality indicator type 4 – Precision estimates:

The advantage of a probability-based sampling design is that estimators of precision can be developed using the formulae given in Section 3 or using bootstrapping methods. Relative standard errors (RSE: standard error divided by mean) for key variables such as discard volumes, catch at age etc. should be calculated to indicate the quality of data in terms of precision. As a minimum where RSE estimates are not available, the effective sample size, or (minimally) the numbers of fishing trips sampled for length or age should be provided (as in Quality indicator 1).

4.3.3 QA report for on-shore sampling schemes

Table 4.2 for onshore sampling was developed based on Table 4.1 (at-sea sampling) but differs in the description of the sampling design, strata (e.g. ports or group of ports such as largest ports only) and quality indicators. The importance of the member states landings and the importance of the different port categories are given to allow estimation of the relative significance of the samples.

Quality indicator type 1 – Target and sampled population:

The **Number of ports** is the total number of on-shore access points for sampling (e.g. ports or port groups), by major stratum (e.g. port size) for each country where commercial landings occur. This metric indicates from how many ports could theoretically have been sampled. In the example, port group 1 consisted of the 3 major ports of the country that contributed 75% to the total landings of the country. A further 12 minor ports contributing 25% of landings were excluded from the frame and not sampled. Further explanations about how ports were stratified for the sampling can be provided under the sampling design section.

The **Number of sampled ports** indicates how many of the possible ports were in fact visited; in the example all 3 major ports were visited, and no visits were made to the 12 minor ports which contributed 25% to the national landings of the given species.

The **Number of visits** records the number of sampling trips during the last year (days, where the PSU is port x day). This indicates the amount of effort applied based on the sampling design given in the description.

The **Number of vessels sampled** records the number of total number of vessel landings that were sampled during the port visits in the given year.

The **Number of unique vessels sampled** is the number of vessels from which landings were sampled only once in the given year (see Section 4.2.2). This number cannot be larger than the number of vessel landings sampled.

The sampling quality indicators provide information on the quantity of samples collected for length or age (otoliths, scales or other structures are not always available or even collected for some stocks). This should record both the number of trips from which length or age data were collected for the stock, and the total number of fish measured or aged. Effective sample sizes should ideally be calculated although this may be difficult, especially with graded catches and these figures may not be readily available from the RDB.

Quality indicator type 2 – Response rates:

Skippers or port master may refuse access to landings to sample – this should be recorded similar to the refusal rate of at-sea sampling programs. If components of the landings are barred or inaccessible due to some intervention by the industry or by other landing practices, this causes a departure from the design of the programme.

Quality indicator type 3 – "Goodness of fit":

The use of Quality indicators type 3 would require adapting the quality indicators of spatio-temporal coverage to the characteristics of an on-shore sampling program. For calculations, one may refer to the COST tool[2] and the SGPIDS 2012 report (ICES 2012a).

Quality indicator type 4 – Precision estimates:

If the effective sample size for an individual parameter can be calculated, that number can be used as quality indicator. It is a good and simple quality indicator for the information content of the sample when properly used. The precision as measured by the relative standard error (RSE) of the final estimates of landings at length or age for stock assessments can be used as a quality indicator.

[2] http://wwz.ifremer.fr/cost

Table 4.1. Example of quality assurance report for regional assessment data from at-sea sampling

	AT-SEA-SAMPLING									
	Stock - Species - Area - Year (e.g. Cod - North Sea - 2011)									
		Country A				Country B		Country C		...
		Design				Design				
		Implemention				Implemention				
	Importance: Contribution to stock landing	40				60				
	Sampling / design effect/diagnostic for randomness... (Description according to best practice)									
	Sampling design	probability based discard sampling				quota sampling of catches				
	Primary sampling unit	Vessel				Trip				
	Sampling frame	quarterly vessel list				annual vessel list				
	Periodicity	ca. 1 sample per week during fishing season				1 sample per month				
	Contact protocol	yes				no				
	Sampling manual available	under preparation								
	...									
	Strata from the sampling frame	Fleet 1	Fleet 2	Fleet 3	...	...				
		e.g. active gear (Trawler)	e.g. passive gear	e.g. seine netters						
	Importance: Contribution to national landing	75%	20%	5%						
	Importance: Contribution to national discards	95%	1%	4%						
	Quality indicator									
1	Total number of vessels in the fleet	60	300	5						
	Number of trips sampled onboard of vessels	30	20	0						
	Number of unique vessels sampled	5	17	0						
	Total number of trips conducted by the fleet	1000	8000	6						
	Number of trips sampled where stock occurred in the discards	25	20	0						
	Number of trips sampled where stock occurred in the landings	11	20	0						
	Age key quality indicator (e.g. Mean number of age samples per trip sampled)	100	50	0						
2	Non-response rate	50%	3%	not determined						
	Industry decline (refusal rate)	40%	2%	not determined						
3	Goodness of fit									
	Bias 1: Spatio-temporal coverage	tested and considered all right								
	Bias 2: Vessel selection	smaller vessels rejected observers								
	Bias 3: ...	comment								
4	Precision levels of e.g. parameter a, b, ...									
	e.g. CV, variance, relative sampling error									
	e.g. Input data for XSA model:									
	maturity at age									
	stock weight									
	catch weight									
	catch at age									

Table 4.2. Example of quality assurance report for regional assessment data from on-shore sampling

	ONSHORE SAMPLING					
	Stock - Species - Area - Year (e.g. Cod - North Sea - 2011)					
		Country A		Country B	Country C	...
		Design		Design		
		Implemention		Implemention		
	Importance: Contribution to stock landing	40		60		
	Sampling (Description according to best pratice)					
	Sampling design	probability based landings sampling		quota sampling		
	PSU	Port		Port		
	Sampling frame	quarterly port list		1 annual list		
	Periodicity	quarter		summer, winter		
	Contact protocol	yes		no		
	Sampling manual available	under preparation		no		
	...					
	Strata from the sampling frame	Port group 1 (e.g. large ports sampled)	(e.g. minor ports - not sampled)	...		
	Importance: Contribution to national landing	75%	25%			
	Quality indicator					
1	Number of ports in the country	3	12			
	Number of ports sampled	3	0			
	Number of visits	12	0			
	Number of vessels sampled	36	0			
	Number of unique vessels sampled	28	0			
	Sampling quality indicators					
	e.g. Mean number of age samples per vessel	50	0			
	e.g. Number of otoliths removed	1800	0			
	e.g. Number of fish length measured	12000	0			
2	Refusal rate (e.g. vessel level)	8%	not determined			
3	Goodness of fit					
	Bias 1: Spatio-temporal coverage	tested and considered all right				
	Bias 2: Vessel selection	no concerns				
	Bias 3: ...	comment				
	...					
4	Precision levels of e.g. parameter a, b, ...					
	e.g. CV, variance, relative sampling error					
	e.g. Input data for XSA model:					
	maturity at age					
	stock weight					
	catch weight					
	catch at age					

4.3.4 Further comment on estimates of precision as quality indicators

In statistical fish stock assessment models the uncertainty of the population and mortality estimates (95% confidence limits, for example) are derived from 1) the data uncertainty and 2) the model/sampling design uncertainty. The uncertainty of the estimates from the assessment can be decreased by improving the precision of the input data through better sampling design (by using stratified random sampling, for example) and by increasing the sample size. Sampling can be optimized over strata and levels of sampling hierarchy by using sizes of variance components and costs at different sampling levels (see Section 2.4).

However, these two necessary components are not enough when determining the minimum sample sizes needed and signing off a statistical sound sampling plan. The missing component is the maximum acceptable uncertainty in the final stock assessment estimates (e.g., spawning stock biomass or fishing mortality (F)) used in harvest control rules.. In the other words: what is the minimum precision (largest relative standard errors) in the final estimates of the stock assessment that is acceptable before the conclusions drawn and decisions becomes erroneous. The target precision of the final estimate should be determined by discussing with the decision makers and end users (in general: with most important interest groups). The precision must be accepted by the decision makers and other end users. The scientist must help decision makers to find out the effective size by learning from simulations and scenarios in a what-if-manner.

After the precision has been decided a statistically sound and cost effective sampling design can be planned. If the targeted precision of the final estimate is not known, the correct sample size cannot be calculated and sampling correctly planned which could result in under or over sampling.

Without the target precision of the final estimate, allocating resources and sampling effort to meet the desired precision in the most cost-effective manner cannot be done at a national or regional scale.

The process and thinking presented above can be applied also to the stocks with "non-statistical" assessment models like Adapt or XSA, in which estimates such as catch at age and weights at age are treated as exact without any data uncertainty added to the assessment model.

The effect size is stock related and its determination process must be gone through for each stock and assessment. The sampling design and sample size need to be determined in a manner that can be anticipated to cover a large range of years i.e. reasonable year-to-year variation. The sampling plan must not be changed yearly.

5 Glossary[3]

DOMAIN: In many surveys, estimates are desired for a number of classes of the target population. Such subpopulations of special interest have been given the name domain of study by the U.N. Subcommision of sampling in 1950 (Cochran 1977, p. 34). Data to identify a domain of interest such as a métier are typically obtained after samples have been collected. This is often done by cross-classifying the sample data from PSUs by one or more predictors such as gear, target species, statistical area, and quarter. For domains that are not strata, sample sizes cannot be specified in advance. Special analysis techniques of domain estimation will be required (e.g., Lumley 2010, p. 32).

FLEET: A physical group of vessels sharing similar characteristics in terms of technical features and/or major activity.

FISHERY: A group of vessel voyages targeting the same (assemblage of) species and/or stocks, using similar gear, during the same period of the year and within the same area.

FLEET SEGMENT: a group of vessels with the same length class (LOA) and predominant fishing gear during the year, e.g. according to the Appendix III of the EU-DCF. Vessels may have different fishing activities during the reference period, but are classified in only one fleet segment.

MÉTIER: A group of fishing operations targeting a similar (assemblage of) species, using similar gear, during the same period of the year and/or within the same area and which are characterised by a similar exploitation pattern. The catches for such a sub-population of fishing operations in a fishery (domain) cannot generally be sampled with known probability since a list of PSUs is not available in advance. Estimates of catch characteristic for a métier (domain) are therefore often based on stratification after selection of PSUs (post-stratification.) EU Commission Decision 2008/949/EC (DCF) provides detailed requirements for Member States to collect economic data by fleet segment, and biological data by fleet métier.

MULTI-STAGE SAMPLING: A sampling method in which the target population (e.g., total fleetwide national landing) in principle first is divided into a number of groups or primary units from which samples can be drawn (sampling stage 1). Examples of first stage units (primary sampling units) are trips, vessels, and port-days. The fish in each primary sampling unit in the sample can then be sub-sampled in the secondary stages, and so on.

PRIMARY SAMPLING UNIT: A sampling unit in the first stage in multi-stage sampling is called a primary sampling unit. Primary sampling units in the most common catch sampling schemes can be trips, vessels, or site-days.

SAMPLE DESIGN: The totality of instructions, protocols, and rules that govern a sampling method.

SAMPLING FRAME: In statistics, a sampling frame is the list of sampling units or device from which a sample is drawn. The sampling frame comprises all the primary sampling units and any stratification of these, and may be based on a vessel registry or list of ports.

SAMPLING UNIT: In order to take a sample from a population, the target population must consist of, or be divided into non-overlapping parts (units). Sampling can then be conducted by selecting units according to a defined sampling scheme. These units are called sampling units in the survey sampling literature. The units that can be selected in

[3] Many of the statistical terms are based on Elsevier's Dictionary of Biometry

catch sampling schemes are typically groups (clusters) of fish, such as the cluster of fish in a landing from a fishing trip, or cluster of fish caught in a fishing operation.

STRATIFICATION: The advance decomposition of a **finite population** of sampling units of size N into k non-overlapping subpopulations (strata) of size N_i.

STRATIFICATION AFTER SELECTION: If a simple random sample is taken from a **finite population** of sampling units of size N the sample may be treated as a stratified sample during the analysis if the post-strata sizes N_i are known. Stratification after selection (post-stratification) is usually applied if the strata to which the selected sampling units belong are only known after the sample is taken. This is often the case for métiers. Standard stratified estimators cannot generally be applied when métiers cuts across strata.

6 References

Aanes, S., and M. Pennington. 2003. On estimating the age composition of the commercial catch of Northeast Arctic cod from a sample of clusters. ICES J. Mar. Sci. 60:297-303.

Aanes and Hirst 2012. Comparing design based and model based estimation of catch at age. Power point presentation available at the WKPICS2 Sharepoint Site.

Barbeaux, S.J., S. Gaichas, J.N. Ianelli, and M.W. Dorn. 2005. Evaluation of biological sampling protocols for at-sea groundfish observers in Alaska. Alaska Fish. Res. Bull. 11(2): 82-101.

Box, G. E. P., Hunter, W. G., and Hunter, J. S. 1978. Statistics for Experimenters. John Wiley, New York.

Chih, C.-P. 2010. Incorporating effective sample sizes into sampling designs for reef fish. Fish. Res. 105:102-110.

Cochran 1977. Sampling Techniques. John Wiley & Sons.

Ebeling, M.W., Natale, F. & Doerner, H. (eds.) 2012. STECF, 2012 Review of Proposed DCF 2014-2020. Part 1 (STECF-12-07). JRC Scientific and Policy reports. Report EUR 25338 EN

Folmer, O. and M. Pennington. 2000. A statistical evaluation of the design and precision of the shrimp trawl survey off West Greenland. Fish. Res. 49:165-178.

Global Adult Tobacco Survey (GATS). 2010. Sample Weights Manual, Version 2.0. Atlanta, GA: Centers for Disease Control and Prevention.

Helle, K., and M. Pennington. 2004. Survey design considerations for estimating the length composition of the commercial catch of some deep-water species in the Northeast Atlantic. Fisheries Research 70:55-60.

Hirst, D.J., Aanes, S., Storvik, G., Huseby, R.B., and Tvete, I.F. 2004. Estimating catch-at-age from market sampling data using a Bayesian hierarchical model. Appl. Stat. 53: 1–14.

Hirst, D.J., Storvik, G., Aldrin,M., Aanes, S. and Huseby, R.B 2005 Estimating catch-at-age by combining data from different sources. Can. J. Fish. Aquat. Sci. 62: 1377–1385.

Hirst, D.J. , G. Storvik, H. Rognebakke, M. Aldrin, S. Aanes and J.H. Vølstad. 2012. A Bayesian modelling framework for the estimation of catch-at-age of commercially harvested fish species. Can. J. Fish. Aquat. 69: 1–13. doi:10.1139/CJFAS-2012-0075

Horwitz, H.G. and D.J. Thompson. 1952. A generalization of sampling without replacement from a finite universe. J. Am. Stat. Assoc. 47: 663-685.

Hjellvik, V., O.R. Godø, and D. Tjøstheim, 2002. The measurement error in marine survey catches: the bottom trawl case. Fish. Bull. 100:720-726.

ICES. 2008. Report of the Workshop on Methods to Evaluate and Estimate the Accuracy of Fisheries Data used for Assessment (WKACCU), 27–30 October 2008, Bergen, Norway. ICES CM 2008\ACOM:32. 41 pp.

ICES, 2009. Report of the Workshop on methods to evaluate and estimate the precision of fisheries data used for assessment (WKPRECISE), 8-11 September 2009, Copenhagen, Denmark. ICES CM 2009/ACOM:40. 72 pp.

ICES 2010. Report of the ICES Workshop on methods for merging metiers for fishery-based sampling (WKMERGE). ICES CM 2010 / ACOM:40

ICES 2011a. Report of the Workshop on Practical Implementation of Statistical Sound Catch Sampling Programmes, 8-10 November 2011, Bilbao, Spain. ICES CM 2011/ACOM:52

ICES. 2011b. Report of the Study Group on Practical Implementation of Discard Sampling Plans (SGPIDS). ICES 2011 /ACOM:50

ICES 2012a. Report of the Study Group on Practical Implementation of Discard Sampling Plans (SGPIDS). ICES 2012 /ACOM:50

ICES 2012b. Report of the Working Group on Recreational Fisheries. ICES CM 2012/ACOM:23.

ICES 2012c. Report of the Planning Group on Commercial Catches, Discards and Biological Sampling (PGCCDBS 2012). 30 January–3 February 2012, Rome, Italy. ICES CM 2012/ACOM:50.

Kish, L. (1965). Survey Sampling. John Wiley & Sons.

Kish and Anderson 1978. Multivariate and Multipurpose Stratification. Journal of the American Statistical Association 73: 24-34.

Korn E. and Graubard B. 1999. Analysis of health surveys. John Wiley & Sons.

Lehtonen, R. And E.J. Pahkinen. 1996. Practical Methods for Design and Analysis of Complex Surveys. 2nd Edition. John Wiley & Sons.

Lumley, T. 2010. Complex Surveys. A guide to Analysis using R. John Wiley & Sons

Miller, T. J., Skalski, J. R., and Ianelli, J. N. 2007. Optimizing a stratifield sampling design when faced with multiple objectives. ICES Journal of Marine Science, 64, 97–109.

Pennington, M., L.-M. Burmeister and V. Hjellvik. 2002. Assessing the precision of frequency distributions estimated from trawl-survey samples. Fish. Bull. 100:74-81.

Pennington, M., and K. Helle. 2011. Evaluation of the design and efficiency of the Norwegian self-sampling purse-seine reference fleet. ICES J. Mar. Sci. 68:1764-1768.

Pennington, M. and Vølstad, J. H., 1994. Assessing the effect of intra-haul correlation and variable density on estimates of population characteristics from marine surveys. Biometrics 50:725-732.

Skinner, C.J., Holt, D., and Smith, T.M.F. (eds.), (1989), Analysis of Complex Surveys, New York: Wiley.

Særndal, C.E., B. Swensson, and J. Wretman. 1992. Model Assisted Survey Sampling. Springer-Verlag, New York.

Thompson 1992. Sampling. John Wiley & Sons

Vølstad, J.H, W.R. Richkus, S. Gaurin, and R. Easton. 1997. Analytical and Statistical Review of Procedures for Collection and Analysis of Commercial Data Used for Management and Assessment of Groundfish Stocks in the U.S. Exclusive Economic Zone Off Alaska. Prepared for the U.S. Department of Commerce, National Marine Fisheries Service, Alaska Fisheries Science Center, Seattle, Washington. 172 pp.

Annex 1: List of participants

Name	Address	Email	Note
Mike Armstrong Chair	United Kingdom	mike.armstrong@cefas.co.uk	
Jon Helge Vølstad Chair	Norway	jon.helge.voelstad@imr.no jonhelge@imr.no	
Sofia Carlshamre	Sweden	sofia.carlshamre@slu.se	
Ana Claudia Fernandes	Portugal	acfernandes@ipma.pt	
Didzis Ustups	Latvia	didzis.ustups@bior.gov.lv	
Gjert Endre Dingsør	Norway	gjert.endre.dingsoer@imr.no	
Edwin van Helmond	Netherlands	Edwin.vanHelmond@wur.nl	
Kirsten Birch Håkansson	Denmark	kih@aquadtu.dk	
Hans Gerritsen	Ireland	hans.gerritsen@marine.ie	
Joel Vigneau	France	joel.vigneau@ifremer.fr	
Jon Elson	United Kingdom	jon.elson@cefas.co.uk	
Josefine Egekvist	Denmark	jsv@aqua.dtu.dk	
Liz Clarke	United Kingdom	E.D.Clarke@marlab.ac.uk	
Marie Storr - Paulsen	Denmark	msp@aqua.dtu.dk	
Michael Pennington	Norway	michael.pennington@imr.no	
Sofia Carlshamre	Sweden	mailto:sofia.carlshamre@slu.se	
Sondre Aanes	Norway	sondre.aanes@imr.no sondre.aanes@nr.no	
Mika Kurkilahti	Finland	mika.kurkilahti@rktl.fi	
Uwe Krumme	Germany	uwe.krumme@vti.bund.de	
Estanis Mugerza	Spain	emugerza@azti.es	
Katja Ringdahl	Sweden	katja.ringdahl@slu.se	
Jose Rodriguez	Spain	jose.rodriguez@st.ieo.es	
Gudmundur Thordarson	Iceland	gudthor@hafro.is	
Beatriz Guijarro	Spain	beatriz@ba.ieo.es	By correspondance
Els Torreele	Belgium	els.torreele@ilvo.vlaanderen.be	By correspondance
Alastair Pout	United Kingdom	a.pout@marlab.ac.uk	By correspondance
Sieto Verver	Netherlands	sieto.verver@wur.nl	By correspondance

Annex 2: Agenda

Tuesday November 6

9:00 – 09:30 Introduction and Workshop Objectives [Mike Armstrong]

9:30 - 10:30 Review and clarification of basic concepts from WKPICS1 for design and practical implementation of statistically-sound catch sampling schemes for large-scale and small-scale fisheries [Jon Helge Vølstad plus discussion].

10:30 - 11:00 Coffee Break

11:00 - 12:30: Start breakout groups for each ToR:

ToR(a): Regional Coordination of catch sampling schemes (by RCM area)

ToR(b): Develop guidelines for design-based and model-assisted estimation methods (for means, proportions, totals) including precision estimation.

ToR(c): Develop and define quality indicators and levels for onshore and offshore sampling schemes and advise on revisions to the WKACCU score cards to accommodate them. [to be dealt with by ToR (a) subgroup]

Lunch 12:30-13:30

13:30 – 18:00 Sub Groups – discussion and report drafting

Wednesday, November 7

9:00 – 12:30. Presentation by ICES on Regional Data Base; Sub Groups– discussion and report drafting

Lunch 12:30-13:30

13:30 – 16:00: Sub Groups– discussion and report drafting

16:00 – 18:00:` Report back by Sub Groups

Thursday, November 8

9:00 – 12:30. Sub Groups– discussion and report drafting

Lunch 12:30-13:30

13:30 – 16:00: Sub Groups– discussion and report drafting

16:00 – 18:00:` Report back by Sub Groups; review of text drafted

Friday, November 9

9:00 – 12:30. Review text drafted by sub groups (continued).

General meeting closes at 12:30

Annex 3: Guidelines for best practice in catch sampling schemes

DOCUMENTATION OF SAMPLING DESIGN, PERFORMANCE OF SAMPLING AND PRODUCTION OF ESTIMATES			
Process that need to be described	Best practice	Comment	Bad practice
Target population	The target population needs to be identified and described. Access to the target population for sampling purposes need to be analysed and documented.		
Primary sampling units (PSUs)	Choice of PSUs should be identified, justified and documented. PSUs could be trips, vessels*time or sites*time (harbours, markets, access points). Size of PSUs should be documented	If PSU is something else than trip, vessel or site the choice need to be thoroughly explained.	
Sampling frame	The sampling frame (list of PSUs) should be a complete list of non-overlapping PSUs. The sampling frame should ideally cover the entire target population.	If it is not possible to cover the entire target population with the sampling frame it is good practice to clearly describe how large the excluded part of the population is and the reason for excluding it.	To exclude large parts of the target population in an ad-hoc way.
Stratification of the sampling frame	Strata should be well defined, known in advance and fairly stable. Clear definitions and justifications of strata should be available. One PSU can only be in one stratum. The min-	If the desired minimum number of samples per stratum is not analytically assessed, the choice needs to be justified and	To over-stratify (few or no samples in each strata) the sampling schemes. Over-stratification results in increased risk for

	imum number of samples within a stratum is dependent on objective, PSU and variance and needs to be calculated. The number of samples within a stratum needs to be justified, in particular if it is below 10.	described. Care needs to be taken to avoid overstratification.	bias, particularly for ratio estimates, and a need to impute data.
Distribution of sampling effort	The way sampling effort is distributed between strata needs to be described. In accordance with best practice, this can be based on analysis of variance or just distributed proportionally. The different sampling inclusion probabilities/weighting need to be documented.	If other methods, such as expert judgment are used, this should be explained and justified.	
Sample selection procedure	In accordance with good practice, the selection of PSUs to sample should be done in a controlled way allowing for estimation of sampling inclusion probabilities for the different samples. In principal this mean that samples shall be chosen randomly (probability based sampling). Random sampling can be either simple random sampling or systematic random sampling. The selection procedure needs to be justified and described	If it is impossible to use probability-based sampling, the samples need to be thoroughly validated for how representative they are. This process need to be described. If a non-probability based sampling design is applied, this needs to be accounted for in the estimation process (e.g model based estimations). This needs to be thoroughly explained. For	Ad-hoc based sampling, without proper documentation to allow estimation of bias, where the sampling inclusion probabilities cannot be estimated.

		small-scale fisheries where there is no census information on the target population, the only way to sample in accordance with good practice is randomly.	
Hierarchical structure in the sampling	All the levels in the hierarchical structure of the sampling scheme need to be documented. Sampling should be random at all levels. Sampling probabilities should be worked out at each level, and information for this needs to be collected (e.g number of boxes)		Failure to account for the different levels of sampling units in the design and estimation processes. (Risk for bias as well as hiding true variation)
Protocol for selection of samples at lower sampling levels (SSU, etc.)	Such protocols should exist in a national repository		
System to monitor performance of sampling schemes - Quality Indicators	Non-response rates should be recorded. Precision of estimates (relative standard error) should be calculated, where relevant. Effective sample size (or appropriate proxy such as number of vessels or trips sampled) should be calculated and recorded.		
Documentation of raising/weighting procedure for national estimates	Data analysis methods should be fully documented, covering: (1) how the multi-stage sample selection is accounted for in the raising/weighting procedures; (2) ancillary information (for example from fleet census		

	data), that is used to adjust sample weights to correct for any imbalance in samples compared to the population; (3) methods of adjustment for missing data and non-responses.		

Annex 4: Summary of the national sampling programmes

Survey questions 1 to 6

1. Probability based - Is the sampling probability-based and if so, how? e.g. random, systematic-random, opportunistic)
2. Sampling frame - What is the sampling frame (e.g. vessel*week, port*day etc).
3. Stable strata - Is the sampling frame stratification stable over the year (e.g. vessel size, week(or is there an attempt to stratify by characteristics that change over time (e.g. fishing area, metier)?
4. Optimisation - What methods are used to optimise the allocation of sampling effort (e.g. across survey strata).
5. Overall target prec. - Do you have a target precision for the overall estimates?
6. Stratum target prec. - Do you have a target precision for the strata or domains?

Scheme	Ctry	Sub-scheme	1. Probability based	2. Sampling frame	3. Stable strata	4. Optimization	5. Overal target prec.	6. Stratum target prec.
Onshore - Large Scale Fleet	Basq ctry	onshore + coastal	opportunistic	none		proportional to effort	DCF	DCF
	DEN	demersal, pelagic + shellfish	opportunistic	none	Yes	none	DCF	no
	UK-E	demersal, pelagic + shellfish	systematic random	bus-route	Yes	fishing activity & resources	DCF	yes
	FRA	commercial species	Systematic non-random (target = n trips /month)	List of ports + gear type	Not stable (Season + Fishery)	kind of Neyman's allocation	DCF	no
	IRE	demersal + shellfish	opportunistic	none	no (fishing area)	proportional to landed weight	DCF	DCF
	LVA	demersal+ pelagic	opportunistic	Trip *quarter	yes	aiming for complete coverage	DCF	DCF
	NLD		opportunistic	none		main auctions	DCF	no
	NOR		systematic random	ports	yes	?	No	No
	UK-S	demersal	yes	market day	yes	adhoc: roughly proportional to landed weight	DCF	DCF
	UK-S	shellfish	opportunistic	none	no (fishing area)	none	DCF	DCF
	UK-S	pelagic	intercept sampling of known vessels	fleet	yes – but seasonal	aiming for complete coverage	DCF	DCF
	ESP	ICES dem + pel	Opportunistic	none		prop to landings	DCF	no
	ESP	Med dem + pel + shell	opportunistic	month * port		prop to landings	DCF	no
	SWE	dem + pel	opportunistic	none	yes	none	DCF	DCF
At sea - Large Scale Fleet	Basq ctry		opportunistic	none		proportional to fishing effort	DCF	DCF
	BEL		opportunistic	none		effort	DCF	DCF
	DEN	demersal + shellfish	yes RANDOM	vessel list	No fishing area and metier	prop to #trips	DCF	no
	UK-E		Random	Vessel x quarter	Predominant fishing activity in previous yr same Q.	proportional to effort, target sample size	DCF	DCF
	FRA	All commercial	Random	List of vessels	Not stable (Season + fishery)	yes - ref given	YES	no
	GER		yes			?	DCF	DCF
	GER	NS/NEA demersal and pelagic	opportunistic	vessel lists otter, freezer, beam trawler, ...	no, (metiers*quarter)	proportional to last years fishing effort	DCF	DCF
	GER	Baltic demersal	yes RANDOM	vessel lists (active,	Fairly stable (area, quarter,	proportional to last years fishing	DCF	DCF

1. Probability based - Is the sampling probability-based and if so, how? e.g. random, systematic-random, opportunistic)
2. Sampling frame - What is the sampling frame (e.g. vessel*week, port*day etc).
3. Stable strata - Is the sampling frame stratification stable over the year (e.g. vessel size, week(or is there an attempt to stratify by characteristics that change over time (e.g. fishing area, metier)?
4. Optimisation - What methods are used to optimise the allocation of sampling effort (e.g. across survey strata).
5. Overall target prec. - Do you have a target precision for the overall estimates?
6. Stratum target prec. - Do you have a target precision for the strata or domains?

Sch eme	**Ctry**	**Sub-scheme**	**1. Probabil-ity based**	**2. Sam-pling frame**	**3. Stable strata**	**4. Optimization**	**5. Overal target prec.**	**6. Stra-tum target prec.**
				passive)	fleet segment)	effort		
	GER	Baltic pe-lagic	Systematic random	Ports * week	Fairly stable (area, quarter, fleet segment)	proportional to last years fishing effort	DCF	DCF
	IRE	demersal + shellfish	opportunistic	none	no, (metiers)	proportional to effort	DCF	DCF
	LVA	observers	opportunistic	none		none	DCF	DCF
	NLD	self-sampling	opportunistic	none		none	DCF	no
	NLD	observers	opportunistic	none		none	no	no
	NLD	reference fleet				prop to ?,	no	no
	NLD	other	opportunistic	none		even over com-panies	no	no
	NOR		opportunistic	none		?	no	no
	NOR	reference fleet	systematic random	reference fleet	yes	?	no	no
	POL	>12m	kind of			target # trips as set in DCF plan	no	no
	PRT	Demersal + custaceans	opportunistic	Vessel list	Fairly stable	Prop to last year's fishing effort	DCF	DCF
	UK-S	demersal	yes	vessel list	yes	adhoc: divided between fleet segments for sample size reasons	DCF	DCF
	UK-S	nephrops	opportunistic	none		ad-hoc: min sample size for strata	DCF	DCF
	ESP	ICES all spp.	Opportunistic + random	Vessel * port	Not stable (season + fishery)	Prop to effort (number of trips)	dcf	dcf
	ESP	Med, de-mersal +shellfish	opportunistic	Month* port		ranking accord-ing to ??	DCF	no
	SWE	Dem + nep + pandalus	Random	vessel list	no, area and fishery	min sample size for strata	no	no
SSF	Basq ctry		opportunistic	none		prop to landings	no	no
SSF	FRA	All com-mercial	Simple ran-dom	List of groups of ports*days	Yes	prop to landings	DCF	no
SSF	ESP	Med, dem +shellfish	opportunistic	Month* port		Prop to landings	DCF	no

Survey questions 7 to 9

7. Quality indicators - What quality indicators do you use to evaluate bias?
8. Key estimates - What are the key estimates that are provided from this sampling scheme?
9. Estimators - What estimators are used to provide these estimates (e.g. ratio estimates)?

Sch eme	**Ctry**	**Sub-scheme**	**7. Quality indicators**	**8. Key estimates**	**9. Estimators**
Large	Basq ctry	onshore + coastal	non-response	landed w-at-age	?

7. Quality indicators - What quality indicators do you use to evaluate bias?

8. Key estimates - What are the key estimates that are provided from this sampling scheme?

9. Estimators - What estimators are used to provide these estimates (e.g. ratio estimates)?

Scheme	Ctry	Sub-scheme	7. Quality indicators	8. Key estimates	9. Estimators
	DEN	demersal, pelagic + shellfish	no	WG ests	ratio
	UK-E	demersal, pelagic + shellfish	spatial & temporal coverage	WG ests	ratio
	FRA	commercial species	WKACCU scorecard on occasion	WG ests	ratio
	IRE	demersal + shellfish	spatial & temporal coverage	WG ests	ratio
	LVA	demersal+ pelagic	no	WG ests	ratio
	NLD		no	WG ests	?
	NOR		for internal use only	WG ests	ratio
	UK-S	demersal	# trips sampled/total, coverage maps, temporal coverage, # lengths, # ages, sampling fractions, non-response rates, refusal rates	WG ests	ratio
	UK-S	shellfish	# trips sampled/total, coverage maps, temporal coverage, # lengths, # ages, sampling fractions	WG ests	ratio
	UK-S	pelagic	# trips sampled/total, coverage maps, temporal coverage, # lengths, # ages, sampling fractions, non-response rates, refusal rates	WG ests	ratio
	ESP	ICES	Number of sanples and individuals; spatial and temporal cover	LF	
	ESP	Med	no	species comp, LF	
	SWE	deml + pel	nothing formal	WG ests	ratio
At sea - Large Scale Fleet	Basq ctry		non-response rates, refusal rates	WG discard ests?	?
	BEL		refusal rates	WG ests	?
	DEN	demersal + shellfish	non-response rates, VMS cover	WG ests	?
	UK-E		non-response rates, refusal rates	?	ratio
	FRA	All commercial	non-response rates	WG ests	Ratio, model
	GER		non-response rates, non-random sampling elements	DCF + QI	?
	GER	NS/NEA dem+pel	non-response rates, refusal rates	DCF + WG ests	ratio
	GER	Baltic demersal	non-response rates, refusal rates, number of vessels per fleet (total, sampled, unique)	DCF + WG ests + QI	ratio
	GER	Baltic pelagic	non-response rates, refusal rates, number of vessels per fleet (total, sampled, unique)	DCF + WG ests + QI	ratio
	IRE	demersal + shellfish	spatial coverage	WG ests	ratio
	LVA	observers	no	DCF	ratio
	NLD	self-sampling	no	DCF	?
	NLD	observers	non-response rates, refusal rates	DCF + DPUE (n/hour)	?
	NLD	reference fleet	LF, no incorrectly sampled trips	DCF + DPUE (n/hour)	?
	NLD	other	spatial coverage	DCF + DPUE (n/hour)	?
	NOR		for internal use only	WG ests	ratio
	NOR	reference fleet	for internal use only	WG ests	ratio
	POL	>12m	no	DCF?	ratio?
	PRT	Demersal + custaceans	non-response rates, refusal rates	Catch (dis+lan) comp, volume, len, age	ratio
	UK-S	demersal	# trips sampled/total, # lengths, # ages, sampling fraction, non-response rates, refusal rates	WG discard ests	ratio
	UK-S	nephrops	# trips sampled/total, # lengths, # ages, sampling fraction	WG discard ests	ratio

7. Quality indicators - What quality indicators do you use to evaluate bias?
8. Key estimates - What are the key estimates that are provided from this sampling scheme?
9. Estimators - What estimators are used to provide these estimates (e.g. ratio estimates)?

Scheme	Ctry	Sub-scheme	7. Quality indicators	8. Key estimates	9. Estimators
	ESP	ICES all spp.	Number of samples, non-response rates	WG discard ests	ratio
	ESP	Med, demersal +shellfish	no	species composition, LF	
	SWE	Dem + nep + pandalus	non-response rate	WG discard ests, catch composition, length distr for all species	ratio
SSF	Basq ctry		non-response rates	catch estimates	?
SSF	FRA	All commercial	CV	Catch and effort estimates, species composition, length structure	Horvitz-Thomson
SSF	ESP	Med, deml +shellfish	no	species composition, LF	

Survey questions 10 to 14

10. Weighting follows sampling - Does the data analysis use weighting procedures that follow the original sampling design?
11. Precision estimate - What methods are used to estimate precision?
12. Weighting Q2- Does your estimation method attempt to estimate sampling probabilities reflecting the true probability of sampling occasions (e.g. market-day), with additional domain estimation where appropriate, or does it ignore some of the actual clustering (such as market-day) and assume trips are randomly selected within the domain of interest, treating these as the sampling strata. If you don't understand the question, please answer "don't know".[4]
13. Post-stratification - Which diagnostics and decision rules are used to decide on post-stratification or aggregation of data?
14. Software - What software is used for data exploration?

Scheme	**Ctry**	**Sub-scheme**	**10. Weighting follows sampling**	**11. Precision estimate**	**12. Weighting Q2**	**13. Post-stratification**	**14. Software**
Onshore - Large Scale Fleet	Basq ctry	onshore + coastal	?	IEO do it		sample size, gaps in ALK	own data base, COST
	DEN	demersal, pelagic + shellfish	yes	?	No	do not post-stratify	SAS, FishFrame, R
	UK-E	demersal, pelagic + shellfish	yes	analytic	No	sample size, raising factors	R, COST, Excel, SQL
	FRA	commercial species	stratified sampling	analytic & bootstrap	No	métiers, similarity of length compositions	COST
	IRE	demersal + shellfish	yes	analytic & bootstrap	no	sample size, gaps in ALK	R, COST, SQL
	LVA	demersal+ pelagic	yes	analytic & bootstrap	no	sample size	R,COST, Excel
	NLD		Yes	?		spatial & temporal coverage	SAS
	NOR		within str- no, across str - yes	bayesian mcmc		small sample size	R, ECA (R-package in programmed in C)
	UK-S	demersal	in prog	bootstrap	in progress	sample size	R, COST data structures
	UK-S	shellfish	in progress	bootstrap	in progress	sample size	R, COST data structures
	UK-S	pelagic	in prog	bootstrap	in progress	sample size	R, COST data structures
	ESP	ICES	Yes	Delta-method	No	Sample size	own data-base, Excel, R
	ESP	Med	yes	Delta-method	No	Sample size	own data-base, Excel
	SWE	dem + pel	yes	bootstrap		none	R, S+
At sea - Large Scale Fleet	Basq ctry		yes	?		N/A	own data base, COST
	BEL		no weighting	analytic & bootstrap		sample size (trips/Q)	R, COST, Access
	DEN	demersal + shellfish	no - raise by landed weight, should raise by # trips	bootstrap	No	do not post-strat	SAS, FishFrame, R
	UK-E		Not always; post- strat by gear	bootstrap	No	RF, gaps in ALK	R, COST, Excel, SQL
	FRA	All commercial	Stratified sampling	analytic, bootstrap	yes	métiers, similarity of length compositions and/or discard behaviour	COST

[4] This question was an attempt to get clearer answers to question 10. Answers suggest that it was poorly understood by respondents, and if the table format is adopted for future use by Regional Coordination Groups or other end users, Q10 & Q12 should be merged and clarified with examples of typical departures of analysis from sampling design, such as lumping of samples across strata without accounting for different sampling probabilities, or ignoring cluster sampling.

10. Weighting follows sampling - Does the data analysis use weighting procedures that follow the original sampling design?
11. Precision estimate - What methods are used to estimate precision?
12. Weighting Q2- Does your estimation method attempt to estimate sampling probabilities reflecting the true probability of sampling occasions (e.g. market-day), with additional domain estimation where appropriate, or does it ignore some of the actual clustering (such as market-day) and assume trips are randomly selected within the domain of interest, treating these as the sampling strata. If you don't understand the question, please answer "don't know".[4]
13. Post-stratification - Which diagnostics and decision rules are used to decide on post-stratification or aggregation of data?
14. Software - What software is used for data exploration?

Scheme	Ctry	Sub-scheme	10. Weighting follows sampling	11. Precision estimate	12. Weighting Q2	13. Post-stratification	14. Software
	GER		?	?		split trips by gear/area	PHP-database,SQL, Excel, Access, R
	GER	NS/NEA demersal and pelagic	no	analytic & bootstrap	don't know	spatial & temporal coverage	R, COST, SQL, Access, Excel
	GER	Baltic demersal	no	analytic & bootstrap	assuming random selection	split trips by gear/subdivision	PHP-database,SQL, Excel, Access, R
	GER	Baltic pelagic	no	analytic & bootstrap	assuming random selection	none	Excel, Access, dBase
	IRE	demersal + shellfish	yes	bootstrap	no	sample size	R, COST, SQL, Access, Excel
	LVA	observers		no	no	none	R,COST, Excel
	NLD	self-sampling	yes	?		spatial & temporal coverage	SAS
	NLD	observers	?	no		none	SAS & R
	NLD	reference fleet	?	no		none	SAS & R
	NLD	other	?	no		none	SAS
	NOR		within str - no, across str - yes	bayesian mcmc		small sample size	R, ECA (R-package in programmed in C)
	NOR	reference fleet	within str - no, across str - yes	bayesian mcmc		small sample size	R, ECA (R-package in programmed in C)
	POL	>12m	?	?		VMS data, last year's distribution pattern	COST, Excel
	PRT	Demersal + custaceans	?	bootstrap		Num days in trip	R, Excel
	UK-S	demersal	yes	bootstrap	yes	sample size	R, COST data structures
	UK-S	nephrops	no	bootstrap	no	sample size	R, COST data structures
	ESP	ICES all spp	yes	analytic	no	Sample size	Own database, R, COST
	ESP	Med, demersal +shellfish	yes	delta method	no	Sample size	own data-base, Excel
	SWE	Dem + nep + pandalus	yes	analytic & bootstrap		do not post-strat	R, Excel
SSF	Basq ctry		in progress	no		in progress	own data-base
	FRA	All commercial	yes	bootstrap	Yes	Métiers, fleet segment	dedicated software
	ESP	Med, dem +shellfish	yes	delta	no	Sample size	own data-base, Excel

Annex 5: WKPICS3: Terms of Reference for the next meeting

To be agreed at PGCCDBS 2013.

Annex 6. List of presentations and working documents

All presentations, Working Documents and national sampling scheme summaries are archived on the WKPICS SharePoint site.