

**Workshop report**

# **Quality Assurance Framework Subgroup Workshop: Training session on methodologies in the Handbook**

**Virtual training from Finland 4.-6.5.2021**

Heidi Pokki & Juha Heikkinen

## Contents

1	Introduction .....	3
1.1.	List of participants .....	3
1.2.	Agenda.....	4
2	Presentations .....	5
2.1.	History of the Handbook.....	6
2.2.	Training sessions on methodologies in the Handbook.....	9
2.3.	Methodological report as part of the work plan .....	24
3	Conclusions .....	28

# 1 Introduction

The discussion on improving methodological aspects of the economic data collection of fisheries started in SGECA-09-02. Since then, further need for sharing the best practises of the sampling strategies, estimation methods and quality assurance has been acknowledged by PGECON for many years (in 2014, 2016, 2017). The first workshop on statistical issues and thresholds was held in 2013 in Helsinki. The workshop proposed that a handbook for best practices in economic data collection should be commissioned and PGECON endorsed this proposition.

A Handbook on 'Methodologies on sampling designs and estimation methods for fleet and aquaculture economic data collection' was produced under WP 2 of the SECFISH project with the aim of strengthening regional cooperation for the collection of social and economic data of the fisheries sector (2017 - 2019). Work Package 2 aimed to harmonize the methodologies for sampling design and estimation methods by providing a practical manual based on the general theory of probability sampling.

In PGECON 2019 the following points were raised regarding the Handbook:

- The Handbook can be used as a reference for National Work Plans to justify the described methodologies.
- It was agreed that the handbook would be very useful, and that each MS should try to follow the suggested procedures, thus using the handbook as a reference. MS could then report back at PGECON 2020 with their user experience(s) and issues encountered, or there might be a separate workshop where the methodologies are explained in detail, and MS can be given the opportunity to work through them.

Furthermore, PGECON 2019 recommended that: 'a Quality Assurance Framework (QAF) subgroup workshop should take place to define the process of quality assessment and assurance and revise the guidelines of the methodological report (with reference to the Handbook). Then as an outcome, PGECON could provide recommendations and guidelines for AR evaluation EWG on how to improve quality evaluation of DCF data and to complement the currently existing quality evaluation procedures.'

The Quality Assurance Framework Subgroup Workshop: Training session on methodologies in the Handbook was originally planned for March 2020 as a physical meeting in Helsinki. Unfortunately, the workshop was postponed until 2021 due to the COVID-19 pandemic. The workshop was organized as a 3-day online training session on 4.-6.5.2021. This timing will allow the member states to incorporate the teaching from the Handbook in the national methodological reports for economic data collection which will be part of the national work plans (2022-2024) drafted in autumn 2021.

The workshop was held using Teams as the platform for disseminating course materials and setting up the online meeting. The practical exercises applying simulated data were conducted using R and R Studio. There were 30 participants from 11 member states, with the Joint Research Center and the European Commission attending the workshop. The list of participants is presented in section 1.1. and the agenda is presented in section 1.2.

## 1.1. List of participants

First name	Last name	Country	Organisation
Paolo	Accadia	Italy	NISEA
Jörg	Berkenhagen	Germany	Thünen-Institute of Sea Fisheries
Brian	Burke	Ireland	BIM

Suzana	Cano	Portugal	DGRM
Judy	Cassar	Malta	Department of Fisheries and Aquaculture
Christos	Danatskos	Greece	FRI- ELGO DIMITRA
Irina	Davidjuka	Latvia	Institute of Food Safety, Animal Health and Environment - BIOR
Adelbert	De Clercq	Belgium	Flanders Research Institute for Agriculture, Fisheries and Food (ILVO)
Monica	Gambino	Italy	NISEA
Juha	Heikkinen	Finland	Natural Resources Institute Finland (Luke)
Lina-Marie	Huber	Germany	Thünen-Institute of Sea Fisheries
Emmet	Jackson	Ireland	BIM
Edvardas	Kazlauskas	Lithuania	AIRBC
Andreas	Kotelis	Malta	Department of Fisheries and Aquaculture
Markku	Kärnä	Finland	Natural Resources Institute Finland (Luke)
Angelos	Liontakis	Greece	Agricultural Economics Research Institute-DEMETER, Hellenic Agricultural Organization
Stamatis	Mantziaris	Greece	Agricultural Economics Research Institute-DEMETER, Hellenic Agricultural Organization
Jurgen	Mifsud	Malta	Department of Fisheries and Aquaculture
Sarah	Neehus	Belgium	European Commission
Heidi	Pokki	Finland	Natural Resources Institute Finland (Luke)
Mika	Rahikainen	Finland	Natural Resources Institute Finland (Luke)
Evelina	Sabatella	Italy	Nisea
Andrew	Sciberras	Malta	Department of Fisheries and Aquaculture
Hanna	Swahnberg	Sweden	Swedish Agency Marine and Water Management
Emmanouil	Tziolas	Greece	Fisheries Research Institute
Irene	Tzouramani	Greece	Agricultural Economics Research Institute
Joonas	Valve	Finland	Natural Resources Institute Finland (Luke)
Jarno	Virtanen	Finland	Joint Research Centre
Ivana	Vukov	Croatia	Ministry of Agriculture

## 1.2. Agenda

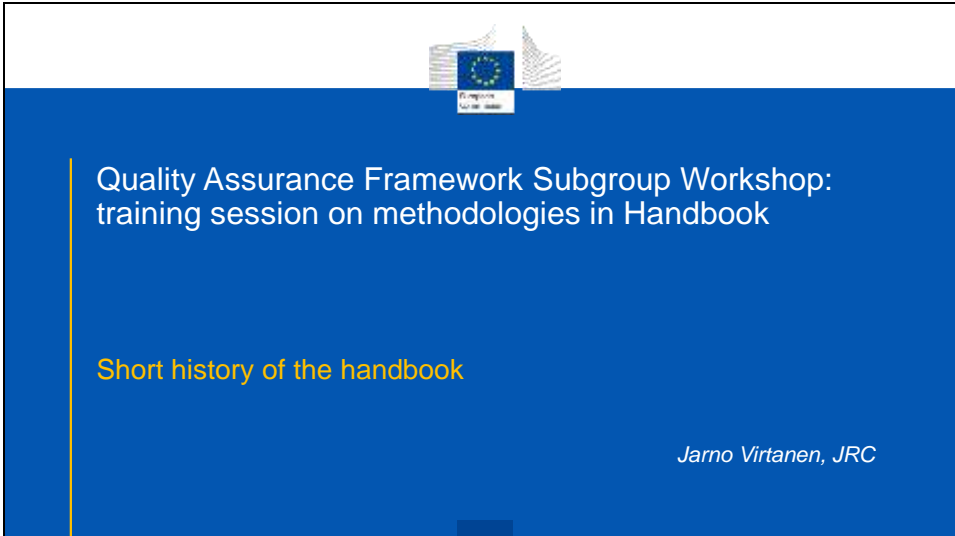
CET	<b>Tuesday 4.5.</b>	<b>Wednesday 5.5.</b>	<b>Thursday 6.5.</b>
9.00-10.30	Welcome & introduction (Heidi Pokki, Jarno Virtanen, Juha Heikkinen)	Simple random sampling demo (Juha Heikkinen)	Model-based inference, balanced sampling, ratio, regression, and calibration estimator, non-response (Juha Heikkinen)


10.30-11.00	<i>Coffee</i>	<i>Coffee</i>	<i>Coffee</i>
11.00-12.30	R crash course (Juha Heikkinen)	11-11:30 Groups: srs.r etc. 11:30- Design-based inference, systematic sampling. (Juha Heikkinen)	Methodological report as part of the work plan (Evelina Sabatella)
12.30-13.30	<i>Lunch</i>	<i>Lunch (12:40 – 13:40)</i> <i>Return to plenary</i>	<i>Lunch</i>
13.30-15.00	R crash course continue (Juha Heikkinen)	Unequal probability sampling, auxiliary information (Juha Heikkinen) <b>14:50 groups</b>	(Comparison of strategies and domain estimation) OR continue discussions on methodological report
15.00-15.30	<i>Coffee</i>	<i>Coffee</i>	<i>Coffee</i>
15.30-17.00	R questions from groups? SIMPOP, sampling basics, simple random sampling theory (Juha Heikkinen)	<b>plenary</b> .... PPS and stratified sampling, comparison of basic sampling methods (Juha Heikkinen)	Summary, reflections, and questions (Juha Heikkinen, Heidi Pokki)

## 2 Presentations

The workshop presentations are introduced in the following sections. After an introduction and welcome by Heidi Pokki (LUKE), Jarno Virtanen (JRC) presented the history of the Handbook since 2009 to outline the workshop (section 2.1). Next Professor Juha Heikkinen (LUKE) started with a crash course on R followed with lectures and demonstrations on how to apply the statistical methods presented in the Handbook. This included several practical examples using R code (section 2.2). Next Evelina Sabatella (NISEA) presented the new methodological report as part of the work plans (section 2.3) and the group discussed the implications. Finally, the workshop was concluded by summarizing the lessons learned and discussing reflections on the training. The workshop was interactive, and some of the practical R code exercises were carried out in smaller groups in Teams.

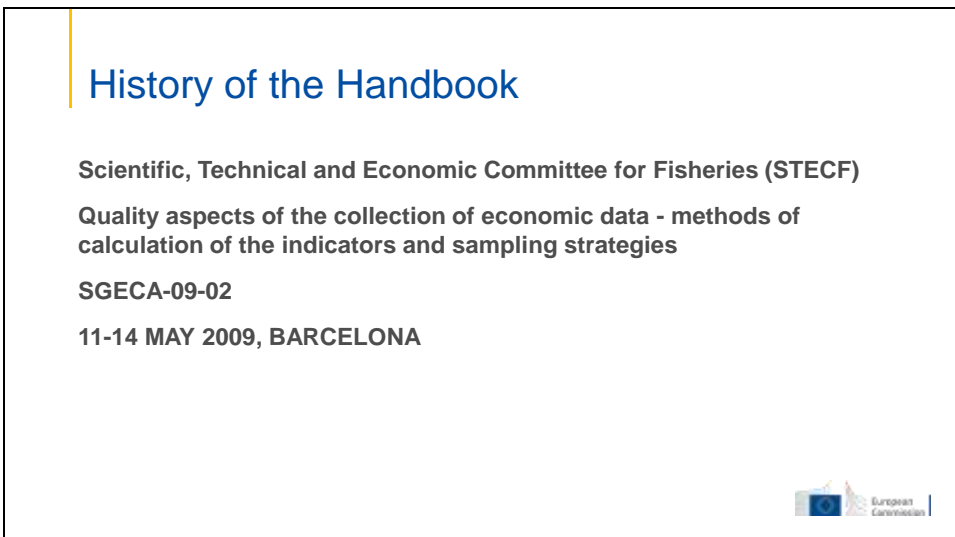
## 2.1. History of the Handbook



  
Quality Assurance Framework Subgroup Workshop:  
training session on methodologies in Handbook

Short history of the handbook

*Jarno Virtanen, JRC*




History of the Handbook

**Scientific, Technical and Economic Committee for Fisheries (STECF)**

**Quality aspects of the collection of economic data - methods of calculation of the indicators and sampling strategies**

**SGECA-09-02**

**11-14 MAY 2009, BARCELONA**





List of participants of the SGECA-09-02 meeting

STECF members: Sabatella Evelina (Chairman), Hatcher Aaron, Van Oostenbrugge Hans  
Virtanen Jarno

External experts: Berkenhagen Jörg, Bertelings Heleen, Collet Isabelle, DeMeo Michele  
Elias Leonor, Goti Leyre, Jonsson Anna, Motova Arina, Van Iseghem Sylvie

JRC experts: Guillen Jordi, Nord Jenny

European Commission: Calvo Angel, Cervante Antonio



## History of the Handbook

- Statistical issues and thresholds” (Helsinki, 2013)
- PGECON (2014) recommends that the handbook for best practices in economic data collection as proposed by the workshop in Helsinki will be commissioned. It should facilitate the enhancement of the survey design and quality of the economic data.
- PGECON (2016) repeats the need for several studies which have been strongly recommended, some of them for several years: Handbook on sampling design and estimation methods for fleet economic data collection



## History of the Handbook

- PGECON 2017 once again stressed the need for the “Handbook on sampling design and estimation methods for fleet economic data collection” as suggested several times before. It would provide a comprehensive reference for MS, thus facilitating the harmonization and comparability of data collection amongst MS



## PGECON 2017 Recommendations:

Ref. No.	Recommendation
1	<p>PGECON recommends that the reporting on the economic data collection and its resultant quality could be best organized by the following documentation:</p> <ul style="list-style-type: none"> <li>• Methodological document, including a detailed description of methods of surveys, structured in accordance with the ESS guidelines (<a href="#">Annex 7</a>) and has references to selected ESS QAF Principles (<a href="#">Annex 6</a>) listed in optimized WP Table 5B. This document can be either incorporated in the WP or used as a standalone document of the WP (<a href="#">Annex 8</a>).</li> <li>• Annual Quality report, with tables with specified quality indicators, taking into account the checklist for quality reporting and structured according to the ESS guidelines (<a href="#">Annex 6</a>).</li> </ul>
2	<p>PGECON recommends that during the EWG on quality assurance, the collected documentation and developed checklist and outline should be used as a basis for further development of the methodological report and the quality report.</p>



## SECFISH project

- Strengthening regional cooperation for the collection of social and economic data of the fisheries sector (2017 - 2019)
- WP 2: Harmonization of methodologies for sampling design and estimation methods for fleet and aquaculture economic data collection

In Work Package 2 the consortium will address the methodologies for sampling designs and estimation methods by providing a handbook including the relevant information.

- Handbook on sampling design and estimation methods for economic data collection in fisheries statistics



## PGECON 2019 recommendations:

- A Quality Assurance Framework (QAF) subgroup workshop should take place to define the process of quality assessment and assurance and revise the guidelines of the methodological report (with reference to the Handbook).
- Then as outcome, PGECON could provide recommendations and guidelines to AR evaluation EWG how to improve quality evaluation of DCF data and to complement the currently existing quality evaluation procedures



# Thank you

© European Union 2020



We noted the reuse of this presentation is authorised under the [CC-BY 4.0](https://creativecommons.org/licenses/by/4.0/) license. For any use or reproduction of elements that are not owned by the EU, permission may need to be sought directly from the respective right holders.


Slide xx: element concerned, source: e.g. [Fotolia.com](https://www.fotolia.com/); Slide xx: element concerned, source: e.g. [iStock.com](https://www.istock.com/)



## 2.2. Training sessions on methodologies in the Handbook

**Quality Assurance Framework  
Subgroup Workshop: training  
session on methodologies in  
Handbook**


Juha Heikkinen  
virtual training from Finland 4.-6.5.2021



1

**This workshop deals with**


- Probability sampling designs for selecting a part of the target population for data collection.
- Methods to estimate population parameters based on probability samples
- Methods to assess uncertainty of estimates based on the sampling design (design-based inference as opposed to model-based)
- Approaches for utilizing auxiliary information (from related registers or earlier surveys) for more efficient sampling and/or estimation
- Sampling simulation to anticipate the efficiency of different strategies (design+estimator) in a specific context.



2 11.-12.6.2019 Heikkinen: Sampling methods ...

**What we learn**

- General properties, advantages, disadvantages of different strategies
- Which estimation methods are suitable for different sampling designs: (approximate) design-unbiasedness.
- What can we say about the (anticipated) design variance before data collection?
- How can we estimate variance after data collection?
- How can we use simulated sampling from a related population for more useful anticipation of design variance etc.
- SIMPOP is also used to "confirm" and illustrate the theoretical results.



3 11.-12.6.2019 Heikkinen: Sampling methods ...

### SIMPOP population

- SIMPOP.xlsx: Artificial population containing records of  $p = 18$  numeric variables for  $N = 120$  fishing vessels; see Handbook Table 3.2.
- Full records for 100 vessels with ACTIVITY=1 and ID=1,2,...,100, partial records for 20 vessels with ACTIVITY=0 and ID=101,102,...,120. Ordered by ID.
- Variable STR3 appearing in Table 3.2. will be constructed later from variable GT.

4

11.-12.6.2019 Heikkinen: Sampling methods ...

### Sampling frame

- The sampling frame consists of *identifiable* units that are attached with unique *labels*, for example the identification code of a registered fishing vessel or the PIN of a person.
- ID codes allow population units to be sampled and contacted for data collection. By using identification codes, information can be extracted from registers and other sources and merged with records of the sampling frame, to be used in sampling and estimation procedures.
- Formally, a frame population is denoted  $\Omega = \{1, 2, \dots, k, \dots, N\}$ , it has  $N$  identifiable elements. In formulae, integer labels from 1 to  $N$  for simplicity. In practise, any unique labels will do.

5

11.-12.6.2019 Heikkinen: Sampling methods ...

### Sample survey

- The information of primary interest attached to the units of target population is denoted with the values of *target variable*  $Y$ .
- Values of  $Y$ ,  $\{y_1, \dots, y_k, \dots, y_N\}$ , are assumed unknown prior to the survey, which is carried out to obtain (practically error-free) measurements of  $Y$  for elements  $k \in \omega$  of the sample  $\omega$  drawn from the frame population.
- In practice, there are usually several target variables, but here we mainly assume just one of primary interest to make things simpler.
- The basic assumption behind sample survey is that accessing values of  $y_k$  for all population units (*census*) is too expensive, but a given number  $n$  (*sample size*) of units can be accessed.

6

11.-12.6.2019 Heikkinen: Sampling methods ...

### Population parameters

- The aim of a (sample) survey is to estimate the unknown values of *population parameters*, functions of all population values of the target variable

$$\theta = f(y_1, y_2, \dots, y_N)$$

- Here (and in the Handbook) mainly consider *population total*

$$\tau = \sum_{k=1}^N y_k = y_1 + y_2 + \dots + y_N$$

- In case of VALUE, this is the total value of the landings over all vessels in the population.
- In case of census, we could simply compute the true value of  $\tau$ , but in case of sample survey, we can only estimate it.

7

11.-12.6.2019 Heikkinen: Sampling methods ...

### SIMPOP frame and simulated surveys

- In case of active vessels in SIMPOP, unit labels are the values of variable ID (1,2,...,N = 100).
- CATCH, VALUE, and TOTAL\_COST represent typical target variables Y.
- Unlike in real-world survey contexts, here we actually know the values of Y for all population units, but in simulated surveys we pretend that we don't.
- However, knowing the "truth" about the whole population (true values of population parameters) allows us to assess the true performance of various sampling strategies (for this particular population).

SIMPOP\_VALUE.R



8

11.-12.6.2019 Heikkinen: Sampling methods ...

### Simple random sampling (SRS)

- Fixed sample size n, specified by the user.
- Simple random sampling (without replacement) selects a random subset  $\omega \subset \{1, 2, \dots, N\}$  of distinct population units (labels) so that all subsets of size n have the same selection probability.
- Inclusion probability of any given population unit k into the sample,  $\pi_k = nTN$ , is the same for all units.
- In general,  $w_k = 1T\pi_k$  is known as the sampling weight of unit k.
- In this case,  $w_k = NTn$  for all units.



9

11.-12.6.2019 Heikkinen: Sampling methods ...

### Horvitz-Thompson estimator

- In general, if we know inclusion probabilities  $\pi_k$ , then population total  $\tau$  can be estimated (design-)unbiasedly (we'll return to meaning of this) by Horvitz-Thompson (HT) estimator

$$\tau_{HT} = \sum_{k \in \omega} \frac{y_k}{\pi_k} = \sum_{k \in \omega} w_k y_k$$

- For SRS (without replacement)

$$\tau_{HT} = \sum_{k \in \omega} \frac{N}{n} y_k = N\bar{y}_\omega$$

where  $\bar{y}_\omega = \sum_{k \in \omega} y_k / n$  is the sample mean of the target variable.



10

11.-12.6.2019 Heikkinen: Sampling methods ...

### Uncertainty in estimation

- Since we assume that  $y_k$  is available without error for the sampled units, the difference between true  $\tau$  and estimated  $\tau_{HT}$  is completely due to sampling, i.e.,  $\tau_{HT}$  is computed from a sample (only a subset of the population) and  $\tau$  from the whole population.
- Estimates of uncertainty, i.e., quantifications of how different  $\tau$  may be from  $\tau_{HT}$ , can be derived from design variance  $V(\tau_{HT})$ .
- $V(\tau_{HT})$  depends both on sampling design (e.g., size n SRS) and on estimator (e.g., HT). Design + estimator = sampling strategy.
- The main idea of the workshop (and the Handbook) is to compare  $V(\tau_{HT})$  of different sampling strategies at fixed costs (n, in this workshop), and figure out how to estimate  $V(\tau_{HT})$  from the sample.



11

11.-12.6.2019 Heikkinen: Sampling methods ...

### Estimation of uncertainty; terminology

$v(\tau_{HT})$  estimator of design variance, whose value (variance estimate) can be computed from the selected sample.

$s.e(\tau_{HT}) = \sqrt{v(\tau_{HT})}$  standard error.

$cv(\tau_{HT}) = \frac{s.e(\tau_{HT})}{\tau_{HT}}$  coefficient of variation, usually expressed as percentage.

$DEFF(\tau_{HT})$  design effect, the ratio of the design variances associated with the target strategy and a reference strategy (usually SRS-HT), with equal costs ( $n$ ); or, ratio of equivalent sample sizes leading to equal variances.

$\tau_{HT} \pm t_{1-\alpha/2, n-1} \times s.e.$  confidence interval (CI). For two-sided 95% CI ( $\alpha = 0.05$ ), can use  $t_{1-\alpha/2, n-1} = 1.96$  if  $n$  is large; more details later.

12

11.-12.6.2019 Heikkinen: Sampling methods ...

### SRS-HT uncertainty

Design variance

$$V_{SRS}(\tau_{HT}) = N^2 \left(1 - \frac{n}{N}\right) S^2$$

where  $S^2$  is the population variance of  $Y$ , can be unbiasedly estimated by

$$v_{SRS}(\tau_{HT}) = N^2 \left(1 - \frac{n}{N}\right) s^2$$

where

$$s^2 = \frac{1}{n-1} \sum_{k \in \mathcal{D}} (y_k - \bar{y})^2$$

is the sample variance of  $Y$ .

13

11.-12.6.2019 Heikkinen: Sampling methods ...

### What did we do? Why does it work?

- Suppose we could repeat the sample survey several – preferably infinitely many – times with the same strategy, leading to estimates  $\tau_{HT}^1, \tau_{HT}^2, \dots$
- In real life we only have one survey (at one time).
- With artificial populations, like SIMPOP, we can do a finite (but as large as we like) number of replications (simulated sampling).
- With statistical theory, we can "do" infinitely many replications.
- Concepts like "design-unbiasedness" and "design variance" tell us what would happen in such replications.
- This is the basis of design-based inference (uncertainty assessment).

14

11.-12.6.2019 Heikkinen: Sampling methods ...

### Design-unbiasedness

- A sampling strategy is design-unbiased, if

$$E\tau_{HT} = \lim_{M \rightarrow \infty} \frac{1}{M} \sum_{m=1}^M \tau_{HT}^m = \tau$$

- Statistical theory says that SRS-HT, as well as many other common strategies, is design-unbiased.
- Design-unbiasedness is nice, because it means that we are not systematically over- or underestimating.
- But it does not tell us anything about uncertainty.
- In many cases, as we shall see, mildly biased (nearly design unbiased, Handbook Section 2.5) are preferable.

15

11.-12.6.2019 Heikkinen: Sampling methods ...

### Design variance

- associated with a given sampling strategy, is simply the variance of  $\tau\bar{\varepsilon}$  between (hypothetical) replications of the strategy.
- Statistical theory says that, under certain assumptions, we can actually estimate this variance from just one survey. This is the essential core of the whole business of statistical inference.
- For example, on slide 13, we estimated variance between surveys using the variance within one survey, between the surveyed units.

16

11.-12.6.2019 Heikkinen: Sampling methods ...

### Confidence interval

- at confidence level  $\gamma = 1 - \alpha$  has probability  $\gamma$  of containing the true value  $\tau$ .
- More precisely, if  $\gamma = 95\%$ , then a valid CI contains  $\tau$  in 95% of the (infinite number of) replicated samples.
- Thus, if we can construct a valid CI, then we have 95% chances of selecting such a sample that the CI constructed from it contains  $\tau$ .
- In practice, CI is always approximative, because it is based on assumptions (like applicability of central limit theorem).
- Quite generally, a well-founded choice is to select the critical value  $t_{1-\alpha/2, n-1}$  (slide 12) to be the  $1 - \alpha/2$  quantile of the  $t$ -distribution with  $n - 1$  degrees of freedom.

17

11.-12.6.2019 Heikkinen: Sampling methods ...

srs\_sim.R

### Systematic sampling (SYS)

- Select sampling interval  $a$  (integer)
- If you want sample size (approximately)  $n$ , then select  $a = N/n$ , rounded to an integer; downwards, if you want to secure sample size  $\geq n$ .
- Create integer labels  $1, 2, \dots, N$  to population units
- Select first sample unit label  $k_1$  at random from  $\{1, 2, \dots, a\}$ ; equal selection probability  $1/a$  for each.
- Then the systematic sample  $\omega = \{k_1, k_2, \dots, k_n\}$  contains units with labels  $k_1, k_2 = k_1 + a, k_3 = k_2 + a$ , etc., until  $k_n$  such that  $k_n + a > N$ .
- If  $N/a$  is an integer, then  $n = N/a$ , else sample size  $n$  is random.

18

11.-12.6.2019 Heikkinen: Sampling methods ...

### Systematic sampling: HT or conditional estimator?

- Inclusion probability  $\pi_k = 1/a$  for all population elements  $k$ .
- Thus, HT estimator of population total is

$$\tau_{HT}^* = \sum_{k \in \omega} \frac{y_k}{\pi_k} = a \sum_{k \in \omega} y_k$$

- Even if  $n$  is random,  $\tau_{HT}^*$  is always design-unbiased.
- But the *conditional estimator*  $N\bar{y}(\mathcal{O}_\omega)$  may be more efficient (may have smaller variance).
- In  $\tau_{HT}^*$ , sampling weights  $w_k = a$  do not depend on the realized sample, but in the conditional estimator, weights  $w_k = N/n$  vary between samples depending on  $n$  (if  $n$  is random).
- In conditional estimator, sum of weights is always  $= N$ .

19

11.-12.6.2019 Heikkinen: Sampling methods ...

### Systematic sampling: useful?

- If integer labels are independent of the target variable, then the SRS variance estimator (slide 13) is applicable.
- If the order of the integer labels correlates positively with the target variable, then systematic sampling is usually more efficient, so that SRS variance estimator should be positively biased (conservative).
- Creating the integer labels by regions (1 to  $N_1$  for units from region 1,  $N_1 + 1$  to  $N_1 + N_2$  for units from region 2, and so on) can ensure good geographic representation in a systematic sample.
- Systematic sampling can be used in some situations, where the population size  $N$  is not known prior to sampling.
- E.g., sample vessels by visiting each harbour of some region in turn.

### Unequal probability sampling

- SRS and SYS give equal inclusion probabilities  $\pi_k = P(k \in \omega) = n/N$  to all population units  $k$ ;  $\pi_k = n/N$  and  $\pi_k = 1/a$ , respectively.
- Unequal probability sampling gives different inclusion probabilities  $\pi_k$  to different units  $k$ .
- Completely ok with any choice of  $\pi_k \in (0,1)$ , but note, in particular, that  $\pi_k > 0$  must hold for all  $k$ .
- HT estimator  $\tau_{HT}^{\Sigma} = \sigma_{k \in \omega} \frac{y_k}{\pi_k}$  remains unbiased.
- But conditional estimator  $\tau_{\text{cond}}^{\Sigma} = \frac{N}{N_{HT}} \tau_{HT}^{\Sigma}$  may be more efficient.

### Variance in unequal probability sampling

- If sampling fraction  $n/N$  is small, then

$$\text{var}(\tau_{HT}^{\Sigma}) = \frac{n \left(1 - \frac{n}{N}\right)}{n-1} \frac{1}{n} \sum_{k \in \omega} \left( \frac{y_k}{\pi_k} - \frac{\tau_{HT}^{\Sigma}}{n} \right)^2$$

usually ok (maybe somewhat conservative).

- Otherwise, there are more cumbersome formulae with second-order inclusion probabilities for more accurate variance estimation.
- If  $\pi_k$ 's are uncorrelated with  $y_k$ 's then the variance is always greater than with SRS (equal inclusion probabilities optimal).
- But if  $\pi_k$ 's approximately proportional to  $y_k$ 's then, unequal probability sampling can lead to extremely efficient estimation (see PPS, a bit later).

### Auxiliary data: unit level, aggregate, multivariate

- Thus SRS is an optimal sampling design **in absence of any auxiliary information**. But if some exist (which is practically always the case), there are several ways to improve.
- To improve **sampling** design, auxiliary data must be available for all sampling units (=population units in element sampling).
- Aggregate auxiliary information can be used by some **estimation methods** (ratio estimation, linear regression estimation). They require unit-level auxiliary data only for the sampled units.
- We begin with sampling designs and assume that the values of some auxiliary variable  $x$  are available for all population units.
- Some methods (like balanced sampling) can utilize several  $x$ -variables.

### Auxiliary data: usefulness, data types

- Auxiliary data does not have to be "correct" in any sense, as long as the same method of producing it was used both for sampled and non-sampled units (or aggregates).
- If  $x$  co-varies with  $y$ , then improvements may be gained. If you are not confident about strong relationship between  $x$  and  $y$ , then it is usually possible to choose strategies that lead to minor losses even in absence of correlation.
- Auxiliary data may be categorical (e.g., stratified sampling), ordinal (e.g., systematic sampling) or (practically) continuous (e.g., PPS sampling). In spatial sampling, coordinates may play the role of auxiliary variables.

24

11.-12.6.2019 Heikkinen: Sampling methods ...

### Auxiliary data: SIMPOP

- In SIMPOP,  $GT$ ,  $kW$ , and  $DAS$ , as well as, their combinations  $GT\_DAS$  and  $kW\_DAS$  represent typical auxiliary variables, which might be available for all population units prior to the survey.
- They are also variables that can be anticipated to co-vary with the target variables. In SIMPOP, such correlations are quite strong (Handbook Section 3.2).
- In real life, can do similar analyses based on data collected in a previous related survey to find good auxiliary variables.

descriptives.R

25

11.-12.6.2019 Heikkinen: Sampling methods ...

### Systematic sampling revisited

- Systematic sampling based on integer labels obtained as ranks of auxiliary variable  $GT$  (equivalently, SIMPOP ordered by  $GT$  and then re-labeled by  $1, 2, \dots, N$ ) is clearly more efficient than SRS-HT, more so with the conditional estimator.
- The problem with systematic sampling is that we don't have a design-unbiased variance estimator that would be needed to be able to report the improved efficiency.

sys\_aux.R

26

11.-12.6.2019 Heikkinen: Sampling methods ...

### PPS sampling

- Unequal probability sampling design, where  $y_k T \pi_k$ , varies (much) less than  $y_k$ , can lead to extremely efficient estimators.
- But  $\tau \mathcal{E}_{HT}$  may be unstable, if  $\pi_k$ 's vary wildly.
- For example, if  $Y$  is CATCH and  $X = DAS$  (effort) then we can expect  $y_k T x_k$  (catch per effort) to vary less than  $y_k$ .
- And if we know  $x_k$  for all population units, then we can choose
 
$$\pi_k = \frac{n}{\sum_{j=1}^N x_j} x_k \propto x_k.$$
- Such unequal probability scheme is often called sampling with (inclusion) *probability proportional to size* (PPS,  $\pi$ PS), where  $X$  is a "size variable".

pps.r

27

11.-12.6.2019 Heikkinen: Sampling methods ...

### Stratified sampling

- Divide population into parts (strata) so that each population unit belongs to one, and only one, stratum.
- Assume known population sizes,  $N_h, h = 1, \dots, H$ , of the  $H$  strata;  $\sum_{h=1}^H N_h = N$ .
- Conduct separate sample survey in each stratum with sample sizes  $n_h$ ;  $\sum_{h=1}^H n_h = n$ .
- This yields estimates  $\tau \bar{y}_h$  for each stratum,  $h = 1, \dots, H$ , and estimated design variances  $v(\hat{\tau \bar{y}})$ .



Heikkinen: Sampling methods ...

### Why stratify?

Population-level estimates may improve:

- Random under/over-representation of different parts of population may lead to poor estimates of population totals.
- If within-stratum variance small in comparison to between-stratum variance, then stratified sampling more efficient than SRS.
- Greater proportional allocation of sample to strata with more variation can lead to large reductions in variance.
- Stratified sampling can also facilitate more efficient domain estimation (to be discussed in the end of the workshop)

More flexible and robust than, albeit not as efficient as, PPS. Can also combine; see Handbook sec. 3.6.4.



29

11.-12.6.2019 Heikkinen: Sampling methods ...

### Stratified estimator

The total of the whole population can be unbiasedly estimated by

$$\tau \bar{y}_{STR} = \sum_{h=1}^H \tau \bar{y}_h$$

and the design variance of this *stratified estimator* by

$$v(\hat{\tau \bar{y}}_{STR}) = \sum_{h=1}^H v(\hat{\tau \bar{y}}_h)$$

Simplicity of these formulae is one reason why we prefer to work with totals rather than means in this workshop.



30

11.-12.6.2019 Heikkinen: Sampling methods ...

### Allocation of stratified sample

- Equal:  $n_h \equiv n/H$ . Makes sense, if stratification is mainly for domain estimation: sample equally spread to all strata.
- Proportional:  $n_h \propto N_h$  (approximately) the same in all strata, i.e.  $n_h \approx nN_h/N$ ; usually can't make them exactly the same. Most natural; gives all sample units approximately the same sampling weight. Proportional allocation is a sensible "default".
- Neyman: increase allocation from proportional in strata with greatest variation in  $y$

$$n_h = n \frac{N_h \sigma_h}{\sum_{h'=1}^H N_{h'} \sigma_{h'}}$$

Optimal, if within-stratum variances  $\sigma_h^2$  of  $y$  are known; better than proportional, if we have good guesses of them.



31

11.-12.6.2019 Heikkinen: Sampling methods ...

### Comparison of sampling designs

- In demonstrations this far, we have utilized artificial population SIMPOP, with features (variances of variables and correlations between them) expected to resemble those of a true target population.
- In particular, we have performed sampling simulations, repeating the planned design several times and getting an idea of the anticipated design variance from the empirical variance of the estimate over the simulations.
- In this summary, we show that it is also (often) possible to compute the theoretical variances without simulation.

32

11.-12.6.2019 Heikkinen: Sampling methods ...

### Comparison: Target and designs

- Consider estimation of population total  $\tau = \sigma_{k=1}^N y_k$  based on the observed values of  $Y$  (CATCH) for a sample  $\omega \subset \{1, 2, \dots, N\}$  of  $n$  population units.
- In designs considered this far,  $\omega$  is selected by SRS, SYS, PPS, or STR. For STR, we consider both proportional allocation STR-p and Neyman allocation STR-Ney in this comparison; SRS within strata.
- All except SRS (can) use auxiliary data for
  - SYS: ordering of population
  - PPS: determination of unequal inclusion probabilities  $\pi_k$
  - STR-p: determination of strata
  - STR-Ney: determination of strata and anticipation of within-stratum variances for optimal allocation

33

11.-12.6.2019 Heikkinen: Sampling methods ...

### Comparison: Auxiliary data

- In comparisons, we use auxiliary variable STR3 to determine the strata: It provides a partition of the population into three strata of approximately equal size through an aggregation of variable GT.
- For other purposes, we compare designs based on two alternative auxiliary variables GT and GT\_DAS.
- We compare variances of the basic estimators associated with each design with sample size  $n = 20$  ( $N = 100$ ).

34

11.-12.6.2019 Heikkinen: Sampling methods ...

### Comparison: Basic estimators

- The basic estimator is  $\tau \bar{\Sigma}_{HT} = \sigma_{k \in \omega} y_k^T \pi_k$  for all other designs,  $\tau \bar{\Sigma}_{STR} = \bar{\sigma}_{h=1}^H \tau \bar{\Sigma}_{HT,h}$  for STR.

- In SRS and SYS,  $\pi_k = nTN$  for all  $k$ , in PPS

$$\pi_k = \frac{n}{\sigma_{j=1}^N x} x_k,$$

where  $X$  is the auxiliary variable<sup>1</sup>, and in STR  $\pi_k = n_h/N_h$  for  $k \in h$ .

- In STR-p,  $n_h \approx nN_h/N$ , in STR-Ney

$$n_h \approx n \frac{N_h \sigma_h}{\sigma_{h=1}^H N_h \sigma_{h'}},$$

where  $\sigma_h$  is the standard deviation of  $X$  within stratum  $h$ .

35

11.-12.6.2019 Heikkinen: Sampling methods ...

### Comparison: Theoretical variances

- For SRS,

$$V_{SRS}(\tau_{HT}) = N^2 \left(1 - \frac{n}{N}\right) S^2$$

where  $S^2$  is the population variance of  $Y$ .

- For SYS,

$$V_{SYS}(\tau_{HT}) = \frac{1}{a} \sum_{k=1}^a (\tau_{HT,k} - \tau)^2$$

where  $a = N/n$  and  $\tau_{HT,k}$  is the HT estimate from systematic sample with  $k_1 = k$ .

36

11.-12.6.2019 Heikkinen: Sampling methods ...

### Summary: theoretical variances (ctd)

- For PPS, we use conservative approximation

$$V_{PPS}(\tau_{HT}) \approx \frac{n}{N} \sum_{k=1}^N \left(\frac{y_k}{n} - \tau\right)^2$$

derived under with-replacement PPS.

- For STR,

$$V_{STR}(\tau_{STR}) = \sum_{h=1}^H V_{SRS}(\tau_{HT,h})$$

where  $V_{SRS}(\tau_{HT,h}) = N_h^2 \left(1 - \frac{n_h}{N_h}\right) S_h^2$  and  $S_h^2$  is the population variance of  $Y$  within stratum  $h$ .

37

11.-12.6.2019 Heikkinen: Sampling methods ...

comp\_des.R

### Summary: theoretical comparisons

des	aux	cv, %	DEFF
SRS		7.10	
SYS	GT	3.52	0.496
SYS	GT_DAS	1.46	0.206
PPS	GT	5.79	0.816
PPS	GT_DAS	3.54	0.498
STR_p	STR03	6.12	0.862
STR_Ney	STR03+GT	6.21	0.875
STR_Ney	STR03+GT_DAS	6.08	0.856

38

11.-12.6.2019 Heikkinen: Sampling methods ...

### Comments on comparison

- SYS was extremely efficient (as it often is), but the serious problem is that it is difficult to make this efficiency visible in practise:
  - No (even approximately) unbiased variance estimator exists
  - If we use SRS variance, then the improved efficiency does not show in our reported uncertainty.
  - Actually SRS variance estimates from a typical SYS sample is even larger than from a typical SRS sample of the same size, because SYS tries to maximize within-sample variance.

39

11.-12.6.2019 Heikkinen: Sampling methods ...

### Further comments on comparison

- In general, STR cannot usually be expected to be particularly efficient with a small number of strata (SYS is similar to STR with  $n$  strata!)
- Nevertheless, advantages of STR include
  - Simplicity and robustness
  - Availability of reliable, design-unbiased variance estimates
  - Possibility to utilize several auxiliary variables, when constructing the strata
  - Possibility to utilize anticipated within-stratum variances that are obtained from sources external to the current population (e.g., previous survey).

40

11.-12.6.2019 Heikkinen: Sampling methods ...

### Auxiliary data in estimation

- This far auxiliary data has been used to improve sampling design.
- It can also be used to obtain a more efficient estimator for a given sampling design.
- Nothing prevents us from using (the same) auxiliary data both in sampling and in estimation (although we do not go into that in this course).
- However, if you are (potentially) interested in several population parameters, then it may not be wise to optimize sampling for one of them, and it is not so easy to optimize it for all of them.
- On the contrary, we can use different estimators and possibly different auxiliary data for the different parameters.

41

11.-12.6.2019 Heikkinen: Sampling methods ...

### Ratio estimator

- Suppose we know the population total  $\tau_x$  of size variable  $x$  and
- observe the values  $y_i$  and  $x_i$  for the sample elements  $i \in \omega$ .
- Then, if our target variable  $y$  is also a size variable, and if  $x$  and  $y$  are correlated, the ( $\pi$  weighted) ratio estimator

$$\tau_{y,\text{rat}}^{\pi} = \frac{\tau_x}{\tau_{x,H}^{\pi}} \tau_{y,\text{HT}}$$

is more efficient than the HT estimator  $\tau_{y,\text{HT}}^{\pi}$  (this is the same as earlier  $\tau_{y,\text{HT}}$ , but notation had to be complicated for obvious reason).

- Ratio estimator corrects HT estimator by the ratio of the true and estimated values of  $\tau_x$ .
- This makes sense, since both  $\tau_{y,\text{HT}}^{\pi}$  and  $\tau_{x,\text{HT}}^{\pi}$  are based on the same sample.

42

11.-12.6.2019 Heikkinen: Sampling methods ...

rat.r

### Balanced sampling

is a restricted version of the basic sampling schemes such that Horvitz–Thompson estimators of the totals of auxiliary variables are the same or almost the same as the true population totals (e.g. Deville & Tillé 2004).

- In particular, this means that the ratio estimator is (almost) identical to the HT estimator.
- This should help, because we can then combine exact unbiasedness of the HT estimator with reduced variance of ratio estimator.
- Unlike other sampling designs this far, can directly utilize several auxiliary variables simultaneously.
- But rely on HT estimator: no additivity problems.

43

11.-12.6.2019 Heikkinen: Sampling methods ...

bal.R

### Local pivotal method (lpm, Grafström & al. 2012)

- Similar idea as in balanced sampling, but
- aim for balance by spreading out the sample as much as possible in the space of the auxiliary variables.
- So not only the means (or HT estimators of totals) are similar in sample and population, but also the whole distribution.
- Approximate variance estimator available.
- Example: Sample of agricultural fields from Field Plot Registry subsampled by lpm with coordinates as auxiliary data.



### Model-based approach

- This far, we have assumed that the uncertainty in estimators is due to probability sampling; values  $y_i$  have been treated as fixed (non-random).
- In model-based approach (e.g., Chambers & Clark 2012),  $y_i$ 's are treated as realizations of random variables  $Y_i$ , and variability between samples that could have been drawn does not affect variance estimation.
- Inferences will then be more influenced by the characteristics of the sample that was actually drawn (vs. sampling scheme in design-based approach)
- Randomized (preferably balanced) sampling is, nevertheless, recommended as a safe-guard against poor inferences caused by model-misspecification.

### Model-based versus design-based

- You can't say that one is right and the other isn't.
- The  $y_i$ 's are what they are, so there is no true randomness in them. But also the sample is what it is.
- So both approaches are based on hypothetical constructions, which give us tools to deal with uncertainty.
- The design-based approach is usually considered more objective, because it relies on random mechanism we have created ourselves.
- On the other hand, model-based approach is more similar to other uses of statistical methods, and more generally applicable, for example, in uncertainty assessment for systematic sampling and in small area estimation.

### The simplest model

- Assume that  $Y_i$ 's are iid (independent and identically distributed) with unknown common expected value  $\mu$  and unknown common variance  $\sigma^2$ .
- Then the sample mean  $\bar{y} = \sigma_{i \in \omega} Y_i / n$  is a model-unbiased predictor of the population mean  $Z = \sigma_{i=1}^N Y_i / N$  for any sample  $\omega$  of size  $n$  no matter, how it was selected, and
- the model-unbiased estimator of the mean square prediction error

$$E(\bar{y} - Z)^2 = \frac{N-n}{n} s^2$$

where  $s^2$  is the sample variance of  $Y_i$ 's, is identical to the design-unbiased estimator of variance for the HT estimator under SRS (Thompson 2002, sec. 2.7).

### Model-based vs model-assisted

- More generally, the model- and design-based approaches sometimes lead to the same calculations, but model-based variances are dependent on realism of the assumed model.
- It is also possible to use (regression) models in the design-based framework (Särndal et al. 2003).
- For example, ratio estimator is efficient, when model  $E(Y_i) = \beta x_i$ ;  $V(Y_i) = \sigma^2 x_i$  fits well, but the validity of its (design-based) variance estimator does not depend on this model assumption.
- Ratio estimator is an example of a general class of regression estimators "generated" by models, or "model-assisted estimators" (Särndal et al. 2003, ch. 6 & 7).

48

11.-12.6.2019 Heikkinen: Sampling methods ...

### Regression estimation & non-response

- Linear regression model generates linear regression estimator.
- In addition to the model fitted to sample values of  $Y$  and  $X$ , regression estimation of the population mean of  $Y$  only requires the population total of  $X$ .
- Generalized linear model  $\rightarrow$  generalized regression estimator. Requires unit-level auxiliary data.
- NOTE: can use several  $X$ -variables!
- For calibration vs. regression estimation, see, e.g., [Lumley \(2008\)](#) or [Lumley & al. \(2011\)](#). See also [Zardetto \(2015\)](#).
- Non-response: Handbook Ch. 5.

49

11.-12.6.2019 Heikkinen: Sampling methods ...

### Comparison

- Continued from slide 38.
- Use GT, GT\_DAS, and GT + DAS (if possible) in
  - PPS (HT estimator)
  - LPM (HT estimator)
  - RAT (SRS sample)
  - REG (SRS sample)
- Compare variances with sample size  $n = 20$  ( $N = 100$ ).
- $V_{PPS}(\tau_{\mathcal{E}_{HT}})$  from slide 37.
- $V_{LPM}(\tau_{\mathcal{E}_{HT}})$  by simulation.
- $V_{SRS}(\tau_{\mathcal{E}_{RA}})$  and  $V_{SRS}(\tau_{\mathcal{E}_{REG}})$ : see next slides

50

11.-12.6.2019 Heikkinen: Sampling methods ...

### Theoretical variance of ratio estimator

Design variance of ratio estimator

$$\tau_{\mathcal{E}_{rat}}^{\mathcal{D}} = \frac{\tau_x}{\tau_{\mathcal{E}_{x,H}} \Sigma} \tau_{y,HT}$$

under SRS can be approximated by

$$V_{SRS}(\tau_{\mathcal{E}_{rat}}^{\mathcal{D}}) = \frac{N-n}{n} \frac{1}{n} \sum_{k=1}^N \left( y_k - \frac{\tau_y}{\tau_x} x_k \right)^2$$

51

11.-12.6.2019 Heikkinen: Sampling methods ...

### Theoretical variance of regression estimator

Design variance of regression estimator

$$\tau \sum_{q=1}^p \frac{\tau_{x_q} - \bar{\tau}_{x_q}}{\tau_{x_q,HT}}^2$$

under SRS can be approximated by

$$V_{SRS}(\tau \sum_{q=1}^p \frac{\tau_{x_q} - \bar{\tau}_{x_q}}{\tau_{x_q,HT}}^2) = \frac{N-n}{n} \sum_{k=1}^N \left( y_k - \sum_{q=1}^p b_q x_{qk} \right)^2$$

### Comparison: theoretical cv's\*

des	aux	cv, %	DEFF
PPS	GT	5.79	0.816
RAT	GT	5.87	0.826
REG	GT	5.92	0.833
LPM	GT	5.98	0.843
PPS	GT_DAS	3.54	0.498
RAT	GT_DAS	3.78	0.533
REG	GT_DAS	3.82	0.538
LPM	GT_DAS	4.62	0.651
REG	GT+DAS	3.91	0.551
LPM	GT+DAS	4.67	0.658

\*except for LPM, which was simulated

### Comments on comparison

- Using two auxiliary variables GT and DAS was not an improvement over using one auxiliary variable GT\_DAS, in this case. The reason might be that CATCH in SIMPOP was actually simulated using GT\_DAS. So this result is not generalizable to real populations.
- RAT and REG work here equally well (or RAT even slightly better), because regression of CATCH vs. predictors goes through the origin. REG might be better for non-linear relationships.
- RAT (and REG) yield nearly the same efficiency as PPS, but have the definitive advantage of allowing different predictors for different targets while PPS can use only one auxiliary variable.

### Comments on comparison (ctd)

- LPM was not particularly efficient here.
- Its advantage is simplicity and robustness: equal probability sampling (unlike PPS) and unbiased HT estimator (unlike RAT and REG).
- It is also possible to use RAT or REG on an LPM sample.

### Planned domains

If estimates are needed for (sub-)domains of population, best to survey them as separate strata:

- Fixed sample size  $n_h$  in each domain (*planned domains*) is much easier to handle properly than random  $n_h$ 's, which typically result from not stratifying (*unplanned domains*).
- Fixed sample size in each domain also allows for better control over the precision of stratum-specific estimators.
- Some small or highly variable domains may require a relatively larger sample.
- Can even use different sampling strategies in different strata.



Heikkinen: Sampling methods ...

### Unplanned domains

- Unplanned domains are like strata, but they were not taken into account in sampling design.
- Population sizes  $N_d$  of unplanned domains  $d$  may be unknown.
- But for each sample unit, we can determine its domain.
- For example, size- $n$  SRS from the whole population leads to random sample sizes  $n_d$  for the domains.
- In such case, it is possible to obtain domain-specific estimates by transforming the original target variable  $Y$  into *extended domain variable*  $Y_d$  so that  $y_{dk} = y_k$ , if unit  $k$  belongs to domain  $d$ , and  $y_{dk} = 0$  otherwise.
- We can determine the value  $y_{dk}$  for each sampled unit, and estimate the population total of  $Y_d$ , i.e. domain total, by HT.



57

11.-12.6.2019 Heikkinen: Sampling methods ...

### Planned vs. unplanned domains

- Planned domains (stratified sampling) generally yield more precise domain estimates than the (unconditional) estimates resulting from unplanned domains (population-level sampling).
- It is not a big sin, as such, to assume stratified sampling even if population-level sampling was actually conducted (c.f. conditional inference and post-stratification).
- Then, if we know the population sizes of the domains, we can use fixed- $n$  HT estimators within domains like in stratified estimation.
- But this leads to domain-specific sampling weights so that estimates of domain totals don't add up to HT estimate of population total.



58

11.-12.6.2019 Heikkinen: Sampling methods ...

### Planned vs. unplanned domains, ctd.

- With one-way division to domains, we can use the stratified estimator for the population total and the additivity remains.
- But different stratifications would lead to different population estimates.
- To some extent, multi-way stratification is also possible with help of balanced sampling ([Falorsi & Righi 2008](#))
- Unplanned-domain approach (sampling weights from the population-level design) guarantees additivity in all divisions to domains.

domains.r



59


11.-12.6.2019 Heikkinen: Sampling methods ...

## 2.3. Methodological report as part of the work plan

Quality Assurance Framework Subgroup Workshop:  
training session on methodologies in Handbook, virtual  
training from Finland 4.-6.5.2021

**METHODOLOGICAL REPORT AS PART OF THE  
WORK PLAN**

Evelina Sabatella



**Revision of DCF Work Plan and Annual Report templates and  
guidelines  
(STECF-20-18)**

**3 WP/AR templates and guidelines**

Next steps:

- 1<sup>st</sup> feedback from NC/RCG: 30 April
- COM interservice consultation: May 2021
- RCG testing new templates & 2<sup>nd</sup> feedback: June 2021 (?)
- MS receive new templates: July 2021
- EMF(A)F committee voting: July 2021 (?)
- Translation & adoption: September 2021 (?)

**Text Box**

SECTION 5: ECONOMIC AND SOCIAL DATA IN FISHERIES  
Text Box 5.2: Economic and social variables for fisheries data collection

SECTION 6: ECONOMIC AND SOCIAL DATA IN AQUACULTURE  
Text Box 6.1: Economic and social variables for aquaculture data collection

SECTION 7: ECONOMIC AND SOCIAL DATA IN FISH PROCESSING  
Text Box 7.1: Economic and social variables for fish processing data collection

ANNEX 1.2 - QUALITY REPORT FOR SOCIOECONOMIC DATA SAMPLING SCHEME

**Tables**

Section 5-7. Economic and social data

Table 5.1	Fleet total population and clustering
Table 5.2	Economic and social variables for fisheries data collection strategy
Table 6.1	Economic and social variables for aquaculture data collection strategy
Table 7.1	Economic and social variables for fish processing data collection strategy

**Revision of DCF Work Plan and Annual Report templates and guidelines (STECF-20-18)**

All quality information was moved from Text Boxes to **Annex 1.2 (quality reports)**.

In order to improve the quality reporting of economic data, the EWG used PGECON's **Data Collection Methodological Document for Economic data**, as described in **Annex 8 of PGECON 2017**.

In addition to the sections suggested by PGECON, the EWG also included the confidentiality considerations currently available in Table 5B.

The inclusion of Annex 1.2 will considerably improve submitted information on data quality and should allow for a more efficient and relevant assessment of the WPs.

**Evaluation of annex 1.2 (quality reports)**

STECF spring plenary (2021):

the proposed dedicated annexes on data quality would improve the quality reporting of biological and economic data as well as allow for a more efficient and relevant assessment of the WPs.

The information provided through the submission of the proposed annexes on data quality will be extensive and expertise on sampling/survey design and data quality is needed for their assessment.

STECF concludes that the evaluation of these annexes should be conducted **prior to the EWG on evaluating WPs (EWG 21-17) through specific ad-hoc contracts during the pre-screening phase**.

The methodological approach is seldom modified within a WP period, and hence STECF considers that the evaluation of the data quality annexes is most likely only needed once for the WPs 2022-2024

**Revision of DCF Work Plan and Annual Report templates and guidelines (STECF-20-18)**

**ANNEX 1.2 - Quality Report for socio economic data sampling scheme**

**ANNEX 1.2 - Quality Report for socio economic data sampling scheme**

General comments:

Annex 1.2 fulfils the requirements of the EU-MAP referring to data to be collected under chapter II, section 3, 5, 6, 7 of the EU-MAP Delegated Decision on:

socio economic data on fisheries,  
aquaculture  
and any complementary data collection of  
fishing activity  
and fish processing

Use this document to describe quality aspects of the data collection process:

design,  
sampling implementation,  
data capture,  
data storage  
and data processing.

The annex should be filled for each sampling scheme.

**The handbook on sampling design (reference with link) should be used as a reference where applicable.**

**ANNEX 1.2 - Quality Report for socio economic data sampling scheme**

**Survey Specifications**

*Sector name refers to socio economic data on fisheries, aquaculture and any complementary data collection of fishing activity and processing as given in the EU-MAP Delegated Decision. Sampling scheme refers to survey technique: by census, by sampling, random or non-random, other (with explanation). If sampling then outline sampling design. Variables refer to Tables 7, 9 and 10 of the EU-MAP Delegated Decision. Supra region refer to Table 2 of the EU-MAP Implementing Decision. If the sampling scheme is the same in all supra regions put 'All Supra regions'.*

**Sector name(s):**

**Sampling scheme:**

**Variables:**

**Supra region(s):**

**ANNEX 1.2 - Quality Report for socio economic data sampling scheme**

**1. Survey planning**

Provide a short description of the population the sampling scheme applies to; e.g. 'less active vessels using passive gears'.

**AR comment:** Have there been any deviations?

**2. Survey design and strategy**

List data sources; e.g. interviews, registers, log books, sales notes, VMS, financial accounts etc.  
Describe how the sample sizes were determined.  
Describe survey methods and distribution; e.g. questionnaire forms by post, by email, on website, by phone etc. access to other datasets etc.  
Describe the role of auxiliary information, if any, in the strategy; e.g. for validation, cross referencing, fall back data source etc.

**AR comment:** Have there been any deviations?

**ANNEX 1.2 - Quality Report for socio economic data sampling scheme**

**1. Estimation design**

Describe method of calculating population estimate from sample.  
Describe method of calculating derived data: e.g. imputed values.  
Describe treatment of nonresponse.

**AR comment:** Have there been any deviations?

**2. Error checks**

Describe potential errors and how and where in the process these are detected, avoided or eliminated e.g., data; duplication, double counting, respondent error, upload error, processing error etc.

**AR comment:** Have there been any deviations?

**ANNEX 1.2 - Quality Report for socio economic data sampling scheme**

**1. Data storage and documentation**

Describe how the data is stored.  
Provide link to webpage where additional methodological documentation can be found, if an

**AR comment:** Have there been any deviations?

**2. Revision**

Describe the frequency of the methodology review e.g., revision of; segmentation, survey method per segment, per variable etc.

**AR comment:** Have there been any deviations?

**ANNEX 1.2 - Quality Report for socio economic data sampling scheme**

**1. Confidentiality**

Are procedures for confidential data handling in place and documented?  
Are protocols to enforce confidentiality between DCF partners in place and documented?  
Are protocols to enforce confidentiality with external users in place and documented?  
Are there any issues with publication of data due to confidentiality reasons? Provide an explanation.

**AR comment:** Have there been any deviations?

**AR comment:** Use this text box for providing any additional comments, if necessary.

### 3 Conclusions

The workshop participants had the opportunity to discuss specific methodological/statistical questions regarding their data during the workshop. Professor Heikkinen also provided the opportunity to seek methodological advice after the workshop. All course materials (presentations, instructions, R codes, simulated data) are available for the participants in Teams also for future reference.

The results of this workshop will be utilized in the process of compiling the regional work plans for economic data collection as part of the Fishn'Co project running 2021-2023. In this process, the first step is to set the level of ambition for harmonizing economic data collection (0-no coordination to 4-Joint data collection) before the RCG ECON 2021 meeting as described in Figure 1.

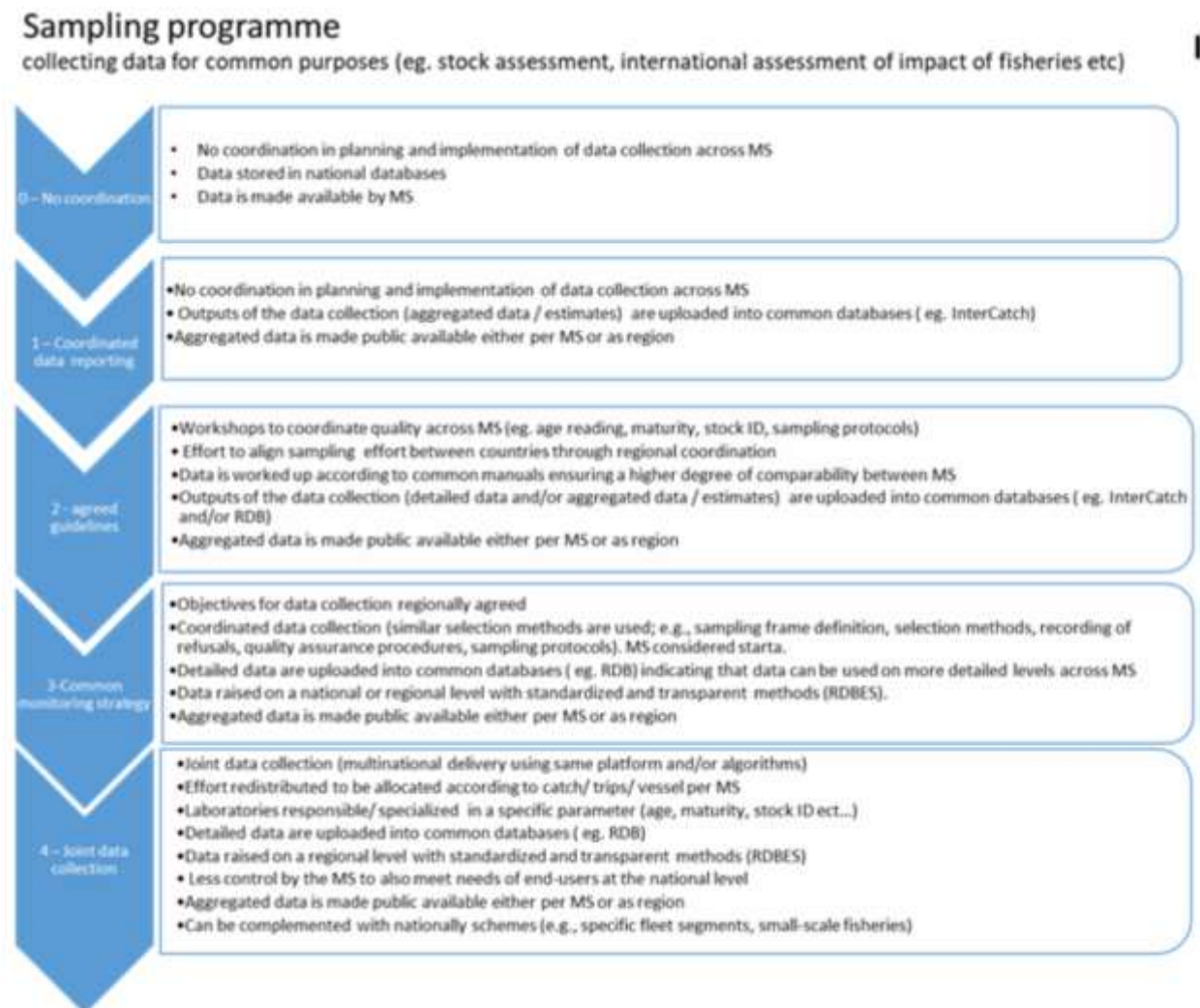


Figure 1. Levels of Ambition for regional coordination from the Fishn'Co project.