WORKSHOP ON ESTIMATION WITH THE RDBES DATA MODEL (WKRDB-EST; outputs from 2019 meeting)

VOLUME 2 | ISSUE 5

ICES SCIENTIFIC REPORTS

RAPPORTS
SCIENTIFIQUES DU CIEM

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

ISSN number: 2618-1371 I

ICES Scientific Reports

Volume 2 | Issue 5

WORKSHOP ON ESTIMATION WITH THE RDBES DATA MODEL (WKRDB-EST; outputs from 2019 meeting)

Recommended format for purpose of citation:

ICES. 2020. Workshop on Estimation with the RDBES data model (WKRDB-EST; outputs from 2019 meeting).
ICES Scientific Reports. 2:5. 106 pp. http://doi.org/10.17895/ices.pub.5956

Editors

Kirsten Birch Håkansson • Nuno Prista

Authors

Andrew Campbell • Chun Chen • Mary Christman • Liz Clarke • David Currie • Laurent Dubroca • Ana Claudia Fernandes • Edvin Fuglebakk • Hans Gerritsen • Kristiina Hommik • Ain Lankov • Twan Leijzer • Richard Meitern • Zuzanna Mirny • Karolina Molla Gazi • Marijus Spegys • Sven Stoetera • Marta Suska • Josefina Teruel Gómez • Ioannis Thasitis • Bart Vanelslander • Julia Wischnewski • Lucia Zarauz

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

Contents

i Executive summary

The RDBES is the new Regional DataBase and Estimation System. The RDBES is expected to replace the previous RDB and InterCatch by the end of 2021 and will bring significant improvements and transparency in the provision of estimates from commercial fisheries to stock assessment and other end-users. The developments of the RDBES meet the EU-MAP requirements of progress towards statistically sound sampling schemes. The RDBES data model and associated database are able to store, among other, sampling data alongside the elements required to describe the sampling design used in data collection. Upload of data to and estimation within the RDBES will require significant adaptation of the data collection processes of national institutes in several areas, including data storage, but also sampling design, field protocols, estimation and data provision to end-users. To secure a soft transition there is a need to intensify internal planning of these adaptations from 2020.

The Workshop on Estimation with the RDBES data model (WKRDB-EST) prepared data for 8 of the 13 upper hierarchies of the RDBES and developed a first set of R-scripts that handles design-based estimation in the RDBES data model. Developments and tests were positive and confirmed the usefulness of the data model for design-based estimation. These developments are publically available in the ICES GitHub (https://GitHub.com/ices-eg/WK_RDBES/tree/master/WKRDB-EST). The RDBES core group will continue the development and produce an R-package that aggregates a) a generic set of estimation functions and b) vignettes documenting design-based estimation in each type of sampling hierarchy.

WKRDB-EST examined and tested version v.1.17 of the RDBES model with feedback being obtained from 15 countries on 8 of the 13 upper hierarchies of the RDBES. The data model can now be considered relatively stable with mostly minor issues being identified. The RDBES core group will discuss these issues and incorporate in a future data model, v.1.18.

Finally, WKRDB-EST discussed the way forward in the development of the estimation component of RDBES. It was agreed that the priority for 2020 should be finalizing the development of design-based estimators. That development should include domain estimation and post-stratification since these aspects are necessary to produce estimates at the spatial and temporal resolution required by a variety of end-users. Development of script for model-assisted and model-based estimation based on the RDBES format should take place in other fora (e.g. SC-RDB-coordinated data workshops, WKs spawned by the ICES Working Group on Commercial Catches (WGCATCH)). A new WKRDB-EST will be suggested to SC-RDB for late 2020 where the developments of design-based estimation will be finalized.

ii Expert group information

Expert group name	Workshop on Estimation with the RDBES data model (WKRDB-EST))
Expert group cycle	Annual
Year cycle started	2019
Reporting year in cycle	1/1
Chair(s)	Nuno Prista, Sweden
	Kirsten Birch Håkansson, Denmark
Meeting venue(s) and dates	30 September – 4 October 2019, Copenhagen, Denmark (25 participants)

1 Introduction

1.1 Overview of RDBES and its development

The RDBES is the new Regional DataBase and Estimation System.

The overarching aims of this system are:

1. To ensure that data can be made available for the coordination of regional fisheries data sampling plans, including for the EU Data Collection Framework (DCF) Regional Coordination Groups (RCGs);
2. To provide a regional estimation system such that statistical estimates of quantities of interest can be produced from sample data;
3. To serve and facilitate the production of fisheries management advice and status reports;
4. To increase the awareness of fisheries data collected by the users of the RDBES and the overall usage of these data.

The RDBES will hold both detailed commercial sampling data and aggregated effort and landings data (Figure 1.1). The system meets several EU-MAP requirements and long-term ICES needs by facilitating the storage of data from statistically sound sampling schemes, both national and regional, alongside core information of the sampling design (e.g. stratification, selection methods and probabilities of inclusion). Estimation algorithms will also be stored within the system so data summaries and estimates produced for a variety of end-users are fully documented. Final integration of the RDBES into the ICES Transparent Assessment Framework (TAF) will secure that RDBES outputs are both transparent and reproducible. The RDBES is expected to replace the previous Regional Database (RDB)[1] and the InterCatch[2] by the end of 2021 (Table 1.1).

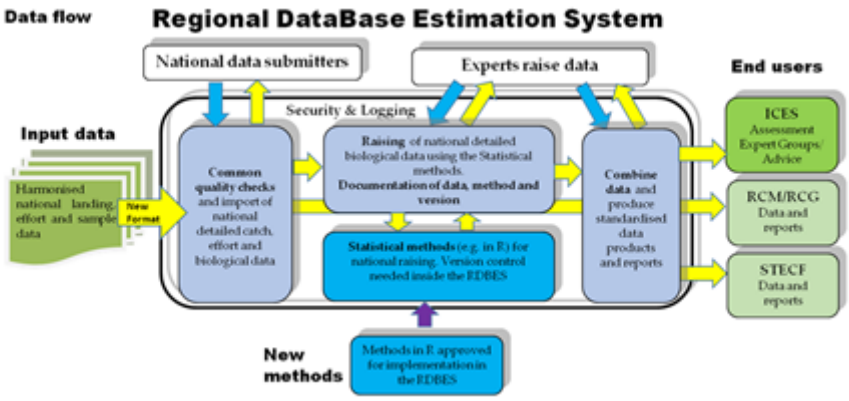


Figure 1.1. Schematic of RDBES (from SC-RDB report, 2018)

[1] The RDB contains detailed commercial fisheries sampling data and aggregated effort and landings data and is hosted and maintained by the ICES Data Centre. The data within the RDB remains the property of the countries that submit the data. https://www.ices.dk/marine-data/data-portals/Pages/RDB-FishFrame.aspx

[2] Intercatch is an ICES database that contains, amongst other, the national level effort and catch estimates used by many ICES Stock Assessment Groups: more details in https://www.ices.dk/marine-data/data-portals/Pages/InterCatch.aspx

	RDB	InterCatch	RDBES
2019	Production Data in/out	Production Data in/out	Development Test data in/out
2020	Production Data in/out	Production Data in/out	Test by selected stocks
2021	Production Data in/out	Production Data in/out	Test by all stocks
2022	Stay alive Data out	Stay alive Data out	Data call for 2021 data
2023	Stay alive Data out	Stay alive Data out	Data call for 2022 and all years
2024	Terminated	Terminated	Production

Table 1.1. Development roadmap for RDBES, RDB, and InterCatch (adapted from SC-RDB report, 2018, pg 8)

The following actors are directly involved in the development of the RDBES (Figure 1.2):

- The **RDB/RDBES Steering Committee (SCRDB)** is the ICES governance group, which oversees the RDB/RDBES;
- The **ICES Data Centre** is part of the ICES Secretariat and is responsible for maintaining and supporting the existing RDB, and developing the new RDBES;
- The **RDBES Core Group** supports the ICES Data Centre in the RDBES development – membership of this group is open to suitably interested and qualified people. It has the following ToRs: 1) Follow, and advise on the development of the project, 2) Provide substantial input to the user requirement specifications, 3) Be responsive to the ICES Data Centre and provide input to issues in the implementation of the RDBES, 4) Testing and approval of developments;
- The **ICES Secretariat** provides secretarial, administrative, scientific, data handling support and develop web systems to the ICES community.

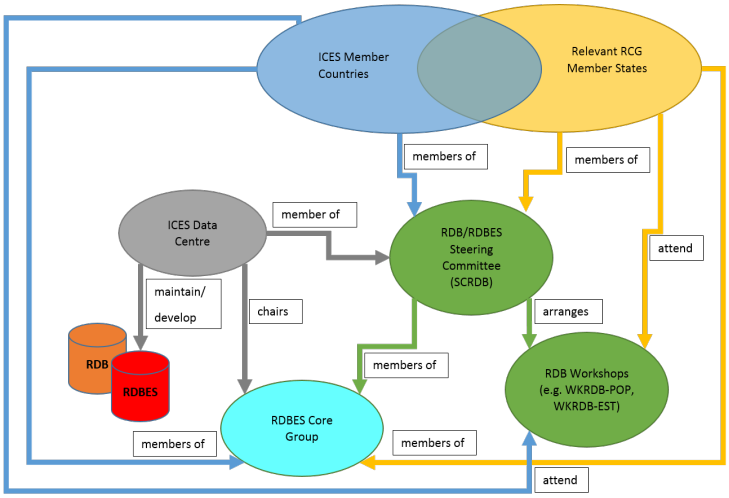


Figure 1.2. Schematic representation of main actors involved in RDBES

Upload of data to and estimation within the RDBES will require significant adaptation of ongoing data collection and data provision processes taking places at national and ICES level. It is expected that areas like national data storage facilities, but also national sampling designs, field protocols, estimation algorithms and routines for data provision to end-users suffer changes. It is also expected that present day ICES roles like Stock Coordinator suffer significant changes, moving towards a more regional, competence- and responsibility sharing, framework.

To secure a soft transition towards the RDBES there is a need to intensify national planning of these adaptations and SC-RDB has planned a series of workshops throughout 2019–2021 with the aim of supporting countries in this transition (see SC-RDB report, 2018, section 3.3). The Workshop on Populating the RDBES data model (WKRDB-POP), chaired by David Currie and Edvin Fuglebakk in February 2019 and the present Workshop on Estimation with the RDBES data model (WKRDB-EST) are two such initiatives.

1.2 Participants and terms of reference for the meeting

The participants list and terms of reference of the Workshop on Estimation with the RDBES data model (WKRDB-EST) are presented in Annex 1 and 2, respectively.

WKRDB-EST was attended by 25 participants from 22 institutes and 15 countries. The countries present included a majority of countries associated to ICES and the European Union but also Cyprus (non-ICES country within the EU), and Norway (ICES country, outside the EU). To support statistical work, ICES funded limited time from an external consultant with vast expertise in statistical estimation of fisheries data (Mary Christman, USA).

1.3 Agenda and structure of the meeting

The agenda adopted in the first day WKRDB-EST is displayed in Annex 4.

In brief, the first day of WKRDB-EST was dedicated to introductory presentations. From then onwards, approximately 1.5 days were spent finalizing the preparation of the datasets and documenting issues in the data model followed by 1.5 days developing the estimation scripts themselves. During the week, a couple of plenary lectures were given by Mary Christman (external consultant) on the topics of "Univariate Estimation" and "Multivariate Estimation". The last half-day of the meeting was spent in plenary discussion on progress achieved and the way forward in terms of development of tools for estimation within the RDBES.

The development work was structured in three subgroups (see Annex 9) with the following ToRs and participants:

Subgroup 1: Upper hierarchies: table DE to SS (excluded)

- Convert objects with RDBES table names to Statistical table names and subset design vars to separate table
- Test function that calculates inclusion probabilities on different examples
- Create code/function that systematically applies functions to tables and generates a list

Subgroup 2: SS, SL and SA tables

- SS, SL table: Discuss format: Commercial species and species selection
- SS, SL table: Generate function that outputs TRUE and FALSE for sampling of a vector of species
- SS, SL table: Code: Generate function to assign inclusion probabilities (if sampling unit)
- SA table: Generate code that can assign inclusion probs to variable degrees of subsampling and stratification

Subgroup 3: Lower Hierarchies: FM and BV

- Create a function where:
    - inputs
        - Table (FM and/or BV)
        - Hierarchy type
            - If B – option "FMvalue" or "BVvalue"

        - Variable (in case of A, C)
    - does
        - Calculates inclusion probabilities
    - outputs
        - Lists (hierarchy dependent)

Subgroups were autonomous with regards to how they organized their work and reported back to plenary at the end of the day.

The present report is structured according to the terms of reference of the meeting. First an overview of the RDBES development is given (Section 1), then term of reference a) "Develop and document R scripts for design based estimation for each hierarchy in the RDBES data model" and b) "identify and document any problems with RDBES data model relating to design based estimation" are covered (Section 2 and 3, respectively). The list of presentations given during the workshop are provided in Section 4. A summary of the discussions held on the way forward for the development of the estimation routines of the RDBES is presented in Section 5.

2 Develop and document R scripts for design-based estimation for each hierarchy in the RDBES data model (ToR a)

WKRDB-EST prepared data for 8 of the 13 upper hierarchies of the RDBES and developed a first set of R-scripts to prepare data ex handle design-based estimation within the system. At the start of the development work an agreement was reached on:

- The preferential use of Base R relative to tidy-verse (to facilitate maintenance);
- The structure of input and output objects of the main estimation objects (matrices with one sample per row) and initial steps of development (see Annex 9);
- Development focusing on producing generic functions capable of handling level-by-level estimation for any of the upper, "middle" (SA, SS, SL) and lower hierarchy tables (FM, BV). (see Annex 9);
- The need to avoidance of hard-coding changes to datasets (e.g. error correction) inside functions (e.g. if the original design is non-probabilistic and SRSWOR needs to be assumed, that assumption should be done by editing the data *prior* to calling any function. This will ensure that assumptions stay recorded explicitly in the preparation scripts;
- The usefulness of having a style guide orienting collaborative process.

The datasets prepared will be used during development, e.g., for testing the estimation with both real sampling data and real population data (section 2.1). A couple of scripts were produced to help prepare the data for upload to the RDBES (section 2.2.1). One script was produced to extract data from the RDBES system and a proof-of-concept developed that tests the capability of a query to the Species List table with regards to the presence, absence and missing information on specific species (section 2.2.2). A set of functions started to be developed that tackle the vast array of estimation issues involved in design-based estimation including, the determination of inclusion probabilities for SRSWR and SRSWOR, generation of estimation objects for upper and lower hierarchies, and a proof-of-concept on the handling of the multiple subsampling levels currently possible on the SA table (Section 2.2.3). A prototype package for estimating catch at age from hierarchy 13 samples with unequal probability selection of hauls, and fish parameters recorded in lower hierarchy C was also produced (Section 2.2.3). This prototype presently runs on a relatively simple dataset. Finally, work was done on a) a script demonstrating the use of simulations to explore post-hoc adjustments of sample probability (section 2.2.4), b) a conceptual approach to the adjustment of inclusion probabilities of ongoing programmes when new vessels enter the fishery (section 2.2.4). Finally, a style guide was produced to guide collaborative work during development work and make collaboration more efficient (section 2.2.5). The data prepared will be kept on the WKRDB-EST SharePoint and permission was obtained from most data providers for the use of the data to further develop the system. All scripts are publically available in either personal or the ICES GitHub[3] (see details in sections below).

Overall, the developments and tests made during WKRDB-EST broadly confirmed the usefulness of the RDBES data model for design-based estimation. However, development is still very much at its beginning and the data model may still suffer some changes (see section 3) which justifies a cautious optimism. The following issues have so far been identified that require further investigation a) the handling of the species list and multiple subsampling levels of SA table (for

[3] https://github.com/ices-eg/WK_RDBES/tree/master/WKRDB-EST

now at proof-of-concept stage), b) the production of a "master" estimation object containing design information from all sampling levels down to individual fish), c) coding adaptations that allow both 2D and 3D matrices, d) handling of situations where individual fish are put directly in the SA table. Adjustments will likely also be needed to adapt the system to the ICES Transparent Assessment Framework (TAF) (see section 4). All these issues need to be solved and quite a lot of programming still needs to be done before the RDBES can be stated as prepared for design-based estimation (see comments on the different functions and list of issues section 3).

The RDBES Core Group will tackle these issues and continue development intersessional. A new WKRDB-EST is forecasted for the end of 2020. The final aim is to produce an R-package that aggregates a) a generic set of estimation functions and b) vignettes documenting design-based estimation in each type of sampling hierarchy.

2.1 Datasets prepared

Hierarchy	Who	Where	All tables ready?	Anonymized?	Upload?	Can be used by core-group for post-WK development?	Description and comments
H1	Ana	SharePoint	Yes	Yes	No	Yes	Subset of At sea sampling data (one fleet, one year); lower hierarchy table needs to be updated
H3	Ana	SharePoint	Yes	Yes	No	Yes	Subset of At sea sampling data (one fleet, one year); some issues in upper hierarchy tables - needs to be checked/updated; lower hierarchy table needs to be updated
H1	Andrew	SharePoint	Yes	Yes	No	No	3 years of port sampling of mackerel from pelagic fleet
H1	Bart	SharePoint	Yes	Yes	No	Yes	Only upper hierarchy for one year at-sea sampling, only top 10 species included
H7	Chun, Karolin, Twan	SharePoint	Yes	Yes	No	Yes	Subset of Demersal Market Sampling (Single species, one year, Q1, Q2, Q3)
H1	David	SharePoint	Yes	Yes	Partial	Yes	At-sea demersal
H5	David	SharePoint	Yes	Yes	Partial	Yes	On-shore port/market sampling
H13	Edvin	SharePoint	Yes	Yes	Yes	Yes	Pilot of Lottery sampling 2018. 14 PSUs, uneq. prob., self-sampling, herring.
H1	Hans	SharePoint	Yes	Yes	Yes	Yes	Subset of the uploaded data
H1	Julia	SharePoint	Yes/No	Yes	Yes	Yes	One year German at sea sampling (2017, North Sea)
H1	Kirsten	SharePoint		Yes	Yes	Yes	Subset of DNK at-sea sampling
H2	Kirsten	SharePoint		Yes	No	Yes	Subset of DNK at-sea sampling
H5	Lucia, Josefina	SharePoint	Yes	Yes	No	Yes	One year and one quarter port sampling for two different sampling programmes

Hier-archy	Who	Where	All tables ready?	Anony-mized?	Upload?	Can be used by core-group for post-WK development?	Description and comments
H8	Kristiin, Ain, Richard	SharePoint	Yes/No	Yes	No	Yes	One year onshore sampling for SPF (missing sprat). Some values need up-dating, missing FM table
H6	Marijus	SharePoint	Yes	Yes	Yes	Yes	One year onshore landing sampling for demersal species. Full biology only for flounder.
H1	Marta, Zuzanna	SharePoint	Yes/No	Yes	No	Yes	Only upper hierarchy for one year onboard sampling, multi-spp
H1	Nuno	SharePoint	Yes	Yes	No	Yes	One year onboard sampling, discard and landings, multi-spp. Some probs need updated
H2	Nuno	SharePoint	Yes	Yes	No	Yes	One year onboard sampling, discards and landings, multi-spp Some probs need updated
H8	Nuno	SharePoint	Yes	Yes	No	Yes	One quarter market sampling, landings, cod.
H1	Sven	SharePoint	Yes	Yes	No	Yes	one year onboard sampling of the stra-tum “western Baltic Sea active trawler”

2.2 R-code developments

2.2.1 Data preparation for upload

Name	Who	Language	What it does	Where it can be found	Comments
MI_RDBES_ExchangeFiles	Dave	R	R project which contains functions to: i) Load data from Irish database, ii) Validate RDBES data frames against ICES xsd files iii) Swap a data frame between R and database column names iv) Create H1, H5, CE, CL exchange file formats v) Create RData files for frames in H1 and H5	https://GitHub.com/davidcurrie2001/MI_RDBES_ExchangeFiles	Code works with my data and I have tried to follow style guide but it is not fully tested yet.
prep_her.R	Edvin	R	converts data from IMR database (NMD biotic) to RDBES v1.17	https://GitHub.com/edvinf/wkrdbest-dataconversion/tree/vWKRDBES-EST	

2.2.2 Extraction from RDBES, data preparation for estimation

Name	Who	Language	What it does	Where it can be found	Comments
v_FishingOperation.sql	Henrik	SQL	Extract data from the RDBES tables into csv files including the reference id fields	WK GitHub	View for a table used for data extract
generate_zeros_in_SA	Nuno & Subgroup2	R	Checks of one spp against SL and generates 0s and NAs in SA table	WK GitHub	Proof-of-concept

2.2.3 Estimation

Name	Who	Language	What it does	Where it can be found	Comments
generate_probs	Nuno	R	Determination of inclusion probabilities	WK GitHub	
handling_of_subsampling_in_SA.R	Nuno	R	Handles subsampling levels of SA into a single line for probInc attribution	WK GitHub	Proof-of-concept
read_sp_data_into_list	Kirsten | Marta	R	Reads in data from the share point, after sync to own computer.		Need to made into a function
generic_su_object_upper_hie	Kristen	R	Generates an estimation object from the upper hierarchies down to the sample table (SA)		
add_probs_to_su_oject	Kirsten	R	Is more or less doing the same as Nuno's generate_probs		
h13estimator	Edvin	R	Prototype package for estimating catch at age from hierarchy 13 samples with unequal probability selection of hauls, and fish parameters recorded in lower hierarchy C.	WK GitHub https://GitHub.com/ices-eg/WK_RDBES/tree/vOct4th2019_WK_end/WKRDB-EST/Personal_folders/EdvinFuglebakk	See Annex 7
Lower	Dave, Andy, Richard, Kristiina, Marijus	R	Function to create design table and probability matrix for lower hierarchies (only hierarchy A at the moment)	https://GitHub.com/ices-eg/WK_RDBES/tree/master/WKRDB-EST/Personal_folders/dave/Lower	Only basic testing done - needs to be linked to upper hierarchy and sample information.

2.2.4 Other scripts

Name	Who	Language	What it does	Where it can be found
Simulation.Rmd	Hans G	R markdown	Simple simulation exploring post-hoc changes to sample probability	https://GitHub.com/ices-eg/WK_RDBES/tree/master/WKRDB-EST/Personal_folders/Hans

2.2.5 Style GuideStyle Guide

Early adoption of a common set of rules for code development (or style guide) is fundamental for efficient collaboration in projects such as the development of estimation routines of the RDBES. Such a common vocabulary and grammar ultimately aims at making editing and testing more collaborative and speeding up the development and building of R-packages.

A proposal of a style guide for R-code development was prepared ahead of WKRDB-EST and reviewed during the meeting. The document proposes a set of style rules to be used in the development of R-code for the RDBES. These style rules were compiled and adapted from different sources, including style guides of bioconductor, google and ICES[4] with a few topics and clarifications being added by WKRDB-EST participants.

The final guide proposed by WKRDB-EST is presented in Annex 8 of this report with a live version being found at the WKRDB-EST SharePoint.

[4] https://github.com/ices-tools-prod/doc/blob/master/README.md

3 Identify and document any problems with RDBES data model relating to design-based estimation (ToR b)

Identification and documentation of issues in RDBES data model that affect design-based estimation was carried out throughout the workshop. Issues became apparent as a consequence of discussions held in subgroups or from individual attempts carried out by participants to populate the data model with new real-life datasets and/or upload their data to the RDBES sandbox (http://sbox.rdbes.ices.dk/). These issues add to the ones detected during the development of the scripts themselves (see section 2).

Issues reported by individual countries can be found in Annex 5. The following main issues were identified in relation to estimation. The issues relate both to modifications and additions to the data model and documentation upload and will be sent to the RDBES Core Group for further discussion.

1. Selection probabilities are needed for the Hansen-Hurwitz (HH) estimator and bootstrapping of variances

    a. Add design variable "selection probability" to all the design tables

    b. In bootstrapping, the sampling units are "sampled" using the sampling design and the selection probabilities. One cannot use inclusion probabilities to do those selections since the inclusion probability refers to the likelihood of the unit being included in the final sample of size n not the likelihood of the unit being selected on a single draw as would be done in the bootstrap.

    c. Update documentation

2. Update the documentation for inclusion probabilities

    a. Improve justification of no need to provide inclusion probabilities when SRSWR or SRSWOR.

        i. Where it presently reads "*In most cases, e.g., when simple random sampling is the selection method of the sample, the column does not need to be filled because the inclusion probabilities can be automatically calculated as n/N where n and N are the sampled and total of each stratum.*" it should read "*in many cases, e.g., when simple random sampling is the selection method of the sample, the column does not need to be filled because the inclusion probabilities can be automatically calculated* ***from the sample size (n) and size (N) reported for*** *each stratum.*"

    b. Mary Christman has provided a document with detailed algorithms for the calculation of both inclusion and selection probabilities, Annex 11. This should go into the data model documentation as an annex.

3. Guidelines are needed on how to populate the species list, when a formal list did not exist for past sampling. Those guidelines should be added to the documentation.

    a. Develop guidelines, consult with WGCATCH and participants of WKRDB-POP

4. A unit identifier is needed for the Horvitz-Thompson (HT) estimator so its formula can be applied in WR situations where a count of unique identifiers is used. At present this identifier is only specified in some of the tables (e.g., TE) but not all (e.g., the VS)

    a. Ensure that a unit identifier is present in all design tables

    b. Include check on identifier when SRSWOR or UPSWOR as in those situations it must be unique (i.e., no duplicates allowed). Core group to check if it is enough identifier to be unique within the sample (no need for uniqueness across samples)

    c. Explain identifier in data model and documentation

5. Documentation of non-responses needs to be improved (see Hans study in chapter 4 showing their importance)

    a. To be discussed in the core group, e.g. at what level is it realistic to include

    b. Update documentation

6. RDBES data model needs adaptation to accurately reflect situations when sample selection is made by commercial name and different biological species can be found within a commercial name

    a. Core-group to incorporate proposal of Annex 6

    b. Update data model and documentation

7. Species selection table (SS). This table is presently default in all hierarchies. It includes design variables but is not always a sampling unit, much more frequently acting like a mere link to a sampling frame (present SL table). This is an inconsistency in the data model but is not unique (the same happens with DE and SD tables). However, given the central position of SS in all upper hierarchies, ambiguity in this table significantly troubles estimation leading to questions such as i) how should it be handled during estimation, when not a SU? ii) is it a sampling unit or a sampling frame? This situation adds to the hierarchical structure in the RDBES data model being a bit ambiguous for the SS and SL table: SS allows linking to SL, and both allow linking to upper-level hierarchy elements, such as FO and LE. For certain applications, like bootstrap, we may need to generate new connections, and it should be clear which foreign keys needs to be filled in.

    a. Core group should consider the above situations and carry out an in-depth analysis of its implications. To the extent possible, it would bring coherence to separate the two concepts, species selection and species frame and only include species selection in the hierarchies that in fact select species. In its present location the table fits one of the main purposes of the data model, allowing the correct

specification of a species frame and generating correctly 0s and missing values at species level (See script "generate_zeros_in_SA.R"). It may however be useful to test a flag indicator that signals if "frequency of occurrency" should be calculated or not from those samples.

b. Even after the previous minor adaptation, the usefulness of SS table as *de facto* Species Selection table is remains to be tested in a hierarchy that requires it. In this regard note that in the present SS table each row is not a sampled unit but an aggregated sample - this situation contrasts that of the remaining selection tables and is likely to hinder estimation. The possibility that a species frame could be specified by some other means and species selection only included in the hierarchies that effectively have species as sample unit should be considered. Any alternative will likely represent a major update to the data model (i.e., a v.2) and should therefore require thorough testing in relation to population and estimation of 0s and missing values.

c. This situation should be analysed in light of the changes proposed for SL table (see section

8. Clarify how to calculate inclusion and selection probabilities when sampling is done by weight not units

    a. See annex 11. Consider adding it as annex in data model documentation.

9. Are adjusted inclusion and selection probabilities needed in the data model?

    a. Evaluate the reasoning behind adjusting and where it should be stated (in data model? In estimation code?)

    b. Discuss importance of adjustments in the context of post-stratification

10. Need to revisit the order of some tables in the upload format namely

    a. Species lists

    b. Auxiliary tables

4 Presentations and lectures

Kirsten Birch Håkansson and Nuno Prista - WKRDB-EST Workshop on Estimation with the RDBES data model

General introduction to the ToRs and Agenda of the WKRDB-EST

Colin Millar, Arni Magnusson - Transparent Assessment Framework (ICES TAF)

The ICES TAF framework was presented. TAF organization into Inputs, Model, Outputs can be applied to estimation scripts. RDBES development should keep close contact with the developments of TAF with of view towards future integration of estimation scripts in TAF. Core group and SC-RDB will have to analyse how that can be done - there are aspects related to data confidentiality, then different types of set-ups for the estimation (see Dave's use roles), and it is not yet clear how these issues can be integrated in TAF (which is at present more centralized - 1 input, 1 model, 1 output)

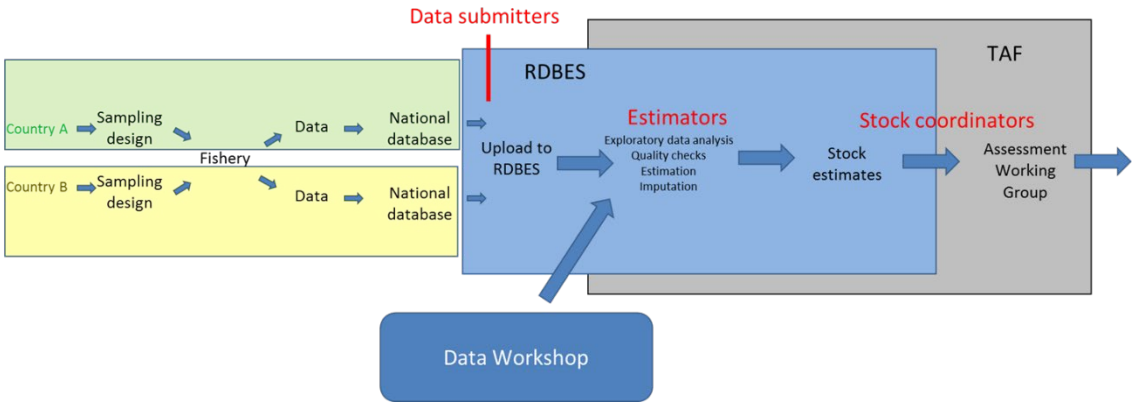
TAF	Design-based Estimation scripts
Input	Download of the data, preparation of views and R objects from RDBES data, assumptions made on data
Model	Estimation routines (determination of inclusion and selection probabilities, estimation via HH and HT estimators, post-stratification
Output	Production of views and graphs for estimated data

David Currie - Never mind the RDB here's the ES

The RDBES will replace both the existing RDB and InterCatch but the majority of the work on the RDBES development to date has involved specifying and populating the data model – what about the estimation?

The RDBES will replace the current system of uploading raised data to InterCatch and the broad work flow will be as follows.

- Before a stock will be included in the RDBES a data workshop must be held.
- This will require the countries that are relevant for that stock to upload data in the new RDBES format. The experts at the workshop will then look at the data and the existing national estimation techniques to agree on the stock estimation techniques that should be used.
- Once the data workshop has been concluded then future data calls for that stock will require countries to have uploaded data to the RDBES by a deadline, rather than submitting raised estimates. The estimates will then be produced by running R scripts on the RDBES data.
- At the data workshop it should be agreed who will create the stock estimates (e.g. will each country create national estimates, which are then combined into stock assessments, will a single person directly create the stock estimates, or will some people create national estimates for multiple countries?)
- The techniques agreed in the data workshop will then be applied within the Transparent Assessment Framework (TAF).
- If there are reasons why the agreed techniques should be modified in a particular year, then it is a national responsibility to inform the estimator(s).
- The stock estimates are then used as an input to the TAF Stock Assessment.



It is a requirement of the RDBES stock estimation that a stock estimate is produced, as well as national estimates for that stock – however national estimates for a country do not necessarily have to be produced by a user from that country.

The following RDBES user roles are relevant for producing stock estimates within the RDBES (note that a single person can hold multiple roles at the same time):

- Data viewer: Can view and export data and estimation scripts for the specific stock area.
- Estimator: Can create and run scripts to create national stock estimates for a specific country and stock area.
- Stock coordinator: Can create and run scripts to produce stock estimates for a specific stock.

Four models of how the user roles could be allocate to different people during RDBES stock estimation were presented using a simple example where only 4 countries contribute data to a stock estimation (Country 1–4). A decision process for how to pick which model to use was also presented

Mary Christman - Design-based Univariate Estimation (Annex 12)

Lecture on univariate estimation

Mary Christman - Design-based Multivariate Estimation (Annex 13)

Lecture on multivariate estimation

Kirsten Birch Håkansson and Nuno Prista - WKRDB - EST: Collaborative SubGroups (Annex 9)

Short presentation outlining the approach followed by development of estimation work during WKRDB-EST, including a discussion of the main types of estimation issues present in upper (all tables to SS, exclusive), "middle" (SS, SL, SA table) and "lower" (lower hierarchy tables, FM and BV) and ToRs for the three subgroups handling these issues. A matrix structure, where rows were samples and columns were parentId, id, and DV variables (and in the case of lower hierarchies also values) was suggested for use in inputs and outputs of the different estimation functions.

Hans Gerritsen - Simple simulation using HH estimator

Some simple simulations were presented, they were based on the examples in Mary Chrisman's presentation: "3) design-based univariate estimation.pptx". The intention was to explore what happens if you adjust your selection probability for estimation purposes (after you have completed your sampling). This might be tempting if the UPS selection probability (e.g. vessel selec-

tion) is based on the landings of each vessel in the previous year. When the current year's landings become available it seems to make sense to adjust the sampling probabilities to the current year's landings.

The simulations seemed to support this, but probably only because the selection probabilities were directly proportional to the variable that was measured. Further explorations showed that this can lead to biased samples.

The main conclusion from the plenary discussion was that it is not appropriate to adjust the selection probabilities as a short-cut to post-stratification. Adjusting the weights is probably a good thing (actual post-stratification) if landings change considerably from year to year or if refusals are high. However this cannot be done inside the HH estimator.

Another discussion point was related to vessels that are removed from the fleet (e.g. after being sold to another country). There are two ways of dealing with this:

- If a vessel that is no longer in the fleet is selected, this can be recorded as a special kind of 'refusal' and the catches of this vessel could be recorded as a real zero.
- Alternatively, the selection probabilities can be adjusted to reflect the zero probability that the vessel is sampled (e.g. increasing the probabilities of the remaining vessels so they sum up to 1 again).

Julia Wischnewski - Vessels selection and Inclusion probabilities: "Reservoir" sampling

A particular attention deserves a situation, when a vessels list at the beginning of the year is still unknown and is expected to change over time. Thus, a vessels sample cannot be selected for the entire year and some sampling rule as well as inclusion probabilities have to be defined for the vessels coming to / leaving the vessels list during the year. Reservoir sampling is one of the sequential sampling algorithms (i.e. real time sampling), which allows to sample from data sampling frame without knowing the size of the sampling frame in advance. The basic idea behind reservoir sampling is following. The vessels sample of the size n(t, M) (i.e. "reservoir"), selected from the available at time t vessels list of size M, can be updated by vessels M+1, M+2, etc., appearing in the vessels list later, at time t+1, t+2, etc. Each new coming vessel is processed one at a time and is either added to the sample or rejected, according the rule of this sampling algorithm. This balances the sample, giving a chance for the new comers to be included as well. Such kind of sampling can be recommended, if multiple generations of the new samples from the updated vessels list during the year are undesirable/impossible, or if a certain new comer is highly desirable to be included in the sample and a quick response from the captain (e.g. about schedule) is expected.

Kirsten Birch Håkansson and Nuno Prista - Way forward in RDBES estimation scripts

Short presentation outlining possibilities for further development of RDBES estimation. Alternatives for design-based estimation and other estimation methods (model assisted), and synergies between RDBES Core Group, SC-RDB and WGCATCH, were discussed. WKRDB-EST participants incentivized to join the RDBES Core Group and participate in further developments.

5 Contribution to the roadmap of development of RDBES estimation

5.1 Training needs for RDBES development

The development of estimation routines for the RDBES aims to be a wide collaborative process, involving not only the core-group of RDBES development, but also those that routinely conduct estimation of commercial catches at national labs. Ideally, a pan-European group of experts will ultimately integrate, collaborate or interact closely in the development of the RDBES estimation routines, either by integrating directly the core-group or by participating in workshops of the type of WKRDB-EST or the Data Workshops currently being planned for 2020 and 2021 (see pg 8 of SC-RDB 2018 report, and section 5.2 of this report).

Efficient collaboration in coding estimation procedures for a multiplicity of sampling hierarchies and sampling designs necessarily requires a common programming language and a common base in terms of R-knowledge, knowledge of statistics (namely survey design and estimation) and knowledge of practices commonly adopted in software development (preparation of packages, use of GitHub, etc.).

WKRDB-EST conducted a small survey among its participants with the aim of identifying to what extent necessary skills were already available within the core-group / WKRDB-EST "community" and identify areas where training is still needed to ensure the pool of knowledge required for development of the estimation routines. While acknowledging that many other people exist at national institutes that have the necessary skills and will ultimately join and contribute to the development[5], it makes sense to consider the 25 participants of WKRDB-EST as a best-case scenario in terms of expertise and the likely starting pool of interested parties uniting the skills, interests and possibly also the time available to further participate in the process.

The survey results are available in the RDBES SharePoint, with aggregated data being displayed below. The table shows that a reasonable group of programmers is available to participate in RDBES development but that skills in survey sampling and estimation, collaboration in software development, and production of R-packages need to be further developed.

Skill set	Comfortable in using	Minor or no use	Non-response
R: base R	18	2	5
R: tidyverse-related pkgs	12	8	5
R: data.table pkg	7	13	5
R: survey pkg	7	14	4
R: sampling pkg	3	18	4
R: writing own functions in R	13	8	4

[5] It is worth noting that the core-group of RDBES development is permanently open to new participants

Skill set	Comfortable in using	Minor or no use	Non-response
Stats: Unequal prob sampling and estimation	3	18	4
Collaboration: GitHub	10	11	4
Packaging: Roxigen2	4	17	4
Packaging: Upload to CRAN	3	18	4

With regards to knowledge on survey design and estimation even though some training can be provided during WKs (see for example, lectures given by Mary Christman during this WK) and other type of EGs **it is recommended that SC-RDB engages with WGCATCH to jointly evaluate the possibility of setting up of a new training cycle in statistically sound sampling design and estimation similar to the one promoted by WGCATCH over the last 5 years.** Training should include domain estimation and post-stratification as these are required to produce estimates at the different levels of spatial and temporal resolution required by end-users. Such training cycle would not only be important to support to the proper development and review of estimation routines within the RDBES but also be beneficial as a support to ICES countries that need to plan statistically sound catch sampling programmes under the EU-MAP and as a support to data providers that need to identify the sampling hierarchy that best fits their countries data before uploading their datasets to RDBES.

With regards to the use of collaborative tools during software development and production of R-packages, that expertise is generally available at the IT departments of European fisheries institutes and is likely also available at the ICES Data Centre itself. Therefore, **WKRDB-EST recommends that SC-RDB and ICES Data Centre jointly evaluate the possibility of conducting, e.g., a series of webinar or short skype courses, to secure that participants of WKRDB-EST and the core group of development of RDBES are proficient in the use of Git and building of R-packages.**

5.2 Way forward in development of estimation within the RDBES

A roadmap for development and production of the RDBES can be found in section 3.3 of the last SC-RDB report. In the same section, a timeline is also put forward whereby the present RDB system and InterCatch will stay in production until 2021, being replaced by the RDBES in 2022 and terminated in 2024. The development of the RDBES itself is planned to take place in 2019-2021 through a series of WKRDB-POP and WKRDB-EST workshops, coupled with test data calls on a set of selected stocks and data provision workshops that use those data to obtain InterCatch level estimates from data extracted from RDBES. Throughout the 2019-2021 period, the core-group of RDBES and ICES Data Centre will keep supporting development needs, interacting with EGs of interest to the development (PGDATA, WGCATCH, WGBIOP) and the RDBES will be progressively integrated in TAF (in development at ICES Data Centre).

The present WKRDB-EST started the development of R-scrips and functions for design-based estimation based on the RDBES data model. In doing this, the capabilities of the data model for design-based estimation have been confirmed. The data model is now considerably stable, with

only a set of relatively minor adjustments being sent for further consideration in the core-group (See section 3). The core-group of development of the RDBES should **develop an intersessional plan for finalization of the development of design-based scripts and functions for the different hierarchies. The plan should include the tackling of domain estimation and post-stratification which are required to produce the spatial and temporal resolution required by end-users. Development work should be open to WKRDB-EST participants and the outcomes presented in a second WKRDB-EST in late 2020. In this workshop,** final tests using real and simulated data should be performed and the possibility of developing an R package will be examined.

Probabilistic sampling designs and design-based estimation are the baseline scenarios underlying the RDBES and estimation from statistically sound sampling schemes. They are a guarantee of unbiased estimates of totals, means, and variances for the quantities of interest attained with a relatively limited number of underlying assumptions. **Accordingly, WKRDB-EST considers design-based estimation should continue to be a priority for future developments at national and regional sampling schemes and RDBES development**. Still, WKRDB-EST recognizes that, out of historical reasons and/or as an attempt to tackle non-probabilistic sampling and/or better control the variance in low sample size data collection programmes, many ICES countries do not presently implement design-based estimation and rather rely on the use of auxiliary variables - such as catch weights or fishing effort - as raising factors within model-assisted and/or model-based estimators. Discussion on the quality of those methods is strongly needed and other fora (e.g., WGCATCH) are planning work to address it and compare the validity of modelling approaches to the design-based approach. Given the pressing timings of RDBES phase-in and the need to ensure, at least of a first step, the continuity of historical time-series, it is likely that some of the data workshops being considered by SC-RDB for 2020 and 2021 will still involve the coding of such present-day modelling alternatives. **WKRDB-EST suggests that those different estimation initiatives are coordinated in a way that methodologies used can be evaluated and R-code safeguarded within TAF so that reproducibility is ensured and opens the way to the future development of validated model-assisted / model-based estimation procedures for the RDBES. To keep development more controlled, test integration in TAF and better promote the RDBES next to data-providers and end-users WKRDB-EST recommends that the number of stocks targeted by the data workshops in 2020 (at present estimated as n=10) is reduced, with emphasis being first put on simpler case-studies involving a smaller amount of countries and participants. WKRDB-EST highlights that the RDBES uploads are made by sampling scheme and that, to adequately generate the zeros during estimation, it is important that all data collected within a sampling scheme is uploaded (i.e., not a subset of species). Additionally, WKRDB-EST highlights that to estimate variables of interest for a specific stock, countries involved will likely need to upload data from several (sometimes all) sampling schemes. Consequently, to keep work-load manageable during test-phase, it will be better that data call focuses on stocks involving a reduced number of countries instead of trying to limit the number of stocks *per se*.**

Given the importance of the above mentioned processes WKRDB-EST strongly encourages ICES countries in general, and particularly those involved in 2020 and 2021 RDBES data calls, to participate in the population workshops currently being planned as a support to the data-call process (WKRDB-POP).

Annex 1: List of participants

Name	Institute	Email
Ain Lankov	Estonian Marine Institute, University of Tartu, Estonia	ain.lankov@ut.ee
Ana Claudia Fernandes	Instituto Português do Mar e Atmosfera, Portugal	acfernandes@ipma.pt
Andrew Campbell	Marine Institute, Ireland	andrew.campbell@marine.ie
Bart Vanelslander	Research Institute for Agriculture, Fisheries and Food (ILVO), Belgium	bart.vanelslander@ilvo.vlaanderen.be
Chun Chen	Wageningen University & Research, The Netherlands	chun.chen@wur.nl
David Currie	Marine Institute, Ireland	David.Currie@Marine.ie
Edvin Fuglebakk (skype)	Institute of Marine Research, Norway	edvin.fuglebakk@hi.no
Hans Gerritsen	Marine Ireland, Ireland	hans.gerritsen@Marine.ie
Henrik Kjems-Nielsen	ICES	henrikkn@ices.dk
Ioannis Thasitis (skype)	Department of Fisheries and Marine Research. Ministry of Agriculture, Rural Development and Environment, Cyprus	ithasitis@dfmr.moa.gov.cy
Josefina Teruel Gómez	Instituto Español de Oceanografia, Spain	josefina.teruel@ieo.es
Julia Wischnewski	Thünen-Institut (TI), Germany	julia.wischnewski@thuenen.de
Karolina Molla Gazi	Wageningen University & Research, The Netherlands	karolina.mollagazi@wur.nl
Kirsten Birch Håkansson (co-chair)	Danish Technical University, Denmark	kih@aqua.dtu.dk
Kristiina Hommik	Estonian Marine Institute, University of Tartu, Estonia	kristiina.hommik@ut.ee
Laurent Dubroca	Ifremer, France	laurent.dubroca@ifremer.fr
Liz Clarke	Marine Scotland, UK	Liz.Clarke@gov.scot
Lucia Zarauz	AZTI, SPain	lzarauz@azti.es
Marijus Spegys	Marine Research Institute, Klaipeda University, Lithuania	marijus.spegys@apc.ku.lt
Marta Suska	National Marine Fisheries Research Institute, Poland	msuska@mir.gdynia.pl

Name	Institute	Email
Mary Christman (skype, ext. consult)	University of Florida, USA	marycchristman@gmail.com
Nuno Prista (co-chair)	Swedish University Agricultural Sciences, Sweden	nuno.prista@slu.se
Richard Meitern	Institute of Ecology and Earth Sciences, University of Tartu, Estonia	richard.meitern@gmail.com
Sven Stoetera	Thünen-Institut (TI), Germany	sven.stoetera@thuenen.de
Twan Leijzer	Wageningen University & Research, The Netherlands	twan.leijzer@wur.nl
Zuzanna Mirny	National Marine Fisheries Research Institute, Poland	zmirny@mir.gdynia.pl

Annex 2: Resolution

2019/2/FRSG48 The **Workshop on Estimation with the RDBES data model (WKRDB-EST)** co-chaired by Nuno Prista, Sweden and Kirsten Birch Håkansson, Denmark, will meet in ICES HQ, Copenhagen, from 30 September to 4 October 2019 to:

a) Develop and document R scripts for design based estimation for each hierarchy in the RDBES data model.

b) Identify and document any problems with RDBES data model relating to design based estimation.

WKRDB-EST will present a written report to ACOM by 20 December 2019.

Priority	This workshop is considered of very high priority. The activities of this workshop will promote the development of a Regional Database and Estimation System, RDBES by producing the algorithms required for design based estimation under the RDBES data model. The RDBES will work as a database for the Baltic Sea, North Sea & Eastern Arctic, and North Atlantic Regional Coordination Groups (RCGs) and produce the estimates used in ICES Fisheries Advice. The development of the RDBES is concentrating on harmonisation, quality assuring, documentation, approved estimation methods and transparency.
Scientific justification	The RDBES will be extensively used by the RCGs and ICES both to store detailed fisheries sample data and to estimate fisheries related variables used in advice. The RDBES data model secures the structure and variables necessary for design-based estimation but algorithms are necessary that implement the estimation methods and produce the final estimates. **ToR a) Develop and document R scripts for design based estimation for each hierarchy in the RDBES data model.** R-scripts will be developed that implement design-based estimation for the upper and lower hierarchies of the RDBES and produce point estimates of fisheries variables such as catch volumes, numbers-at-length and number-at-age. Development will be based on a set of populated data sets from the different hierarchies, compiled prior to the meeting. The R-code will be documented with associated statistical formulas and used in RDBES documentation. The development of scripts for other estimation methods (e.g., ALK-based estimation, Ratio-Estimation) will not be addressed during the WK (they are left for a future occasion. **ToR b) Identify and document any problems with RDBES data model relating to design based estimation.** The development of R scripts for design-based estimation based on the RDBES data model is an important test point within the development of the RDBES. If during the WK issues are identified that limit the application of design-based estimation in the RDBES, these will be documented and forwarded to the RDBES development group for further discussion.
Resource requirements	The two co-chairs, and potentially the rest of the 5 active members of the RDBES Development Support Core Group will be requested to participate and coordinate algorithm development.
Participants	Max 20 people. Participants should be proficient in writing own scripts and functions in R language.
Secretariat facilities	ICES HQ meeting room and facilities

Financial	No financial implications.
Linkages to advisory committees	There are no direct linkages with the advisory committees, but there is a link to WGCATCH, WGBIOP, WGBYC, PGDATA and the stock assessment Working Groups that will ultimately use the estimates produced within the RDBES.
Linkages to other committees or groups	
Linkages to other organizations	The RDBES estimates are connected to regional data collection defined by the RCGs under the European Commission, EC. The RDBES will also support the ICES countries in providing data for assessment. In the case of EU MS, the RDBES is expected to facilitate and improve the quality of provision of commercial catch data requested under different data calls.

Annex 3: Recommendations

Recommendation	Recipient	Has this recommendation be communicated to the recipient?
Reduce the number of countries involved in 2020 test data calls on RDBES format and data workshops. (see more details in section 5.2)	SC-RDB	Yes
Discuss the possibility of setting up of a new training cycle in statistically sound sampling design and estimation (see more details in section 5.1)	SC-RDB	Yes
Discuss the possibility of conducting training in Git and building of R-packages (see more details in section 5.1)	SC-RDB and ICES Data Centre	Yes
Consider a series of workshops and training courses dedicated to model-assisted and model-based estimation (see more details in section 5.1)	WGCATCH	Yes

Annex 4: Agenda

Agenda (Monday and Tuesday)

Date	Time		Topic
30/09	13:00	Plenary	Welcome 1.Housekeeping **(Lise | Henrik)** 2.What are we here for **(Nuno | Kirsten)** 3.Extended round-the-table **(Nuno | everyone)** 1. Who you are 1. R and Stats knowledge 2. What did you bring 1. Data 2. Estimation scripts 3. Knowledge of R and R-packages 4.Report **(Kirsten | Nuno)**
	14:00	Plenary	Common grounds 1.TAF **(Colin | Ani)** 2.The 'ES' of the RDBES **(Dave)** 3.R Input, set-up and Outputs **(Kirsten | Nuno)** 4.Upload - tips **(Henrik | Dave)**
	15:30		Coffee break
	16:00	Plenary	Design-based Estimators of univariate population parameters **(Mary Christman)**
	18:00		End of meeting - Ice breaker in ICES

01/10	09:00	Plenary	Starting to script designed-based estimators 1.Examples (Kirsten | Nuno | anyone else?)
	10:00	Sub-group	Hands on own scripts and data
	11:00		Coffee break
	11:30	Sub-group	Hands on own scripts and data
	13:00		Lunch
	14:00	Sub-group	Hands on own scripts and data
	15:30		Coffee break
	16:00	Plenary	Design-based Estimators of Multivariate population parameters **(Mary Christman)**
	18:00		End of meeting
	19:30		Social dinner

Science for sustainable seas

Agenda (Wednesday to Friday)

02/10	09:00	Plenary	Round-the-table: work so far **(everyone)**
	10:00	Sub-group	Hands on collaborative scripts and data
	11:00		Coffee break
	11:30	Sub-group	Hands on collaborative scripts and data
	13:00		Lunch
	14:00	Sub-group	Hands on collaborative scripts and data
	15:30		Coffee break
	16:00	Sub-group	Hands on collaborative scripts and data
	18:00		End of meeting

03/10	09:00	Plenary	Round-the-table: work so far **(by subgroup)**
	10:00	Sub-group	Hands on collaborative scripts and data
	11:00		Coffee break
	11:30	Sub-group	Hands on collaborative scripts and data
	13:00		Lunch
	14:00	Sub-group	Hands on collaborative scripts and data
	15:30		Coffee break
	16:00	Plenary	Round-the-table: work so far **(by subgroup)**
	18:00		End of meeting

04/10	9:00	Plenary	Wrap-up; identify needs; WK conclusions and future work / roadmaps for other estimators
	13:00	end of meeting	

Science for sustainable seas

Annex 5: Identify and document any problems in converting national data formats to the RDBES format and uploading to the Sbox

A5.1 Belgium

What (variable | table |hierarchy)	Where (conversion | upload)	Describing the problem
FOhaulNumber	conversion	not clear what this should be, is it different from FOid?
SLid	conversion	confusing description, my original thought when I read SL table description in the "RDBES Data Model.xlsx" was that each record with a species x CatchFrac should have a unique ID, but this is not true
Slyear	conversion	should be “SLyear” (capital letter L)
SL & SA tables	conversion	variable name for catch fraction: SLCatchFrac in SL SAcatchCat in SA would be better to use the same name for both tables
SA table	conversion	Not clear what the difference is between these variables: SAtotalWtLive SAsampWtLive SAtotal SAsampled SAtotalWtMes SAsampWtMes

A5.2 Cyprus

What (variable | table |hierarchy)	Where (conversion | upload)	Describing the problem

A5.3 Denmark

What (variable | table |hierarchy)	Where (conversion | upload)	Describing the problem
VDtype	Conversion	In Denmark, we have 29 different vessels types. Unsure how to convert most of them, so the vast majority will end up as 4=other boats, which is not particularly meaningful.
SDcountry	Conversion	The description in the data model don't fit the codes, which are 'ISO 3166-1 alpha-2 and region codes used by ICES'
VDflagCountry	Conversion	The description in the data model don't fit the codes, which are 'ISO 3166-1 alpha-2 and region codes used by ICES'
SLcatchFraction	Conversion	Why not catchCategory? as in SS and SA
SL	Conversion	The present structure does not work - a solution already accepted, but not implemented
codes in general	Conversion	It would be nice with an overall naming convention for codes, then it would be easier to get them right.

A5.4 Estonia

What (variable | table |hierarchy)	Where (conversion | upload)	Describing the problem
FishingTrip table	Conversion	It would be nice to have FishingArea in this table.

A5.5 France

What (variable | table |hierarchy)	Where (conversion | upload)	Describing the problem

A5.6 Germany, Bremerhaven

What (variable | table |hierarchy)	Where (conversion | upload)	Describing the problem
SS, SL | H1	Conversion	All species are sampled
FOendTime	Conversion	Only start time is available, stop time = start time + duration

A5.7 Germany, Rostock

What (variable | table |hierarchy)	Where (conversion | upload)	Describing the problem
SA | SAtotalWeight	Sample weights (columns 17 18 v. 27, 28)	this seems to be redundant, both sets asks the same information (total W and sampled W) for the aggregation haul (in FO)
FO | FOStatRec	rectangles in EB don't match the codelist	StatRec.xsd lines 545 to 562, code list rectangles contain spaces

A5.8 Ireland

What (variable | table |hierarchy)	Where (conversion | upload)	Describing the problem
Exchange file format	Upload	The Exchange file formats do not specify a very sensible order for the rows. This makes it unnecessarily complicated to 7produce the exchange files and might be introducing errors. The Core Group needs to work with the ICES Data Centre to correct these formats.
Validation data	Upload	I think some of the xsd files do not correctly follow the data model. As mistakes are identified we need to let ICES know so that they can be corrected.

A5.9 Lithuania

What (variable | table |hierarchy)	Where (conversion | upload)	Describing the problem
Exchange file format	Conversion	The most of description in the data model codes don't fit. Need better codes list
FTdepartureTime	Conversion	Mandatory, but not always possible get this data

A5.10 Norway

What (variable | table |hierarchy)	Where (conversion | upload)	Describing the problem
BVStratifictaion	upload	Gave illegal value error, worked when changed to 909
BVunitValue (when BVtype="Age")	upload	Appropriate unit Year, not in reference list. Worked when set to mm.

What (variable | table |hierarchy)	Where (conversion | upload)	Describing the problem
BVunitValue (when BVtype=”sex”)	conversion/upload	Field is mandatory, but no appropriate unit in reference list. See issue above, whether this field should be mandatory. Worked when set to mm
FOrectangle	upload	Gave illegal value error, for apparently legal values. Note whitespace in reference-list (StatRec.xsd). Worked when excluded.
FOfunctionalUnit	upload	M font. Sandbokx requires it present.
SAspecState	upload	Error message on number of record. Had to remove this field for upload to accept SA lines.
FOendTime	conversion	Mandatory, but source data only has startTime. Filled in startTime for endTime.
BV table	conversion	Fields for coding of reason not sampled are not present at the BV table. The particular example that I wanted to handle was ages sampled, but not read because of readability issues with the aging structure (scale or otolith). The SA table contains fields for coding reasons that BV and FM are not sampled *Possible solution:* Quality index for readability. WGBIOP is working with this for e.g. age, maturity
SA table		The way I interpreted the SA table, it always implies some level of clustering in sampling below the ultimate sampling unit above the SL table in the hierarchy. The range of options for this clustering is provided by the reference list for SAunitType. FOr SAunitType “number”" and “Kg”, it is however possible to get samples in an unclustered way. Consider for instance the selection of 30 fish selected at random from the catch as it is transported on a conveyer belt, and the selection of 30 consecutive fish selected from a random location on the same conveyer belt). Later discussions have revealed a reasonable interpretation of the data model where SAunitType “number” and “Kg” might represent unclustered sampling, at this level. Possible solution: clarify documentation.
BVunitValue		BVunitValue is mandatory, but it is unclear to me if it is relevant for categorical variables coded with

What (variable | table |hierarchy)	Where (conversion | upload)	Describing the problem
		values from reference lists (such as sex or maturity).
SA		The distinction between the WtMes and WtLive fields on the SA table is not clearly documented.

A5.11 Poland

What (variable | table |hierarchy)	Where (conversion | upload)	Describing the problem
SAcommSpp	Upload	some species from Polish database are missing on the list (ABK, ACC, ASU, ELP, ENC, FBM, FCC, FPE, FPI, FPP, FRO, FSC, LUM, MOT, MXV, NBU, SME, TRR, TSD, VIV, YEZ)
H1	Upload	too many columns in SA, after deleting SAtotalWtMes, SAsampWtMes, SAconFacMesLive columns, file passed XML conversion
FOstatRec	Upload	In some rectangles in the code list there is additional space
FOnoSampReason	upload	spelling mistake in tRS_Reason For-NotSampling (Not Avalible)

A5.12 Portugal

What (variable | table |hierarchy)	Where (conversion | upload)	Describing the problem
SLid	conversion	All species present in a species list name are included in the table - not sure if SLid is supposed to be unique per species (one line one code) or per name of species list (one code for all lines).
SAid	conversion	Different columns for the same weights from the same aggregation level (e.g. FO): SAtotalWtLive vs SAtotalWtMes and SAsampWtLive vs SAsampWtMes. Not sure if possible differences may justify the presence of all variables in this table.
General: variable names and codes	conversion	Consistency of names/codes across tables. There are different names for the same codes. E.g. catchCategory in tables SA & SS is named catchFraction in table SL; There are the same names for different codes. E.g. 'sampler' in

What (variable | table |hierarchy)	Where (conversion | upload)	Describing the problem
		some tables is for the name of the institution and in others is for 'observer', 'self-sampling', etc.

Spain, IEO and AZTI

What (variable | table |hierarchy)	Where (conversion | upload)	Describing the problem
SDinstitution	conversion	In the list of accepted codes, there are one code per institute and location (i.e. AZTI Sukarrieta and AZTI Pasaia). But we would need just one code per institution (one for IEO and one for AZTI)
OStotal	conversion	It is still not clear to us how shall we deal with the PSU which would be part of a stratum, but are not in our sampling frame (i.e. ports which are not being sampled because of practical constraints). If we want to raise the sampling estimates to those ports to get the totals somewhere. And this situation need to me handle also in the estimation functions
LEsequenceNumber	conversion	each LEid will have a unique LEsequenceNumber? need it to be sequential in a pre-determined order?
SAcommercialSpecies	conversion	In Spain the groups of species which are landed mixed, with the same commercial name is an important issue. We would like to participate in the solutions that are being proposed and test them with our data

A5.14 Sweden

What (variable | table |hierarchy)	Where (conversion | upload)	Describing the problem
SA | SApresentation	Conversion	A code is required for parent samples which subsamples have a mixture of presentation (e.g., whole and gutted). Suggested "Mixture" or "Not Applicable"
SA | SAsex	Conversion	Presently M. Not applicable to many parent samples (e.g., basket of discards)?
All | XXclusterName	Conversion	Name of var not consistent. Suggest "cluster" so it is consistent with "stratum"

What (variable | table |hierarchy)	Where (conversion | upload)	Describing the problem
All	Documentation (xls)	varTypes not fully consistent across tables
FT | FTarvLoc	Documentation (xls)	Basic checks should be “Code list”
FT |SampleDetailsID	Documentation (xls)	Appears as “SampleDetails” in column “Short Description”
OS | SamplingLocation	Documentation (xls)	Basic checks should be “Code list”
LE | LEmixedTrip	Documentation (xls)	yes/no should be “Code list” or set to logical. This is a general issue affecting other variables.
SS | CatchCategory	Documentation (xls)	Basic checks should be “Code list”
SA | SAselectionMethod	Conversion	NotApp and NotSam not defined in code list
SA | SAreasonNotSampledFM	Conversion	options in code list not really adapted to SA/FM table
SA | SAreasonNotSampledBV	Conversion	options in code list not really adapted to SA/FM table
SA | Species code	Conversion	is M. Not applicable to unsorted samples e.g., baskets of discards
FM | FMtype	Conversion	use “Carapace Length” instead of “carapace length”
BV | BVunitScaleList	Documentation (xls)	Field name and R name do not match
BV | BVtype, BVmethod, BVMeasurementEquipment	Conversion	Present code lists incomplete

A5.15 The Netherlands

What (variable | table |hierarchy)	Where (conversion | upload)	Describing the problem
SS, SL | H1	Conversion	We sample whatever is present.
LEhaulNum | LE | H7	Conversion	No link with total haul number
SA | SAtotalWeightMeasured	Conversion	The difference with SAtotalWeightLive is not clear. It seems like both should contain the same weight but if there is both live and measured weight in a stratum the weight of the corresponding stratum is the sum of the live and converted measured weight which is not recorded anywhere.

A5.16 UK and Scotland

What (variable | table |hierarchy)	Where (conversion | upload)	Describing the problem

Annex 6: WKRDB-EST Proposal of new format for SL that accommodates for commercial names

wo long-standing objectives of the RDBES data model are a) to allow the accurate distinction between true zeros and missing values during the estimation and b) to correctly accommodate for sampling of units that may contain a mixture of species (e.g. a box of monkfishes that contained *Lophius piscatorius* and *Lophius budegassa*) and/or sampling units which a species content may not directly correspond to the commercial name of the unit (e.g. a box of *Clupea harengus* that is found to contain S*prattus sprattus*).

In the previous RDB data model only the positive observations (i.e., the species registered in the samples) could be uploaded. This feature led to ambiguity in the distinction of true zeros and missing values for species not registered in a sample, leaving it unclear if the species were effectively absent from the sample (true zero), or that not required to sample (i.e., a missing value), or both. Ambiguous interpretation of such cases can greatly impact the quality of estimates, leading to significant biases in, e.g., landings, discards and frequency of occurrence of, e.g., incidental by-catch species. Additionally, the previous RDB data model did not allow the upload of nested sampling details such as the species content in weight of supra-specific or mixed-species groups found in, e.g., boxes at market. This situation is prone to ambiguities in the identification of the sampling units and to cause error in the calculation of sampling probabilities of collected samples. It also prevents adequate estimation of the proportions of species involved in such groups and the disaggregation of grouped national totals, a calculation that frequently takes place in groups such as monkfishes or rays, and lacks transparency and quality standards.

The Species List table of the RDBES was introduced in the data model with the intention of tackling both previously mentioned quality concerns. In the data model the SL table acts as a species frame, where countries can declare the species they systematically identify and quantify in their sampling programmes. Its presence in the data model, just above the SA table, is meant to allow the accurate determination not only of the species present but also the true absences (i.e., the true zeros) and missing values such as those generated by the use of reduced species sampling objectives (e.g., short or species-focuses species sampling frames). Importantly, this is achieved avoiding the need to individually upload 0s for all absent species. With regards to the statement of supra-specific and mixed species units, a "SLcommSpp" column was included in the model with a link to an aphiaID code list.

WKRDB-EST concluded that in its present form the SL table can be combined with SS and SA tables and correctly identify positive, zero and missing values. A demonstration of the capability of this rearrangement (and the RDBES data model in general) to generate true zeros and NAs using the SL as a sampling frame is provided in script "generate_zeros_in_SA.R". With regards to the supra-specific and mixed species units a re-arrangement of the table is proposed that includes the link between these lower resolution groups and their species content. Details of the re-arrangement proposed and a simplified format suggested for uploads are displayed below.

Species List Details (re-arranged – SLD?)

Order	Short Description	Field Name	R Name	Type	Required	Basic checks	Description	Short Description
1	SpeciesListID	Slid	SLid	int	M		Automatic ID - PK of table	Automatic ID - PK of table
2	Record type	SLrecordType	SLrecType	String	M		Fixed value ('SL')	Fixed value ('SL')
3	SpeciesListName		SLlistName	String	M		The name of the species list	The name of the species list
4	Year	Slyear	Slyear	int	M		Year	Year
7	CatchFraction	SLcatchFraction	SLCatchFrac	String	M		Which catch fraction is this list valid for? (Catch/Lan/dis)	Which catch fraction is this list valid for? (Catch/Lan/dis)

Species List Content (new – SLC?)

Order	Short Description	Field Name	R Name	Type	Required	Basic checks	Description	Short Description
1	SpeciesListContentID	SLCid	SLCid	int			Automatic ID - PK of table	Automatic ID - PK of table
2	SpeciesListID	SLid	SLid	int	M		Foreign key/link to the SpeciesList table	Foreign key/link to the SpeciesList table

Order	Short Description	Field Name	R Name	Type	Required	Basic checks	Description	Short Description
5	Taxa code	SLCtaxaCode	SLCtaxaCode	String	M	Code list	The Aphiaid code of the taxa actually selected as given by www.marinespecies.org. The taxa actually selected can be species (e.g., 126436 = Gadus morhua, i.e., Atlantic cod) but also more general taxonomic groups corresponding to commercial designations present at ports or markets and that may include multiple species. In the latter case use the aphiaID that best approximates the commercial name you selected for (e.g., 125802 = Lophius spp for "Monkfishes nei"; Rajiformes for "Skates nei"; 368409 = Batoidea for "Skates and Rays nei")	The Aphiaid code of the taxa

Species (new – SPP?)

Order	Short Description	Field Name	R Name	Type	Required	Basic checks	Description	Short Description
1	SpeciesID	SPPid	SPPid	int	M		Automatic ID - PK of table	Automatic ID - PK of table

2	SpeciesListConten-tID	SLCid	SLCid	int	M		Foreign key/link to the SpeciesList Con-tent table	Foreign key/link to the SpeciesList Con-tent table
5	Species Code	SPPspeciesCode	SPPsppCode	String	M	Code list	The Aphiaid code of the species con-tained in selected taxa given by www.marinespe-cies.org (127160 for Solea solea).	The Aphiaid code of the species

Example of simplified format suggested for uploads to RDBES: combines Species List Content (new) and Species (new) into a easier to read format

SpeciesListNameID	Species list name	Year	CatchFraction	SpeciesListContentsID	Taxa code	SpeciesID	Species Code
1	List XYZ	2015	Lan	1	125802	1	126555
1	List XYZ	2016	Lan	1	125802	2	126554
1	List XYZ	2017	Lan	2	126436	3	126436

Annex 7: Norwegian case study – estimator for hierarchy 13

GitHub link: https://GitHub.com/ices-eg/WK_RDBES/tree/master/WKRDB-EST/Personal_folders/EdvinFuglebakk

WKRDB-EST hierarchy 13 with unequal probability sampling
Edvin Fuglebakk 10/4/2019

During the workshop I have been developing a package for estimating catch at age in numbers from hierarchy 13 sampling with unequal probability selection of hauls. The RDBES data model v 1.17 was used in the development. Towards ToR a) This may serve as a prototype estimator for this specific sampling scheme, but I have identified few generic coding constructs immediately applicable to other schemes. A few ideas that may be considered for the RDBES estimation specification has come up. These will be discussed below. Towards ToR b) the implementation has served to identify some minor issues, discussed below, but more importantly it verifies that there are very few obstacles to design based estimation of this kind in the data model.

Data model issues identified

- Fields for coding of reason not sampled are not present at the BV table. The particular example that I wanted to handle was ages sampled, but not read because of readability issues with the aging structure (scale or otolith). The SA table contains fields for coding reasons that BV and FM are not sampled
- The hierarchical structure in the RDBES data model is a bit ambiguous for the SS and SL table. SS allows linking to SL, and both allow linking to upper-level hierarchy elements, such as FO and LE. I found that when considering bootstrap methods, it may be important to have a stricter definition of how SS foreign keys should be used, if it is possible to maintain links to higher levels in the hierarchy in either SS or SL, that would simplify this considerably.
- The way I interpret the SA table, it always implies a some level of clustering in sampling below the ultimate sampling unit above the SL table in the hierarchy. The range of options for this clustering is provided by the reference list for SAunitType. FOr SAunitType “number”" and “Kg”, it is however possible to get samples in an unclustered way. Consider for instance the selection of 30 fish selected at random from the catch as it is transported on a conveyor belt, and the selection of 30 consecutive fish selected from a random location on the same conveyer belt). I think the data model, or the data model documentation should be considered to clearly distinguish these two. The issue came up when considering assuming unclustered sampling for sampling that was actually clustered, and discovering that I didn’t know if the data model could tell me how it was actually sampled.
- BVunitValue is mandatory, but it is unclear to me if it is relevant for categorical variables coded with values from reference lists (such as sex or maturity).
- The distinction between the WtMes and WtLive fields on the SA table is not clearly documented.

Ideas for RDBES estimation specification

In order to separate assumptions and approximations from the design based functions, I implemented them with quite strict checks on data, and added data manipulation functions to deal with assumptions. For instance a Hansen-Hurwitz estimator assuming sampling with replacement will stop with an error for samples selected

with without replacement, and a separate function "assumeSelectionMethod" manipulates the selectionMeth fields prior to executing the Hansen-Hurwitz. In this way, assumptions are made quite clear in the final code that puts together the estimators at different levels. See code appendix. I think this was worked well, and would suggest considering the principle when developing the RDBES estimation system.

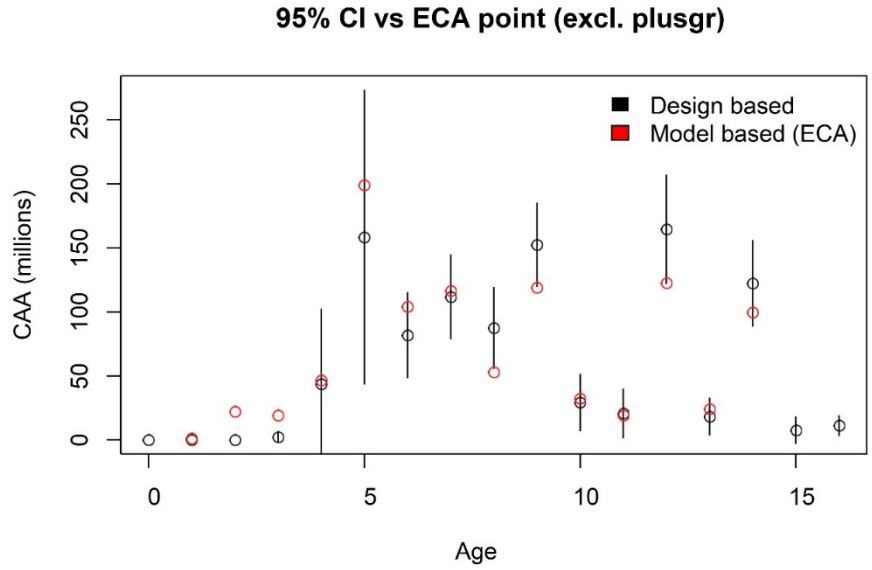


Figure 1 Design based estimates from the Herring lottery pilot (2018) with 95% confidence intervals. Point estimates from model based (ECA) estimates for same stock and year is overlaid.

Evaluation of data model

The data model was well suited for the design based estimation. I implemented an estimator for the Norwegian lottery-sampling, and made estimates from the pilot sampling, with rather low sample size (14 PSUs). Implementation was done with reasonable assumptions on replacement selection at the PSU level, and a possibly less reasonable assumption of the within-haul variances. Those assumptions were suggested or imposed by the sampling design, and not by any restriction in the data model. Below, design based estimates are compared with model based estimates previously obtained for the same stock and year (but with additional samples). The agreement is in my opinion quite good, particularly considering the low sample size in the pilot study.

Code appendix

```
h13estimator:::herringlottery_workflow
## function() {
##             data <- herringlottery
##             proportionsAtAgeBV <-
##             calculateBVProportions(data$BV,
##             "Age",
##             stratified = F)
##             meanWeightsBV <-
##             calculateBVmeans(data$BV,
##             "Weight",
##             stratified = F)
##
##             data$SA <-
##             assumeSelectionMethod(data$SA,
##             "SYSS",
##             "SRSWR")
##             sampleTotals <-
##             estimateSAcaa(
##             data$SA,
##             data$SS,
##             data$SL,
##             "126417",
##             proportionsAtAgeBV,
##             meanWeightsBV,
##             stratified = F
##             )
##
##             data$FO <-
##             assumeSelectionMethod(data$FO,
##             "UPSWOR",
##             "UPSWR")
##
##             haulTotals <-
##             estimateFOCatchAtAge(data$FO,
```

```
##             data$SS,
##             data$SA,
##             sampleTotals,
##             stratified = F)
##
##             grandTotals <-
##             estimateTotalHH(data$FO,
##             haulTotals)
##
##             FOvarZero <-
##             assumeFOconstantVar(haulTotals,
##             0,
##             ages = grandTotals$age)
##             covar <-
##             estimateTotalHHVar(data$FO,
##             grandTotals,
##             haulTotals,
##             FOvarZero)
##
##             report <-
##             makeReportTable(grandTotals,
##             covar)
##
##             return(report)
## }
## <bytecode: 0x7fbf589b1850>
##  <environment: namespace:h13estimator
```

Annex 8: R-Style Guide to be used in RDBES development

styleGuide: codeInR

Contents

1. Style guide

1.1 Why a style guide

The importance of having a style guide orienting the development of RDBES R-code is to facilitate a common vocabulary and grammar in the code that makes editing and testing more collaborative and speeds up the process of package building.

There are many different styles and rules proposed for R-coding (see a sample in section ***References***). ICES itself provides a few general guidelines. The style proposed in this document aims to find a consensus between the main general global style rules proposed by a set of different sources [in brackets] and the personal opinions and preferences of some of those involved in RDBES development.

1.2 Main style advice [2,5,11]

"When coding, use common sense and most of all BE CONSISTENT".

This means a) write your code in consistent and predictable style; and b) if you are editing code made by others, adopt the local style, i.e., take a few minutes to look at the code around you and learn its style. Example: If others use spaces around their *if*-clauses, you should do that too; If their comments have "little boxes of stars around them", make your comments have "little boxes of stars around them" too.

1.3 Dependencies

Be careful when introducing libraries that are not already used in a project. Some libraries come with many additional dependencies, and changes to dependencies can challenge code-maintenance. If a problem can be efficiently solved using dependencies already introduced, consider doing so instead.

For instance, many people prefer coding in *tydiverse*, rather than in *base R* and we certainly don't want to miss out on their inputs. However, code in *tydiverse* will ultimately be more difficult to maintain as it requires multiple dependencies and that adds risk to the projects in terms of long-term maintenance.

So, if you know both *tidyverse* and *base R*, code your functions in *base R*. If you do not know or do not feel as efficient when using *base R*, then stick to *tidyverse* so you can keep contributing. Just be ready to accept your code is translated to *base R* at some stage of the packaging progress.

Also consider license constraints on dependencies. Some libraries come with licenses that restrict the kind of licenses the dependent code may be published under.

2. Naming

2.1 File names [1,2,3,10,11]

- End the file name with ".R" [caps!]
- Separate words in the file name with "_" not "." or " ".
- Be concise and meaningful in the naming
    - name the script of a function with the name of that function;
    - put one function per file
    - if the script prepares data, name it "01_Data_Preparation.R" not script.R.
- If your scripts are part of a larger set of steps pre-fix their filenames with numbers so they are easily sorted
    - e.g., "01_Preparation.R", "02_Model.R" not "Preparation.R" and "Model.R"

2.2 Variable names [1,2,3,4,9]

- Prefer nouns when naming your objects (e.g., results)
- Concise and meaningful
- Use camelCaps: initial lower case, then alternate case between words. [2,3,4, RDBES convention]
- Avoid using names of existing objects (RDBES tables, r-variables, r-functions, r-libraries) [1,9]

2.3 Function names [2,3,4,9]

- Concise and meaningful
- Prefer verbs when naming your functions (e.g., summarize)
- Avoid using names of existing objects (RDBES tables, r-variables, r-functions, r-libraries)
- Use camelCaps: initial lower case, then alternate case between words. [2,3,4, RDBES convention]
- Note: according to the above functions like read.csv today would likely be named readCSV

2.4 Package versions [11]

The tradition for R packages is to use version numbers that consist of three counters, for example 1.2-3. It's practical to have the three counters indicate the nature of changes between releases:

- The first counter (major) is incremented when existing user scripts will not give the same output as before. Breaking backward compatibility with a major release can be inconvenient for users, but is sometimes done to adopt an improved overall design.
- The second counter (minor) means new functions, new arguments, or the like. A minor release suggests that it's worthwhile for the user to read about the new functionality.
- The third counter (patch) is used for other improvements. A patch release may introduce bug fixes, improved documentation, etc.

3. General Syntax

3.1 Spaces [1,2,3,4]

- Use space after comma
- Use space around =, ==, +, -, /, *, <- [1,2]
- No space around :, ::, :::
- Use space before left parenthesis (e.g., in if or operations)
- No space before function calls [1,2,3]
- No space around "=" in function arguments [3,4]

3.2 Curly braces

- ident code inside the braces
- first curly brace on first line (not on its own line) [1,2,3]
- only short statements on the same line (and no curly braces)
- *else* statement surrounded by curly braces [2,3]

3.3 Semicolons [2]

- do not use them - separate instructions into different lines

3.4 Line length [1,2]

- Limit 80 characters so it is readable

3.5 Indentation [1,3,4,6,9]

- use spaces [3,4,9]
- indent the multiple instructions within functions or cycles [1,6]

3.6 Assignment

- use <- not =

3.7 Comments [1,2,3,5]

- use ####### to make sections more visible [1,5]
- add 1 space after #
- use capitals for aspects in development or that require special attention
- short comment above commands [2]
- inline comments separated by 2 spaces [3]

4. Functions

4.1 comment at the start of the function [2]

- If possible, use *roxygen2* to document your functions from the start
- Otherwise, functions should contain a comments section immediately below the function definition line. These comments should consist of a one-sentence description of the function; a list of the function's arguments, denoted by *Args:*, with a description of each (including the data type); and a description of the return value, denoted by *Returns:*. The comments should be descriptive enough that a caller can use the function without reading any of the function's code.

4.2 Order of arguments

- Data inputs first, control options after [2]
- To the extent possible, avoid hard-coding alterations of the input data inside the functions. If you do that, flag them with visible prints so assumptions can later be documented.

4.3 Dependencies [11]

- To the extent possible, avoid calling other packages within functions as these are sometimes not maintained. Ideally require only the core R packages, like base, graphics, and stats [11]
- If you have to call external packages, explicitly name them when you call the function (e.g., call “reshape2::melt” instead of “melt”)

4.4 Testing

- Write a simple test for each function. Whenever a bug comes up, add a test for that bug before fixing it. This makes the code easier to maintain in the long run, and makes it much easier to reimplement code without reintroducing bugs. When programming packages there are good solutions for running automated tests (“testthat” with devtools, for instance)

4.5 Outputs/returns

- Ensure that the data types returned from functions are consistent. So that the functions can easily be included in programs, and not only in interactive-mode. E.g. avoid sometimes returning a matrix, and other times a vector depending on whether ncols or nrows are 1. Document clearly when non-expected data types might be returned (including NULL).

4.6 Global variables

- Variables defined outside the function: use them sparingly, and avoid writing functions that change them.

5. Plots

5.1 Order of arguments

- inputs at start, titles and labels next, general formatting last (cex, las, col, etc)

6. General layout of scripts

6.1 Opening [2,5,6]

- Copyright statement comment
- Author comments
- File description comment, including purpose of program, inputs, and outputs

6.2 Libraries and sources [2,5,6]

- *source()* and *library()* statements
- name all requirements and dependencies

6.3 Executed statements [2]

- Load data
- Transform data
- Outputs

7. Some additional Do's and Don'ts

7.1 Annotations during development

- Use Capitals so others see
  - Warnings: WARNING
  - Specific tunings: ATTENTION
  - Tips for improvements: WISHLIST (better at script start)

7.2 attach()

- avoid using it - risk of confounding variables [2]

7.3 errors

- Use *stop()* inside your functions to check assumptions on inputs (e.g., data type, dimensionality, presence of NAs)
- While developing a new function, signal with *stop()* options that you have not yet developed
- Use clear explicit messages in your stop arguments so others know why code broke or what needs to be developed [2]

7.4 setwd()

- limit its use - better to use a project directory named upfront [5] and take filenames etc. as function arguments. When *setwd()* is used to access resources like conversion-tables or GIS-files, consider packing as R-package, and put resource files in the package directory "inst" or "data" or in sysdata.rda. This avoids having to update *setwd()* when code is transferred to another computer.

7.5 Temporary object names

- v1, v2, v3 - vector
- t1, t2, t3 - table
- m1, m2, m3 - matrices
- df1, df2, df3 - data frames
- ls1, ls2, ls3 – lists
- Remove objects immediately when you don't need it further

7.6 Memory usage

- use *gc()* to clean memory fully - *rm()* cleans the object but not necessarily the memory associated to it

7.7 Loops

- use *apply*, *lapply* etc to avoid for loops [3]

7.8 Saving

- Don't use default workspace to save objects *.RData* [5,6]
- Save *sessionInfo()* so you can remember later what version of R and packages you used to run

that specific code

7.9 Passwords

- Do not store any passwords in the script (alternative, e.g., package keyring)

7.10 Other

- avoid mixing S3 and S4 [2]
- Review and test your code rigorously – once your code is ready, ensure that you test it rigorously on different input parameters. Ensure that the logic used in statements like *for*-loop, *if* statement, *if else* statement are correct. It is a nice idea to get your code reviewed by your colleague to ensure that the work is of high quality. [6]
- vectorize your code [5,6]. See example http://www.win-vector.com/blog/2019/01/what-does-it-mean-to-write-vectorized-code-in-r/

8. References

1

Hadley hickam

http://adv-r.had.co.nz/Style.html

2

google R style

https://google.github.io/styleguide/Rguide.xml

3

R style guide

http://jef.works/R-style-guide/

4

Bioconductor style guide

https://www.bioconductor.org/developers/how-to/coding-style/

5

Best Practices for Writing R

https://swcarpentry.github.io/r-novice-inflammation/06-best-practices-R/

6

R Best Practices: R you writing the R way!

https://www.quantinsti.com/blog/r-best-practices-r-you-writing-the-r-way/

7

John Myles White - Writing Better Statistical Programs in R

http://www.johnmyleswhite.com/notebook/2013/01/24/writing-better-statistical-programs-in-r/

8

CamelCase vs underscores: Revisited

https://whathecode.wordpress.com/2013/02/16/camelcase-vs-underscores-revisited/

CamelCase vs underscores: Scientific showdown

https://whathecode.wordpress.com/2011/02/10/camelcase-vs-underscores-scientific-showdown/

Consistent naming conventions in R

https://www.r-bloggers.com/consistent-naming-conventions-in-r/

9

R Style. An Rchaeological Commentary

https://cran.r-project.org/web/packages/rockchalk/vignettes/Rstyle.pdf

10

R style guide

https://csgillespie.wordpress.com/2010/11/23/r-style-guide/

11

R Package Development at ICES

https://github.com/ices-tools-prod/doc/blob/master/README.md

Annex 9: Proposed format for initial development of estimation code

WKRDB - EST Collaborative SubGroups

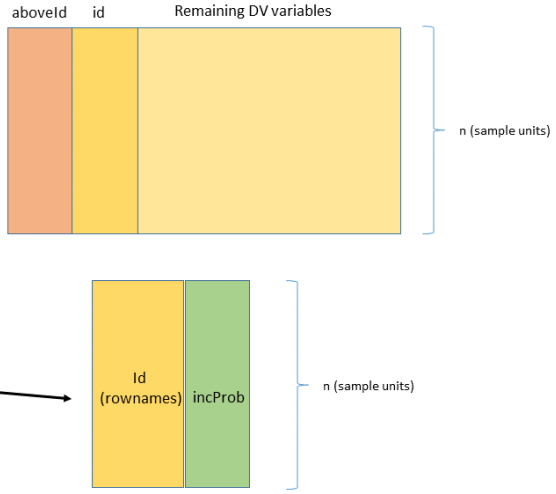
Wednesday and Thursday

ToRs Group 1

- Upper hierarchies: table DE to SS (excluded)
  - Convert objects with RDBES table names to Statistical table names and subset design vars to separate table (test kirsten's function)
  - Test function that calculates inclusion probabilities on different examples (Nuno's function)
  - Create code/function that systematically applies functions to tables and creates a list (suggestion: add that info to output of kirsten's function)

Outputs Group 1

- List
  - Hierarchy
    - Integer [e.g., 8]
  - PSU
    - tableName
      - String [e.g., "TE"]
    - DVtable
      - Data.frame
    - incProbMatrix
      - Matrix
  - SSU
    - …



ToRs Group 2

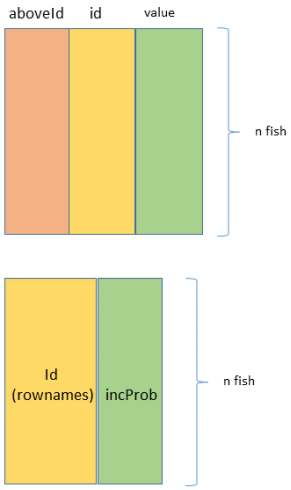
- SS table
  - **Discuss format: Commercial species and species selection**
  - **Code: Generate function that outputs TRUE and FALSE for sampling of a vector of species**
  - Code: Generate function to assign inclusion probabilities (if sampling unit)
- SA table
  - Generate code that can assign inclusion probs to variable degrees of subsampling and stratification

ToRs Group 3

- Lower Hierarchies: FM and BV
  - Create a function:
    - inputs
      - Table (FM and/or BV)
      - Hierarchy type
        - If B – option "FMvalue" or "BVvalue"
      - Variable (in case of A, C)
    - Calculates inclusion probabilities
    - Output
      - Lists (hierarchy dependent)

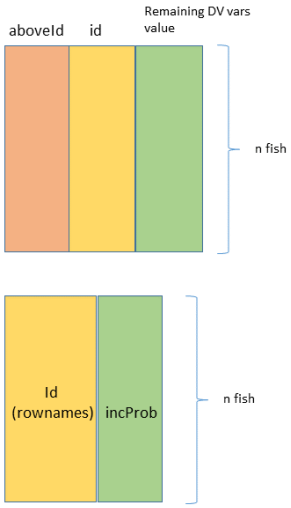
Outputs Group 4: B

- List
  - Hierarchy
    - String ["B"]
  - Value
    - String [FMtype]
  - PSU
    - tableName
      - String ["FM"]
    - designTable
      - Data.frame
    - incProbMatrix
      - Matrix



Outputs Group 4: C

- List
  - Hierarchy
    - String [”C”]
  - PSU
    - tableName
      - String [”BV”]
    - Value
      - String [BVtype]
    - designTable
      - Data.frame
    - incProbMatrix
      - Matrix

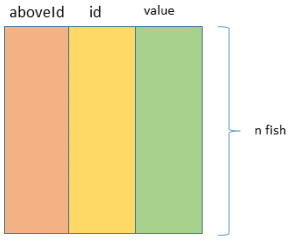


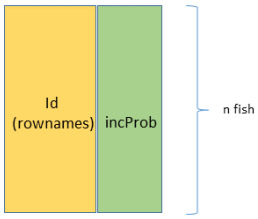
Outputs Group 4: D

- List
  - Hierarchy
    - String [”D”]
  - PSU
    - tableName
      - String [”NA”]
    - designTable
      - Data.frame (empty)
    - incProbMatrix
      - Matrix (empty)

Outputs Group 4:A [option FMvalue]

- List
  - Hierarchy
    - Integer [e.g., B]
  - PSU
    - tableName
      - String ["FM"]
    - designTable
      - Data.frame
    - incProbMatrix
      - Matrix
  - SSU
    - tableName
      - String ["NA"]
    - designTable
      - Data.frame (empty)
    - incProbMatrix
      - Matrix (empty)

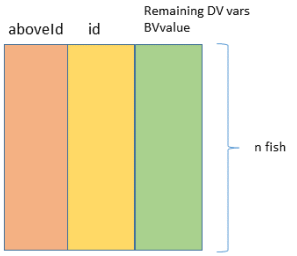


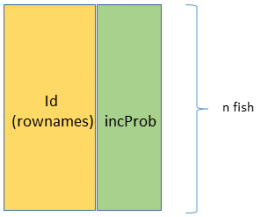


Output format
Equal to B

Outputs Group 4:A [option BVvalue]

- List
  - Hierarchy
    - Integer [e.g., B]
  - PSU
    - tableName
      - String ["FM"]
    - designTable
      - Data.frame
    - incProbMatrix
      - Matrix
  - SSU
    - tableName
      - String ["NA"]
    - designTable
      - Data.frame (empty)
    - incProbMatrix
      - Matrix (empty)





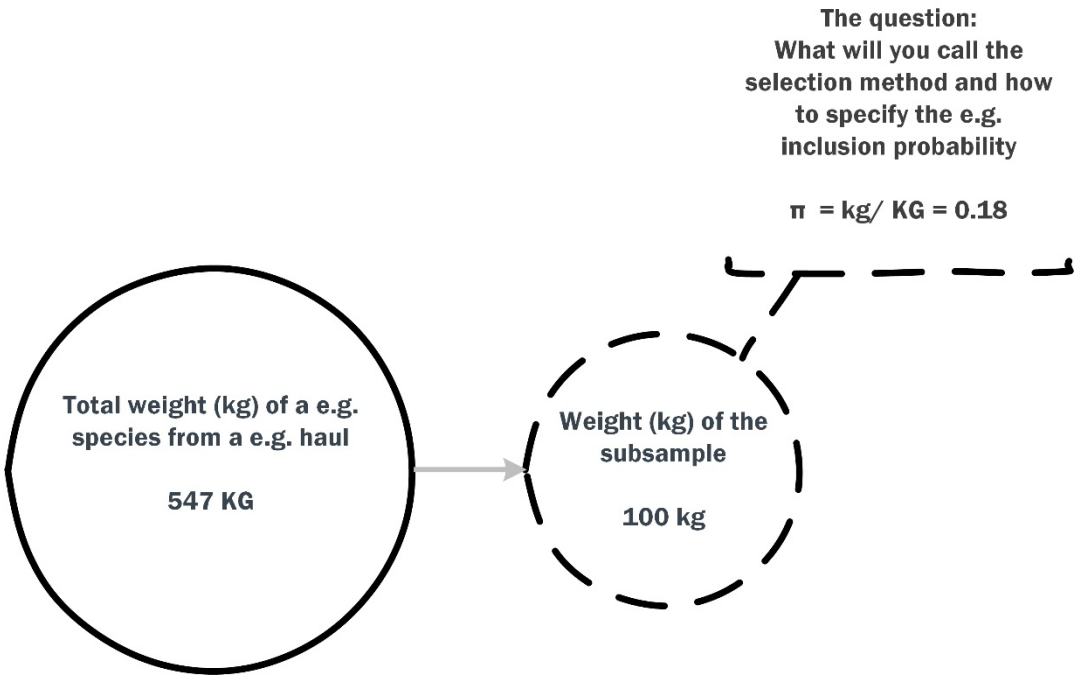
Output format
Equal to C

Annex 10: Sampling weight not unit

Kirsten Birch Håkansson and Mary Christman

A common approach to sampling a pool of fish from e.g. a haul at-sea would be to take an amount of fish out of the total amount of fish. This is not a probabilistic selection of a unit e.g. basket or time, so how should we handle this? In respect to …

1. Selection method
2. Selection / inclusion probabilities
3. Estimation of CV's



First let me note that your $\pi = \frac{kg}{KG}$ *is not an inclusion probability but just a proportion. Recall that an inclusion probability is the probability that a sampling unit (here the* kg *subsample) is included in the final sample of* n *sampling units. See my comments below for further clarification.*

Say, the haul consists of total weight TW and the sample consists of SW, where SW ≤ TW.

To my mind there are 2 possible ways to handle such an issue.

1) *The haul consists of approximately* $N = \left\lfloor \frac{TW}{SQ} \right\rfloor$ *(where* $\lfloor a \rfloor$ *is the floor function, the largest integer smaller than* a*) possible sampling units. One unit, i.e.* $n = 1$*, was selected at random and without replacement (SRSWOR). Hence the selection and inclusion probabilities are the same and equal to* $\frac{1}{N}$*. There is no such thing as a joint inclusion probability since there is only 1 unit was sampled. And, the variance is 0. This is probably the more accurate approach in the sense that it reflects the actual sampling design, but you don't have any way of measuring sampling variability.*

So for the example above;

$$N = floor(547/100) = 5 \text{ and } \pi = 1/5 = 0.2 \text{ and } p1 = 1/5 = 0.2$$

2) *The haul, TW, consists of N fish, where N is unknown. N would be estimated using either (a) length-weight relationships and proportions at length from the sample or (b) by dividing the total weight TW by the mean weight of a fish. Now, SW consists of n sampled fish where n is known but a random number. What is often done is to assume (a) the number of fish to be sampled, n, is a fixed number (i.e. you planned it) and (b) the fish were sampled at random and with replacement (SRSWR). If these 2 assumptions are reasonable, then the selection probability is $\tau_i = \frac{1}{N}$, the inclusion probability is $\pi_i = 1 - \left(1 - \frac{1}{N}\right)^n$, and the pairwise joint inclusion probability is given by $\pi_{ij} = 1 - 2\left(1 - \frac{1}{N}\right)^n + \left(1 - \frac{2}{N}\right)^n$. In this approach all of the numbers used in the calculations of the probabilities are approximate but you can also estimate a sampling variance for the estimated quantities of interest.*

Questions:

Assumption a)

In Denmark we have many self-sampling schemes for small pelagic a la the Norwegian herring lottery (unfortunately presently without the smart selection of the hauls). In the sampling protocols for the fishermen we ask them to take a subsample of the hauls. The size of the subsample is planned, but specified in kg e.g. in the sandeel fishery we ask for 3 kilos (very small fish). The size of the subsamples are set per species so we get approximately x number of fish. We never get exactly x number of fish, but often more fish, so

1. So is it better to always take 100 fish from the subsample instead of taking the whole sample of x kg?
2. Why is it so important that the number of fish is planned?

The reason that the sample size being random could be an issue is that its randomness is an additional source of sampling variability that is not accounted for in the estimators for variance. Recall that variance is calculated assuming that the sampling approach is done identically the same way every time. Since each sampling event would yield different numbers of fish, the sampling scheme as assumed here to be SRSWR of individual fish is not the same exact procedure every time. Hence random sample size adds a bit of variability to the estimates that wouldn't be there otherwise (actually, so does the estimated total number of fish in the haul, N, but we are ignoring that here).

In one sense, asking the fishers to select 100 fish at random would be the better approach since n would now be fixed so that the sampling is identical every time. On the other hand, herring within a haul probably do not vary too much in size, so I expect that the number of fish is not very variable from subsample to subsample. Hence, it really isn't too much of an issue in my mind. It might introduce additional variability but that will be small relative to the total variability.

A question for you: can you be sure that the selection of fish by the fisher is random? Since they are handling individual fish, could they be tempted to get a "representative sample" of sizes (hence overweighting of the rarer small and larger fish)? Instead, asking them to sample X kgs at random is likely more "random" in a sense since they are less likely to notice the fish lengths in the subsample.

Assumption b)

So here it will be the usual rule – ok to assume WR if n/N < 10%.

Exactly.

Annex 11: Algorithms for the calculation of both inclusion and selection probabilities

By Mary Christman

Probabilities at a given level in a sampling scheme

Type of Sampling	**Selection Probability**	**Inclusion Probability (π_i)**
SRSWR	$\frac{1}{N}$	$1-\left(1-\frac{1}{N}\right)^n$
SRSWOR	See below	$\frac{n}{N}$
UPSWR	p_i	$1-(1-p_i)^n$
UPSWOR	See below	$\sum_{k=1}^{n} p_i(k)$

Notation: let

N = number of elements in the population, i.e. the population "size",

n = the number of elements to be sampled, i.e. the sample size,

WITH REPLACEMENT

p_i = the probability of selection with replacement of the i^{th} element in the population, $i = 1, \dots, N$, where $\sum_i p_i = 1$

WITHOUT REPLACEMENT

k = the order in which the i^{th} element is sampled from the population when sampling is without replacement

$p_{i_j}(k)$ = probability that the i^{th} unit is selected on the j^{th} draw when k draws have occurred, $j \leq k$. For example, $p_{i_1}(2)$ is the probability that the i^{th} unit was selected on the 1^{st} draw (hence the i_1) when $k = 2$ draws have occurred so far

$\frac{p_{i_k}(k)}{1-p_{i_1}(k)-p_{i_2}(k)-\cdots-p_{i_{(k-1)}}(k)} = \frac{p_{i_j}(k)}{1-\sum_{m=1}^{(k-1)} p_{i_m}(k)}$ =conditional probability that the i_j^{th} element is selected on the k^{th} draw given that the $i_m^{th}, m = 1,2, \dots, (j-1)$ elements where selected prior to the k^{th} draw, e.g. $\frac{p_{i_2}(2)}{1-p_{i_1}(2)}$ is the probability that the i_2^{th} element is selected on the 2^{nd} draw given that the i_1^{th} element was selected on the first draw.

$p_i(k) = \sum_{(k-1),i} p_{i_1}(1) \times \frac{p_{i_2}(2)}{1-p_{i_1}(2)} \times \frac{p_{i_3}(3)}{1-p_{i_1}(3)-p_{i_2}(3)} \times \cdots \times \frac{p_{i_k}(k)}{1-p_{i_1}(k)-p_{i_2}(k)-\cdots-p_{i_{(k-1)}}(k)} =$ the unconditional probability that the i^{th} element is selected at the k^{th} draw where the summation notation indicates that the sum is over all possible sets of $(i_1, i_2, \dots, i_{(k-1)})$ where the i_m are different integers between 1 and N and none equal i. For each k, $\sum_i p_i(k) = 1$

Example calculations using SRSWOR

1) The probability of the i^{th} element being selected on the first draw is

$$p_i(1) = \frac{1}{N}$$

and the probability of not being selected in the first draw is

$$p_i^C(1) = 1 - \frac{1}{N}.$$

2) The probability that the i^{th} element is not selected on the first draw but is selected on the second draw is

$$p_i(2) = \Pr(i^{th} \notin 1^{st}\ draw) \times \Pr(i^{th}\ in\ 2^{nd}\ draw | \notin 1^{st} draw) = \left(1 - \frac{1}{N}\right)\frac{1}{N-1} = \frac{1}{N}$$

3) The probability that the i^{th} element is selected on the k^{th} draw ($2 < k \leq n < N$) is

$$\begin{aligned} p_i(k) &= \Pr(i^{th}\ not\ in\ 1^{st}\ k-1\ draws) \times \Pr(i^{th}\ in\ k^{th}\ draw | not\ in\ (k-1)\ draws) \\ &= \left(1 - \frac{1}{N}\right)\left(1 - \frac{1}{N-1}\right)\left(1 - \frac{1}{N-2}\right)\cdots\left(1 - \frac{1}{N-k-1}\right)\left(\frac{1}{N-k+1}\right) \\ &= \frac{1}{(N-k+1)}\prod_{j=0}^{k-2}\left(1 - \frac{1}{N-j}\right) = \frac{1}{N} \end{aligned}$$

4) The inclusion probability for the i^{th} element is

$$\sum_{k=1}^{n} p_i(k) = \sum_{k=1}^{n} \frac{1}{N} = \frac{n}{N}$$

Note that the probabilities are identical only because every element has the same chance of being selected at each draw, so the equations for the probabilities at each draw are simplified. This is not the case for unequal probabilities.

Also see the power point presentation "2) Selection and Inclusion Probabilities.pptx".

Annex 12: Design-Based Univariate Estimation presentation

By Mary Christman

3) Design-Based Univariate Estimation 093019.pptx

HANSEN-HURWITZ ESTIMATORS

SINGLE STAGE DESIGNS

HANSEN-HURWITZ (HH) ESTIMATORS

- The HH estimator can be used
  - When sampling is **with replacement and for either equal or unequal selection probability** designs.
  - It uses all observations regardless of whether they have been observed once or multiple times (same unit) in sampling.
  - Advantages include:
    - Relatively easy estimation, even in multi-stage designs
    - Always have variance estimate from the sample

HANSEN-HURWITZ (HH) ESTIMATOR OF THE POPULATION TOTAL: 1 – STAGE UNSTRATIFIED DESIGN

- The HH estimator of the population total (T) and its variance estimator are given by

$$\hat{T}_{HH} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i}$$

$$\widehat{var}(\hat{T}_{HH}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{p_i} - \hat{T}_{HH}\right)^2$$

- The HH estimator of the population mean (μ) and its variance estimator are given by

$$\hat{\mu}_{HH} = \frac{\hat{T}_{HH}}{N} = \frac{1}{Nn}\sum_{i=1}^{n}\frac{y_i}{p_i}$$

$$\widehat{var}(\hat{\mu}_{HH}) = \frac{1}{N^2}\widehat{var}(\hat{\mu}_{HH})$$

HORVITZ-THOMPSON ESTIMATORS

SINGLE STAGE DESIGNS

HORVITZ-THOMPSON (HT) ESTIMATORS

- The HT estimator can be used
  - When sampling is **with or without replacement and for either equal or unequal selection probability** designs.
  - It is based on only the distinct units in the sample, i.e. data from units that may have been sampled more than once, as might occur in with replacement sampling, appear only once in the estimator.
  - Advantages include:
    - Variance of the HT estimator is sometimes smaller than that of the HH estimator

HORVITZ-THOMPSON ESTIMATOR OF THE POPULATION TOTAL: 1 – STAGE UNSTRATIFIED DESIGN

- The HT estimator of the population total (T) and its variance estimator are given by

$$\hat{T}_{HT} = \sum_{i=1}^{\nu} \frac{y_i}{\pi_i}$$

$$\widehat{var}_{HT}(\hat{T}_{HT}) = \sum_{i=1}^{\nu} \left(\frac{1}{\pi_i^2} - \frac{1}{\pi_i}\right) y_i^2 + 2 \sum_{i=1}^{\nu} \sum_{j>i} \left(\frac{1}{\pi_i \pi_j} - \frac{1}{\pi_{ij}}\right) y_i y_j$$

where ν $(\leq n)$ is the number of distinct units that were measured in the sampling

HH VS HT ESTIMATOR OF THE POPULATION TOTAL FOR A 1 – STAGE UNSTRATIFIED DESIGN

- The HH estimator of the population total (T) and its variance estimator are given by

$$\hat{T}_{HH} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{p_i}$$

$$\widehat{var}(\hat{T}_{HH}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{y_i}{p_i} - \hat{T}_{HH}\right)^2$$

- The HT estimator of the population total (T) and its variance estimator are given by

$$\hat{T}_{HT} = \sum_{i=1}^{\nu} \frac{y_i}{\pi_i}$$

$$\widehat{var}_{HT}(\hat{T}_{HT}) = \sum_{i=1}^{\nu} \left(\frac{1}{\pi_i^2} - \frac{1}{\pi_i}\right) y_i^2 + 2 \sum_{i=1}^{\nu} \sum_{j>i} \left(\frac{1}{\pi_i \pi_j} - \frac{1}{\pi_{ij}}\right) y_i y_j$$

VARIANCE ESTIMATION FOR THE HT ESTIMATOR OF T

- Some sampling designs are problematic –
  - Variance estimator ($\widehat{var}_{HT}$) on the previous page may be
    - Undefined, e.g. when one or more $\pi_{ij} = 0$; or
    - Negative when calculated from a sample
  - Sampling designs may not provide easy way to calculate π_{ij}
- Alternative variance estimators are available

ALTERNATIVE VARIANCE ESTIMATOR FOR THE HT ESTIMATOR OF T

- Brewer and Hanif (1983):

$$\widehat{var}_{BH}(\hat{T}_{HT}) = \frac{(N-\nu)}{N\nu(\nu-1)} \sum_{i=1}^{\nu} \left(\frac{\nu y_i}{\pi_i} - \hat{T}_{HT} \right)^2$$

where $\frac{(N-\nu)}{N}$ is the finite population correction factor that can be ignored if sampling is with replacement

- May be positively biased in small samples and so is conservative
- Always non-negative

ALTERNATIVE VARIANCE ESTIMATOR FOR THE HT ESTIMATOR OF T

- Yates and Grundy (1953):

$$\widehat{var}_{YG}(\hat{T}_{HT}) = \sum_{i=1}^{\nu} \sum_{j<i} \left(\frac{\pi_i \pi_i - \pi_{ij}}{\pi_{ij}} \right) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2$$

  - Can only be used when ν is fixed, not random
    - Implies that this should only be used for WOR designs
  - Requires that all $\pi_{ij} > 0$

HORVITZ-THOMPSON ESTIMATOR OF THE POPULATION MEAN: 1 – STAGE DESIGN

- The HT estimator of the population mean (μ) and its variance estimator are given by

$$\hat{\mu}_{HT} = \frac{\hat{T}_{HT}}{N} = \frac{1}{N} \sum_{i=1}^{\nu} \frac{y_i}{\pi_i}$$

$$\widehat{var}(\hat{\mu}_{HT}) = \left(\frac{1}{N} \right)^2 \widehat{var}(\hat{T}_{HT})$$

where $\widehat{var}(\hat{T}_{HT})$ is based on one of the variance estimators for $\hat{T}_{HT}$

HH VS HT ESTIMATOR OF THE POPULATION MEAN FOR A 1 – STAGE UNSTRATIFIED DESIGN

- The HH estimator of the population mean (μ) and its variance estimator are given by

$$\hat{\mu}_{HH} = \frac{\hat{T}_{HH}}{N} = \frac{1}{Nn}\sum_{i=1}^{n}\frac{y_i}{p_i}$$

$$\widehat{var}(\hat{\mu}_{HH}) = \frac{1}{N^2}\widehat{var}(\hat{\mu}_{HH})$$

- The HT estimator of the population mean (μ) and its variance estimator are given by

$$\hat{\mu}_{HT} = \frac{\hat{T}_{HT}}{N} = \frac{1}{N}\sum_{i=1}^{\nu}\frac{y_i}{\pi_i}$$

$$\widehat{var}(\hat{\mu}_{HT}) = \left(\frac{1}{N}\right)^2 \widehat{var}(\hat{T}_{HT})$$

COMPARISON OF HH, HT, AND GHT ESTIMATORS

HH Estimator

- Advantages:
  - Easy to use
- Disadvantages:
  - Only for WR designs
  - Poor variance if p_i not correlated with y_i

HT Estimators

- Advantages:
  - Can be used with WR and WOR designs for which can calculate inclusion probabilities
- Disadvantages:
  - Calculation of joint inclusion probabilities can be problematic
  - Variance estimation can eb difficult
  - Poor variance if π_i not correlated with y_i

GENERALIZED HORVITZ-THOMPSON ESTIMATOR OF THE POPULATION MEAN: 1-STAGE DESIGN

- Useful when
  - The inclusion probabilities, π_i, are not well related to the variable of interest
  - The population size, N, is unknown
- Generalized HT estimator is given by

$$\hat{\mu}_{GHT} = \frac{\sum_{i=1}^{\nu} \frac{y_i}{\pi_i}}{\sum_{i=1}^{\nu} \frac{1}{\pi_i}}$$

  - $\sum_{i=1}^{\nu} \frac{1}{\pi_i}$ is an unbiased estimate of N
  - $\hat{\mu}_{GHT}$ has slight bias which tends to decrease with increasing sample size

VARIANCE ESTIMATOR OF $\hat{\mu}_{GHT}$ FOR 1-STAGE DESIGN

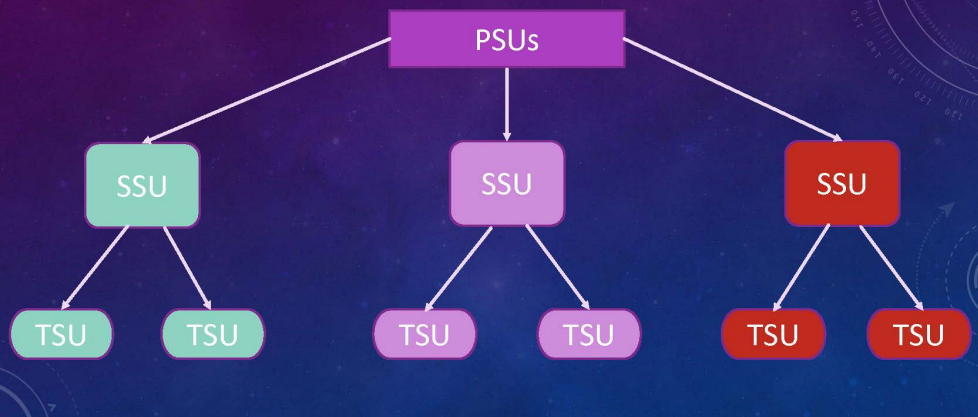
- The estimator of the variance of the estimated mean is given by

$$\widehat{var}(\hat{\mu}_{GHT}) = \frac{1}{N^2}\left[\sum_{i=1}^{\nu}\left(\frac{1-\pi_i}{\pi_i^2}\right)(y_i - \hat{\mu}_{GHT})^2 + \sum_{i=1}^{\nu}\sum_{j\neq i}\left(\frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j}\right)\frac{(y_i - \hat{\mu}_{GHT})(y_j - \hat{\mu}_{GHT})}{\pi_{ij}}\right]$$

- Requires all $\pi_{ij} > 0$
- If N is unknown replace with $\sum_{i=1}^{\nu} \frac{1}{\pi_i}$

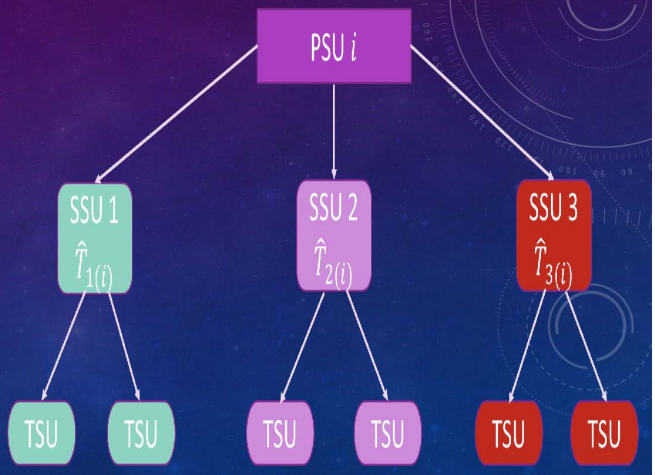
3 – STAGE DESIGNS

REMINDER OF A 3-STAGE DESIGN



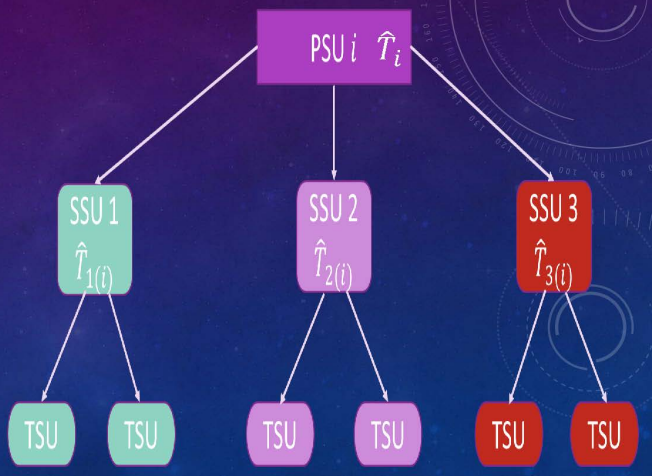
GENERAL APPROACH FOR ESTIMATING THE POPULATION TOTAL

1. Start by estimating the total for each sampled SSU based on the random sample of TSUs in each SSU following the design used to select the TSUs in each SSU
   - Denote these as $\hat{T}_{j(i)}$, $j = 1, \ldots, m_i;\ i = 1, \ldots, n$



GENERAL APPROACH FOR ESTIMATING THE POPULATION TOTAL

2. Estimate the totals for each sampled PSU based on the random sample of SSUs within each PSU following the design used to select the SSUs in each PSU
   - Denote these as $\hat{T}_i, i = 1, \ldots, n$



GENERAL APPROACH FOR ESTIMATING THE POPULATION TOTAL

- Having done:
  1. Estimating the total for each sampled SSU $\hat{T}_{j(i)},\ j = 1, \dots, m_i;\ i = 1, \dots, n$ based on the random sample of TSUs in each SSU following the design used to select the TSUs in each SSU
  2. Estimating the totals for each sampled PSU $\hat{T}_i, i = 1, \dots, n$ based on the random sample of SSUs within each PSU following the design used to select the SSUs in each PSU
- Finish by
  3. Estimating the population total $\hat{T}$ based on the random sample of PSUs following the design used to select the PSUs
     - $\hat{T}$ can then be used for estimating means

HANSEN-HURWITZ ESTIMATORS

3 – STAGE DESIGNS

HANSEN-HURWITZ (HH) ESTIMATOR OF THE POPULATION TOTAL: 3 – STAGE DESIGN

- For a 3–stage unstratified WR sampling design, the HH estimator of the population total (T) is given by

$$\hat{T}_{HH} = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{T}_{i,HH}}{p_i} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{p_i}\frac{1}{m_i}\sum_{j=1}^{m_i}\frac{\hat{T}_{j(i)}}{p_{j(i)}} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{p_i}\frac{1}{m_i}\sum_{j=1}^{m_i}\frac{1}{p_{j(i)}}\frac{1}{u_{j(i)}}\sum_{k=1}^{u_{j(i)}}\frac{y_{k(j,i)}}{p_{k(j,i)}}$$

Each sampled PSU has m_i sampled SSUs

$= \hat{T}_{i,HH}$ = HH Estimator of the total in the i^{th} PSU

Each sampled SSU has $u_{j(i)}$ sampled TSUs

$= \hat{T}_{j(i)}$ = HH Estimator of the total in the j^{th} SSU in the i^{th} PSU

HANSEN-HURWITZ (HH) ESTIMATOR OF THE POPULATION TOTAL : 3 – STAGE WR DESIGN

- Variance estimator for $\hat{T}_{HH}$

$$\widehat{var}(\hat{T}_{HH}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{\hat{T}_{i,HH}}{p_i} - \hat{T}_{HH}\right)^2$$

- This simple form is valid regardless of the number of stages in the design as long as
  - The PSUs are selected independently and WR
  - Subsampling of units selected at any stage t is independent between different units selected at stage t
  - The estimators $\hat{T}_{i,HH}$ are unbiased for the total Y-values of the i^{th} sampled PSU, $i = 1, \dots, n$
  - See Rao (1975), Sankhya C, 37: 133 – 139 for more detail.

HORVITZ-THOMPSON ESTIMATORS

3 – STAGE DESIGNS

HORVITZ-THOMPSON (HT) ESTIMATOR OF THE POPULATION TOTAL: 3 – STAGE DESIGN

- For a 3–stage unstratified sampling design with UPSWR of PSUs and SRSWOR at the other 2 stages, the HT estimator of the population total (T) is given by

$$\hat{T}_{HT} = \sum_{i=1}^{n} \frac{\hat{T}_i}{\pi_i} = \sum_{i=1}^{n} \frac{1}{\pi_i} \underbrace{\frac{M_i}{m_i} \sum_{j=1}^{m_i} \hat{T}_{j(i)}}_{= \hat{T}_i \text{ where } \pi_{j(i)} = \frac{m_i}{M_i} \ \forall j(i)} = \sum_{i=1}^{n} \frac{1}{\pi_i} \sum_{j=1}^{m_i} \underbrace{\frac{U_{j(i)}}{u_{j(i)}} \sum_{k=1}^{u_{j(i)}} y_{k(i,j)}}_{= \hat{T}_{j(i)} \text{ where } \pi_{k(j,i)} = \frac{u_{j(i)}}{U_{j(i)}} \ \forall k(j,i)}$$

HORVITZ-THOMPSON (HT) ESTIMATOR OF THE POPULATION TOTAL: 3 – STAGE DESIGN

- Generically the variance estimator for $\hat{T}_{HT}$ can be decomposed into three parts reflecting the three stages of sampling as three different sources of variation:

$$\widehat{var}\left(\hat{T}_{HT}\right) = \widehat{var}_{PSU} + C_{PSU}\widehat{var}_{SSU} + C_{PSU,SSU}\widehat{var}_{TSU}$$

- For SRSWOR at Stage 3

$$\widehat{var}_{TSU} = \sum_{j=1}^{m_i} \frac{U_{j(i)}(U_{j(i)} - u_{j(i)})}{u_{j(i)}(u_{j(i)} - 1)} \sum_{k=1}^{u_{j(i)}} \left(y_{k(j,i)} - \bar{y}_{j(i)}\right)^2$$

- For SRSWOR at Stage 2

$$\widehat{var}_{SSU} = \frac{M_i(M_i - m_i)}{m_i(m_i - 1)} \sum_{j=1}^{m_i} \left(\hat{T}_{j(i)} - \overline{\hat{T}_{j(i)}}\right)^2$$

- For UPSWR at Stage 1, use the appropriate estimator from those shown for single stage

ESTIMATING POPULATION MEANS FOR MULTI-STAGE DESIGNS

ESTIMATORS OF POPULATION MEANS: 3 – STAGE DESIGN

- There are several population parameters that might be of interest here. These include:
  - μ_3= average value of a TSU (e.g. average weight of a fish landed);
  - μ_2 = average value of an SSU (e.g. mean number of discards/fishing operation in the fishery); and,
  - μ_1 = average value of a PSU (e.g. average annual landings/port in the industry).
- The means are easily estimated using the estimated population total, $\hat{T}_{HH}$.

ESTIMATORS OF POPULATION MEANS: 3 – STAGE DESIGN

- Estimators using either the HH or the HT estimator of the population total :

$$\hat{\mu}_3 = \frac{\hat{T}}{Q} \qquad with\ \widehat{var}(\hat{\mu}_3) = \frac{\widehat{var}(\hat{T})}{Q^2}$$

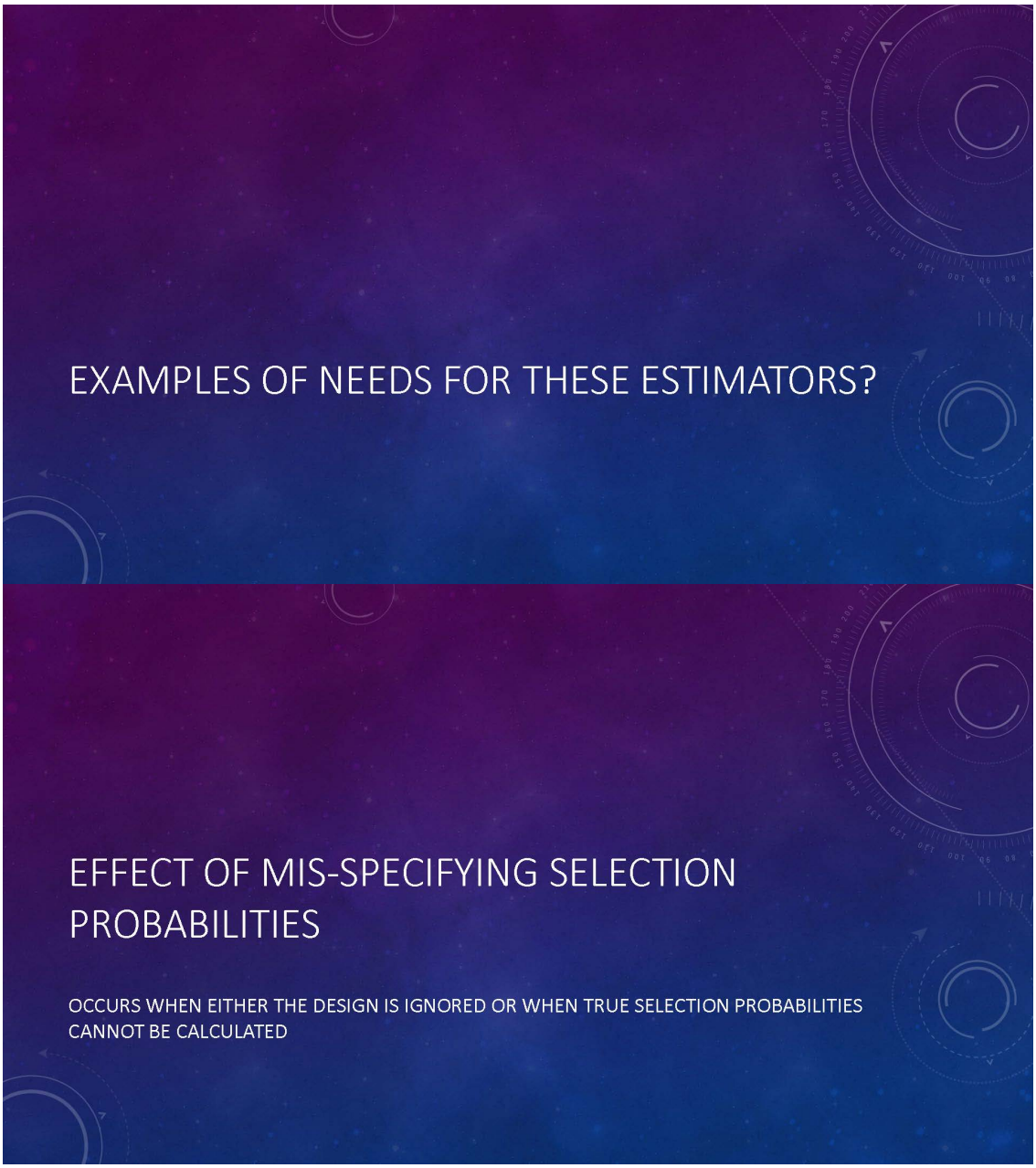
where $Q = \sum_{i=1}^{N} \sum_{j=1}^{M_i} Q_{j(i)}$ is the total number of TSUs in the entire population

$$\hat{\mu}_2 = \frac{\hat{T}}{M} \qquad with\ \widehat{var}(\hat{\mu}_2) = \frac{\widehat{var}(\hat{T})}{M^2}$$

where $M = \sum_{i=1}^{N} M_i$ is the total number of SSUs in the entire population

$$\hat{\mu}_2 = \frac{\hat{T}}{N} \qquad with\ \widehat{var}(\hat{\mu}_2) = \frac{\widehat{var}(\hat{T})}{N^2}$$

where N is the total number of PSUs in the entire population


EXAMPLES OF NEEDS FOR THESE ESTIMATORS?
EFFECT OF MIS-SPECIFYING SELECTION PROBABILITIES
OCCURS WHEN EITHER THE DESIGN IS IGNORED OR WHEN TRUE SELECTION PROBABILITIES CANNOT BE CALCULATED

ASSUMING SRSWR WHEN SAMPLING IS UPSWR

- $\mathbf{Y} = \{1, 2, 3, 4, 50\}$
- UPSprobs $\mathbf{U} = \{0.24, 0.24, 0.24, 0.24, 0.04\}$
- SRSprobs $\mathbf{P} = \{0.20, 0.20, 0.20, 0.20, 0.20\}$
- $\theta = 60$
- $n = 2$
- $\hat{T}_{HH} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i}$
- $E[\hat{T}_{HH}] = \sum_S \Pr(S)\hat{T}_{HH,S}$

Sampling	Mean of $\hat{T}_{HH}$	SE($\hat{T}_{HH}$)	RSE
UPSWR – UPSWR	60	172.1	2.87
UPSWR – SRSWR	22.05	33.19	1.51
SRSWR - SRSWR	60	17.02	0.28

Sample (I)	Pr(I|USPWR)	$\hat{T}_{HH}$|UPSWR	Pr(I|SRSWR)	$\hat{T}_{HH}$|SRSWR
{2, 0, 0, 0, 0}	0.057	4.17	0.04	5.0
{1, 1, 0, 0, 0}	0.116	6.25	0.08	7.5
{1, 0, 1, 0, 0}	0.116	8.33	0.08	10.0
{1, 0, 0, 1, 0}	0.116	10.42	0.08	12.5
{1, 0, 0, 0, 1}	0.019	627.08	0.08	127.5
{0, 2, 0, 0, 0}	0.057	8.33	0.04	10.0
{0, 1, 1, 0, 0}	0.116	10.42	0.08	12.5
{0, 1, 0, 1, 0}	0.116	12.50	0.08	15.0
{0, 1, 0, 0, 1}	0.019	629.17	0.08	130.0
{0, 0, 2, 0, 0}	0.057	12.50	0.04	15.0
{0, 0, 1, 1, 0}	0.116	14.58	0.08	17.5
{0, 0, 1, 0, 1}	0.019	631.25	0.08	132.5
{0, 0, 0, 2, 0}	0.057	16.67	0.04	20.0
{0, 0, 0, 1, 1}	0.019	633.33	0.08	135.0
{0, 0, 0, 0, 2}	0.002	1250.00	0.04	250.0
Totals	1.00	---	1.00	---

ASSUMING SRSWR WHEN SAMPLING IS UPSWR

- $\mathbf{Y} = \{1, 2, 3, 4, 50\}$
- UPSprobs $\mathbf{U} = \{0.10, 0.10, 0.10, 0.10, 0.60\}$
- SRSprobs $\mathbf{P} = \{0.20, 0.20, 0.20, 0.20, 0.20\}$
- $\theta = 60$
- $n = 2$
- $\hat{T}_{HH} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i}$
- $E[\hat{T}_{HH}] = \sum_S \Pr(S)\hat{T}_{HH,S}$

Sampling	Mean of $\hat{T}_{HH}$	SE($\hat{T}_{HH}$)	RSE
UPSWR – UPSWR	60	20.06	0.33
UPSWR – SRSWR	36.67	29.43	0.80
SRSWR - SRSWR	60	17.02	0.28

Sample (I)	Pr(I|USPWR)	$\hat{T}_{HH}$|UPSWR	Pr(I|SRSWR)	$\hat{T}_{HH}$|SRSWR
{2, 0, 0, 0, 0}	0.01	10	0.04	5.0
{1, 1, 0, 0, 0}	0.02	15	0.08	7.5
{1, 0, 1, 0, 0}	0.02	20	0.08	10.0
{1, 0, 0, 1, 0}	0.02	25	0.08	12.5
{1, 0, 0, 0, 1}	0.12	46.7	0.08	127.5
{0, 2, 0, 0, 0}	0.01	20	0.04	10.0
{0, 1, 1, 0, 0}	0.02	25	0.08	12.5
{0, 1, 0, 1, 0}	0.02	30	0.08	15.0
{0, 1, 0, 0, 1}	0.12	51.7	0.08	130.0
{0, 0, 2, 0, 0}	0.01	30	0.04	15.0
{0, 0, 1, 1, 0}	0.02	35	0.08	17.5
{0, 0, 1, 0, 1}	0.12	56.7	0.08	132.5
{0, 0, 0, 2, 0}	0.01	40	0.04	20.0
{0, 0, 0, 1, 1}	0.12	61.7	0.08	135.0
{0, 0, 0, 0, 2}	0.36	83.3	0.04	250.0
Totals	1.00	---	1.00	---

GUESSING PROBS FOR UPSWR WHEN SAMPLING IS UPSWR

- $\mathbf{Y} = \{1, 2, 3, 4, 50\}$
- UPSprobs $\mathbf{U} = \{0.24, 0.24, 0.24, 0.24, 0.04\}$
- GPSprobs $\mathbf{P} = \{0.20, 0.25, 0.15, 0.36, 0.04\}$
- $\theta = 60$
- $n = 2$
- $\hat{T}_{HH} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i}$

Sampling	Mean of $\hat{T}_{HH}$	SE($\hat{T}_{HH}$)	RSE
UPSWR	60	172.1	2.87
GPSWR	60.6	171.8	2.83

Sample (I)	Pr(I|USPWR)	$\hat{T}_{HH}$|UPSWR	Pr(I|GPSWR)	$\hat{T}_{HH}$|GPSWR
{2, 0, 0, 0, 0}	0.057	4.17	0.040	5.00
{1, 1, 0, 0, 0}	0.116	6.25	0.100	6.50
{1, 0, 1, 0, 0}	0.116	8.33	0.060	12.50
{1, 0, 0, 1, 0}	0.116	10.42	0.144	8.06
{1, 0, 0, 0, 1}	0.019	627.08	0.016	627.50
{0, 2, 0, 0, 0}	0.057	8.33	0.063	8.00
{0, 1, 1, 0, 0}	0.116	10.42	0.075	14.00
{0, 1, 0, 1, 0}	0.116	12.50	0.179	9.56
{0, 1, 0, 0, 1}	0.019	629.17	0.020	629.00
{0, 0, 2, 0, 0}	0.057	12.50	0.022	20.00
{0, 0, 1, 1, 0}	0.116	14.58	0.108	15.56
{0, 0, 1, 0, 1}	0.019	631.25	0.012	635.00
{0, 0, 0, 2, 0}	0.057	16.67	0.129	11.11
{0, 0, 0, 1, 1}	0.019	633.33	0.029	630.56
{0, 0, 0, 0, 2}	0.002	1250.00	0.002	1250.00
Totals	1.00	---	1.00	---

EXAMPLES?

CONCLUSIONS?

- Mis-specifying the sampling design causes bias in both the estimate and its variance
  - E.g. sampling using UPS and estimating using SRS
- Using unequal probabilities that are not correct but which mimic the sampling design has little effect on the estimate or its variance
  - What does this say if the estimated probabilities are not close to the ones used in the actual sampling?
    - Now in a situation similar to UPS selection but SRS estimation
  - What does this say about using estimated mimic probabilities in UPS designs?
    - Not much if the actual selection probabilities are based on estimated "sizes", i.e. these are the probabilities in use
    - E.g. $p_i = \frac{weight\ of\ a\ haul}{last\ year's\ total\ landings}$
      - here the only issue is that the sum of the weights of all hauls $\neq$ last year's landings, i.e. $\sum p_i \neq 1$
      - Not a problem since this is requires only a scalar constant; the probabilities are still appropriately of values in relation to each other

Annex 13: Design-Based Multivariate Estimation presentation

By Mary Christman

4) Design-Based Multivariate Estimation 093019.pptx

Notation for **single stage** sampling

3

- $i = 1, \ldots, n$ where n is the number of sampled units
- $j = a, \ldots, A$ where a is the minimum age and A is the maximum age of a fish
- w_{ij} = weight of the i^{th} sampled fish at the j^{th} age
- l_{ij} = length of the i^{th} sampled fish at the j^{th} age
- n_j = number of fish sampled for weight and length at the j^{th} age
- $n = \sum_{j=a}^{A} n_j$ = total number of fish sampled
- N = total number of fish in the population
- p_i = the selection probability for the i^{th} fish
  - assume to be with replacement (WR) since $n \ll N$
  - Use the HH estimators

4

Single stage WR (usually SRSWR)

1-stage SRSWR sampling

5

- Here n is fixed and the $n_j, j = a, \ldots, A$, are random
- Observe lengths $\{l_1, l_2, \ldots, l_n\}$ and ages $\{age_1, age_2, \ldots, age_n\}$ on n fish
- Re-label lengths to associate each with age:
  $\{l_{1a}, l_{2a}, \ldots, l_{n_a,a}, l_{1(a+1)}, l_{2(a+1)}, \ldots, l_{n_{a+1},(a+1)}, \ldots, l_{1A}, l_{2A}, \ldots, l_{n_A,A}\}$
- The estimator of mean length at age is the vector

$$\hat{\boldsymbol{\mu}}_L = \left\{\frac{\sum_{i=1}^{n_j} l_{ij}}{n_j}\right\}_{j=a}^{A} = \begin{bmatrix} \bar{l}_a \\ \bar{l}_{a+1} \\ \cdots \\ \cdots \\ \bar{l}_A \end{bmatrix}$$

Covariance matrix for the vector $\hat{\boldsymbol{\mu}}_L$

6

- $cov(\hat{\boldsymbol{\mu}}_L)$ is a symmetric $A \times A$ matrix with diagonal elements $var(\bar{l}_j)$ and off-diagonal elements $cov(\bar{l}_j, \bar{l}_{j'})$ where $cov(\bar{l}_j, \bar{l}_{j'}) = cov(\bar{l}_{j'}, \bar{l}_j)$

$$cov(\hat{\boldsymbol{\mu}}_L) = \begin{bmatrix} var(\bar{l}_a) & cov(\bar{l}_a, \bar{l}_{a+1}) & \cdots & cov(\bar{l}_a, \bar{l}_A) \\ cov(\bar{l}_{a+1}, \bar{l}_a) & \ddots & & cov(\bar{l}_{a+1}, \bar{l}_A) \\ \vdots & & & \vdots \\ cov(\bar{l}_A, \bar{l}_a) & cov(\bar{l}_A, \bar{l}_{a+1}) & \cdots & var(\bar{l}_A) \end{bmatrix}$$

Covariance matrix for $\hat{\boldsymbol{\mu}}_L$ under SRSWR

- Assume $n_j, j = a, \ldots, A$, are fixed – recommended by Thompson (2002, pg. 124-125)
- $cov(\hat{\boldsymbol{\mu}}_L)$ has
  - Diagonal elements $\widehat{var}(\bar{l}_j|n_j), j = a, \ldots, A$ where under SRSWR

$$\widehat{var}(\bar{l}_j|n_j) = \frac{s_j^2}{n_j} = \frac{1}{n_j(n_j-1)}\sum_{i=1}^{n_j}(l_{ij} - \bar{l}_j)^2, j = a, \ldots, A$$

Covariance matrix for $\hat{\boldsymbol{\mu}}_L$ under SRSWR

- The off diagonal values are the covariances of pairs of means:

$$cov(\bar{l}_j, \bar{l}_{j'}) = \sum_{i=1}^{n_j}\sum_{i'=1}^{n_{j'}}\frac{1}{n_j}\frac{1}{n_{j'}}cov(l_{ij}, l_{i'j'}) \quad , j \neq j'; \quad j, j' = a, \ldots, A$$

where $l_{ij}, l_{i'j'}$ are the lengths of two units labeled ij and $i'j'$ from different age classes (j and j')

Covariance matrix for $\hat{\mu}_L$ under SRSWR 9

- To calculate $cov(l_{ij}, l_{i'j'})$:
  - Let a_i = number of times unit i in age j is included in the sample of size n_j. Then, the $\{a_1, \ldots, a_N\}$ have a multinomial distribution where

$$E[a_i] = \frac{n}{N},\ var[a_i] = \frac{n(N-1)}{N^2}, covar[a_i, a_j] = -\frac{n}{N^2}$$

Covariance matrix for $\hat{\mu}_L$ under SRSWR 10

- Then the covariance can be re-written in terms of the a_i and $a_{i'}$:

$$cov(\bar{l}_j, \bar{l}_{j'}) = \sum_{i=1}^{N_j}\sum_{i'=1}^{N_{j'}} \frac{l_{ij}}{n_j}\frac{l_{i'j'}}{n_{j'}} cov(a_{ij}, a_{i'j'})$$

$$= \sum_{i=1}^{N}\sum_{k\neq i}^{N} I[i,k]\frac{l_i}{n_j}\frac{l_k}{n_{j'}} cov(a_i, a_k)$$

$$= -\sum_{i=1}^{N}\sum_{k\neq i}^{N} I[i,k]\frac{l_i}{n_j}\frac{l_k}{n_{j'}}\frac{n}{N^2}$$

where $I[i,k]$ is the indicator that unit i is in the sample and in age group j and unit k is in the sample and age group j'

NOTE: these can be vanishingly small covariances, so I consider covariance to be 0

11

Single stage WOR (usually SRSWOR)

12

1-stage WOR sampling

- Re-label lengths to associate each with age:

$$\{l_{1a}, l_{2a}, \dots, l_{n_a,a}, l_{1(a+1)}, l_{2(a+1)}, \dots, l_{n_{a+1},(a+1)}, \dots, l_{1A}, l_{2A}, \dots, l_{n_A,A}\}$$

  where l_{ij} is the length of the i^{th} fish in the j^{th} age

- Let π_{ij} be the inclusion probability for the i^{th} fish in j^{th} age class
- The estimator of mean length at age is the vector

$$\hat{\boldsymbol{\mu}}_L = \left\{ \frac{\sum_{i=1}^{n_j} \frac{l_{ij}}{\pi_{ij}}}{\sum_{i=1}^{n_j} \frac{1}{\pi_{ij}}} \right\}_{j=a}^{A} = \begin{bmatrix} \hat{\mu}_{L,a} \\ \hat{\mu}_{L,a+1} \\ \cdots \\ \cdots \\ \hat{\mu}_{L,A} \end{bmatrix}$$

Example: Herring in the Baltic

13

- Actual Sampling Design:
  - SRSWOR of PSU (Vessel×Trip)
  - SRSWR of fish in last fishing operation are sampled for length
  - Stratified SRSWOR of fish are subsampled for weight and age
    - Strata are length classes
    - Stratum sample size = min(observed number of fish in length class, 5)
- For this example, let the fish sampled in the last haul of the Vessel×Trip be the population, therefore this is a 1-stage sample
  - Estimate the mean length at age for these fish

Example: Herring in the Baltic

- Stratified random sampling WOR
- If the number of fish in a length stratum in the "population" < 5, then

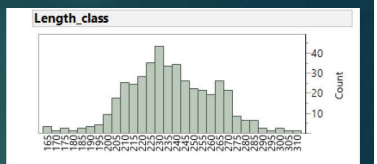
$$\pi_s = 1$$

i.e. the inclusion probability for an observation in length class stratum s

- If the number of fish in a stratum ≥ 5, then

$$\pi_s = \frac{5}{n_s}$$

where n_s is the number of fish of length class stratum s in the population



Frequencies

Level	Count	Prob
165	3	0.00704
170	1	0.00235
175	2	0.00469
180	1	0.00235
185	2	0.00469
190	3	0.00704
195	4	0.00939
200	9	0.02113
205	17	0.03991
210	25	0.05869
215	24	0.05634
220	28	0.06573
225	35	0.08216
230	43	0.10094
235	33	0.07746
240	34	0.07981
245	26	0.06103
250	22	0.05164
255	21	0.04930
260	19	0.04460
265	26	0.06103
270	21	0.04930
275	8	0.01878
280	6	0.01408
285	6	0.01408
290	2	0.00469
295	1	0.00235
300	2	0.00469
305	1	0.00235
310	1	0.00235
Total	426	1.00000
N Missing	0	
30 Levels		

Population

Example: Herring in the Baltic

Length (mm)	Age Class 1	2	3	4	5	6	7	8	9	10
165	l_{11}, l_{21}			l_{14}						
170	l_{31}									
175		l_{12}		l_{24}						
...										
310							l_{17}			
Mean Length at Age	$\frac{\sum_{i=1}^{2} \frac{l_{i1}}{\pi_{i1}}}{\sum_{i=1}^{2} \frac{1}{\pi_{i1}}}$	$\frac{\sum_{i=1}^{5} \frac{l_{i2}}{\pi_{i2}}}{\sum_{i=1}^{5} \frac{1}{\pi_{i2}}}$	$\frac{\sum_{i=1}^{13} \frac{l_{i2}}{\pi_{i2}}}{\sum_{i=1}^{13} \frac{1}{\pi_{i2}}}$			...		...		$\frac{\sum_{i=1}^{2} \frac{l_{i10}}{\pi_{i10}}}{\sum_{i=1}^{2} \frac{1}{\pi_{i10}}}$

$$\hat{\mu}_L = \left\{ \frac{\sum_{i=1}^{n_j} \frac{l_{ij}}{\pi_{ij}}}{\sum_{i=1}^{n_j} \frac{1}{\pi_{ij}}} \right\}_{j=a}^{A} = \begin{bmatrix} \hat{\mu}_{L,a} \\ \hat{\mu}_{L,a+1} \\ \cdots \\ \cdots \\ \hat{\mu}_{L,A} \end{bmatrix}$$

Length (mm)	Age Class 1	2	3	4	5	6	7	8	9	10	Total
165	2			1							3
170	1										1
175		1		1							2
180		1									1
185		1		1							2
190		1	2								3
195			4								4
200		1	3	1							5
205			3	2							5
210			3	2							5
215			5								5
220			4		1						5
225			1	4							5
230			2	2	1						5
235			3	2							5
240			1	2	2						5
245				2	2	1					5
250			1	1	3						5
255					4	1					5
260			1			4					5
265				2	1	2					5
270					1	3	1				5
275					4	1					5
280					1	4					5
285						1	2	2			5
290								2			2
295							1				1
300									1	1	2
305										1	1
310							1				1
Total	3	5	33	23	20	17	5	4	1	2	113

Example: Herring in the Baltic

	Age Class 1	2	3	4	5	6	7	8	9	10
Mean Length at Age	$\frac{\sum_{i=1}^{2} \frac{l_{i1}}{\pi_{i1}}}{\sum_{i=1}^{2} \frac{1}{\pi_{i1}}}$	$\frac{\sum_{i=1}^{5} \frac{l_{i2}}{\pi_{i2}}}{\sum_{i=1}^{5} \frac{1}{\pi_{i2}}}$	$\frac{\sum_{i=1}^{13} \frac{l_{i2}}{\pi_{i2}}}{\sum_{i=1}^{13} \frac{1}{\pi_{i2}}}$			...		...		$\frac{\sum_{i=1}^{2} \frac{l_{i10}}{\pi_{i10}}}{\sum_{i=1}^{2} \frac{1}{\pi_{i10}}}$
Numerator	335	1090	13580	13290	11575	6963	2081	632	300	605
Denominator	3	5.8	59.8	57	47	26.6	7.4	2.2	1	2
Mean	111.7	187.9	227.1	233.2	246.3	261.8	281.2	287.3	300.0	302.5
Sample Mean	167.5	186.0	221.2	214.3	253.0	266.9	290.0	287.5	300.0	302.5

Length (mm)	Age Class 1	2	3	4	5	6	7	8	9	10	Total	π_s
165	2			1							3	1
170	1										1	1
175		1		1							2	1
180		1									1	1
185		1		1							2	1
190		1	2								3	1
195			4								4	1
200		1	3	1							5	0.556
205			3	2							5	0.294
210			3	2							5	0.200
215			5								5	0.208
220			4		1						5	0.179
225			1	4							5	0.143
230			2	2	1						5	0.116
235			3	2							5	0.152
240			1	2	2						5	0.147
245				2	2	1					5	0.192
250			1	1	3						5	0.227
255					4	1					5	0.238
260			1			4					5	0.263
265				2	1	2					5	0.192
270					1	3	1				5	0.238
275					4	1					5	0.625
280					1	4					5	0.833
285						1	2	2			5	0.833
290								2			2	1
295							1				1	1
300									1	1	2	1
305										1	1	1
310							1				1	1
Total	3	5	33	23	20	17	5	4	1	2	113	

Covariance Matrix for $\hat{\boldsymbol{\mu}}_L$ under WOR 17

- Assume $n_j, j = a, \ldots, A$, are fixed
- The vector $\hat{\boldsymbol{\mu}}_L$ has an estimated variance-covariance matrix where
  - The diagonal elements, $\widehat{var}(\hat{\mu}_{L,j}|n_j)$, $j = a, \ldots, A$, are

$$= \frac{1}{\left(\sum_{i=1}^{n_j} \frac{1}{\pi_{ij}}\right)^2}\left[\sum_{i=1}^{n_j}\left(\frac{1-\pi_{ij}}{\pi_{ij}^2}\right)(y_{ij}-\hat{\mu}_{L,j})^2 + \sum_{i=1}^{n_j}\sum_{i'\neq i}\left(\frac{\pi_{ij,i'j}-\pi_{ij}\pi_{i'j}}{\pi_{ij}\pi_{i'j}}\right)\frac{(y_{ij}-\hat{\mu}_{L,j})(y_{i'j}-\hat{\mu}_{L,j})}{\pi_{ij,i'j}}\right]$$

  - The off-diagonal elements are

$$cov(\hat{\mu}_{L,j}, \hat{\mu}_{L,j'}) = \sum_{i=1}^{n_j}\sum_{i'=1}^{n_{j'}} \frac{l_{ij}}{\pi_{ij}}\frac{l_{i'j'}}{\pi_{i'j'}}(\pi_{ij,i'j'} - \pi_{ij}\pi_{i'j'})$$

Why Care About Covariances? 18

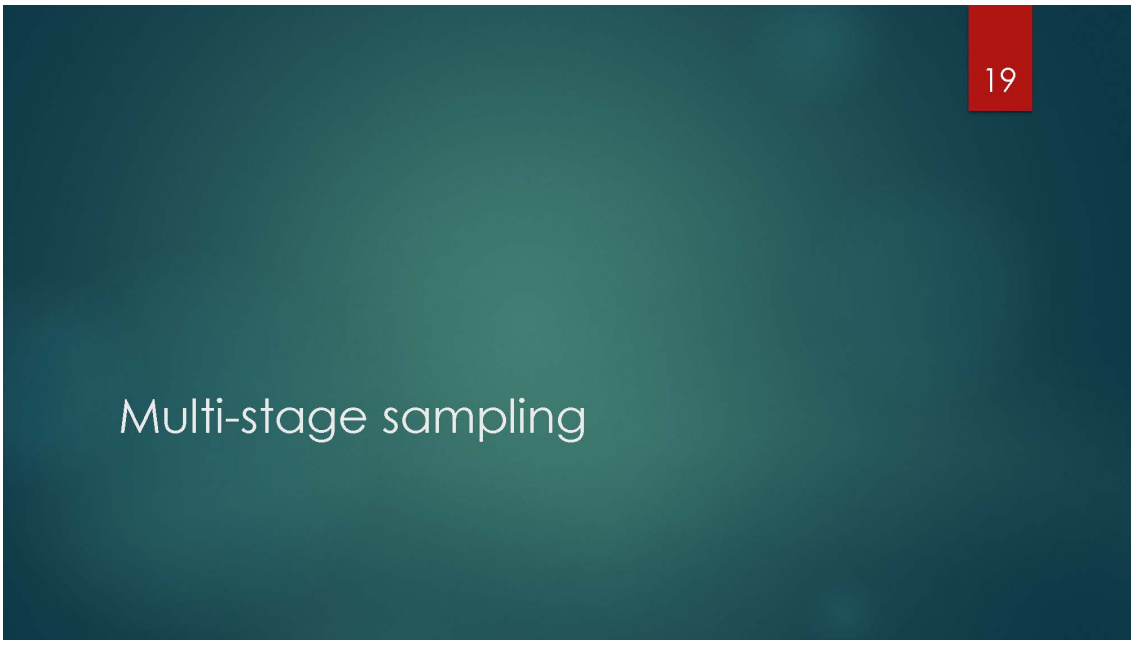
- Suppose you wish to use the mean length at age to estimate the grand mean length of landed fish
- The estimator would be a weighted mean:

$$\hat{\mu}_L = \sum_{j=a}^{A} w_j \hat{\mu}_{L,j}$$

  where w_j is the proportion of the landings in age j

- The variance of $\hat{\mu}_L$ is

$$var(\hat{\mu}_L) = \sum_{j=a}^{A} w_j^2 var(\hat{\mu}_{L,j}) + \sum_{j=a}^{A}\sum_{j'=a}^{A} w_j w_{j'} cov(\hat{\mu}_{L,j}, \hat{\mu}_{L,j'})$$


19
Multi-stage sampling

3-stage sampling for $\hat{\boldsymbol{\mu}}_L$, mean length at age

20

- We can write each mean length at age j, $\hat{\mu}_{L,j}, j = a, \ldots, A$ in the vector $\hat{\boldsymbol{\mu}}_L$ as the ratio

$$\hat{\mu}_{L,j} = \frac{\hat{T}_{L,j}}{\hat{N}_{L,j}}$$

where

- $\hat{T}_{L,j}$ is an estimate of the population total for the j^{th} age and
- $\hat{N}_{L,j}$ is an estimate of the population size for the j^{th} age
- Both estimators would be either HH or HT estimators depending on the sampling design

3-stage sampling

21

- Hence
  - The numerator could be either $\hat{T}_{HH,j}$ or $\hat{T}_{HT,j}$ for the elements in the population of age j depending on the sampling design
  - The denominator would be the equivalent estimator of the total number of elements in the population of age j
  - To obtain $\hat{N}_{HH,j}$ or $\hat{N}_{HT,j}$
    - Replace the $\sum_{k=1}^{u_{j^*i,j}} l_{k(j^*,i),j}$ (the sum of the lengths for fish at age j in the j^{*th} SSU in the i^{th} PSU in the $\hat{T}_{HH,j}$ or $\hat{T}_{HT,j}$ equations) with $u_{j^*i,j}$, the number of fish at age j in the j^{*th} SSU in the i^{th} PSU obtain $\hat{N}_{HH,j}$ or $\hat{N}_{HT,j}$ (see next slide)

Recall the 3-stage estimators

22

$$\hat{T}_{HH} = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{T}_{i,HH}}{p_i} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{p_i}\frac{1}{m_i}\sum_{j=1}^{m_i}\frac{\hat{T}_{j(i)}}{p_{j(i)}}$$
$$= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{p_i}\frac{1}{m_i}\sum_{j=1}^{m_i}\frac{1}{p_{j(i)}}\frac{1}{u_{j(i)}}\sum_{k=1}^{u_{j(i)}}\frac{y_{k(j,i)}}{p_{k(j,i)}}$$

$$\hat{T}_{HT} = \sum_{i=1}^{n}\frac{\hat{T}_i}{\pi_i} = \sum_{i=1}^{n}\frac{1}{\pi_i}\frac{M_i}{m_i}\sum_{j=1}^{m_i}\hat{T}_{j(i)} = \sum_{i=1}^{n}\frac{1}{\pi_i}\frac{M_i}{m_i}\sum_{j=1}^{m_i}\frac{U_{j(i)}}{u_{j(i)}}\sum_{k=1}^{u_{j(i)}}y_{k(i,j)}$$

Variance of the 3-stage estimator

23

- Here we have ratios where both the numerator and denominator are random
- Need to discuss briefly ratio estimators

24

Aside: ratio estimators

25

Ratio estimators: single stage design

- For SRSWR the ratio estimator of $R = \frac{T_y}{T_x}$ is:

$$r = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}$$

where y is the survey variable, $T_y = \sum_{i=1}^{N} y_i$, x is the auxiliary variable, and $T_x = \sum_{i=1}^{N} x_i$.

- Ratio estimators tend to be biased in small samples
- Here our auxiliary information is sample size not other information like effort
- The estimated variance/MSE of r when sampling is SRSWR is approximately

$$\widehat{var}(r) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (y_i - r x_i)^2$$

- If sampling is SRSWOR, multiply $\widehat{var}(r)$ with $\frac{(N-n)}{N}$

26

Variance of 3-stage design under WR sampling

27

Variance of 3-stage WR sampling using HH estimators

- The variance estimator for r implies that if the mean length at age estimators are the ratio of HH estimators, an estimated variance for mean length at age j is approximately

$$\widehat{var}(\hat{\mu}_j) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{\hat{T}_{ij}}{p_i} - \hat{\mu}_j \frac{\hat{N}_{ij}}{p_i} \right)^2$$

where $\hat{T}_{ij}$ is the estimated total length for the j^{th} age class in the i^{th} PSU and $\hat{N}_{ij}$ is the estimated number of fish in the j^{th} age class in the i^{th} PSU

- Note that
  - $\frac{\hat{T}_{ij}}{p_i}$ is an estimate of the population total length for the j^{th} age class
  - $\frac{\hat{N}_{ij}}{p_i}$ is an estimate of the total number of fish at age j in the population

28

Example

SHOWS SOME OF THE ISSUES WITH ACTUAL IMPLEMENTATION

29

Example: Norwegian Spring Spawning Herring For 2016

- 71 PSUs (fishing operations) sampled using PPSWOR where the "size" metric is the estimated size of the haul

$$p_i = \frac{estimated\ fish\ weight\ of\ the\ operation}{estimate\ of\ the\ total\ landings\ in\ weight\ for\ the\ year}$$

- For each PSU, fish (SSUs) were obtained by randomly selecting approximately 20 kg of fish
  - Number of fish sampled for biological purposes was random but averaged around 40 to 50 fish/PSU
  - All fish were measured for length and weight and subset were measured for age

Example: NSSH

30

- If the actual sampling had been a random selection of individual fish rather than a single "bucket" of fish in each PSU, then sampling is 2-stage and the estimator of mean length at age would be

$$\hat{\boldsymbol{\mu}}_L = \left\{ \frac{\sum_{i=1}^{n} \frac{1}{p_i} (M_{ij} \bar{l}_{ij})}{\sum_{i=1}^{n} \frac{1}{p_i} M_{ij}} \right\}_{j=a}^{A}$$

where M_{ij} = number of fish in the i^{th} PSU of age j and $\bar{l}_{ij}$ is the mean length of fish (SSUs) sampled in the i^{th} PSU of age j.

- The numerator $\frac{1}{n}\sum_{i=1}^{n} \frac{1}{p_i} (M_{ij} \bar{l}_{ij})$ is the HH estimator of the population total length at age j
- The denominator $\frac{1}{n}\sum_{i=1}^{n} \frac{1}{p_i} M_{ij}$ is the HH estimator of $M_j = \sum_{i=1}^{N} M_{ij}$, the number of fish in the population at age j.

Example: NSSH

31

$$\hat{\boldsymbol{\mu}}_L = \left\{ \frac{\sum_{i=1}^{n} \frac{1}{p_i} (M_{ij} \bar{l}_{ij})}{\sum_{i=1}^{n} \frac{1}{p_i} M_{ij}} \right\}_{j=a}^{A}$$

- PROBLEM: Do we know M_{ij}? Could we estimate M_{ij}?
  - Requires:
  - $\hat{\mu}_W$ be the average weight of a fish calculated from the entire sample using a DB estimator or external data to estimate the mean weight at age for the species $\hat{\boldsymbol{\mu}}_W$
  - W_i be the observed weight of all fish in the i^{th} PSU
  - $\hat{\boldsymbol{\mu}}_{A,i} = \{\hat{\mu}_{A,ij}\}_{j=a}^{A}$ = vector of proportions at age in the i^{th} PSU
  - So, an estimate of the number of fish at age in the i^{th} PSU would be

$$\bar{\mathbf{M}}_i = \frac{1}{\hat{\mu}_W} W_i \hat{\boldsymbol{\mu}}_{A,i} = \frac{1}{\hat{\mu}_W} \{W_i \hat{\mu}_{A,ij}\}_{j=a}^{A}$$

Example: NSSH

32

- Because the actual sampling is a single observation of a "bucket" of fish from the fishing operation was randomly selected, the 2 stage sampling design is treated as a single stage design for analyses
  - i.e., I am treating each "bucket" as a PSU composed solely of the fish sampled for length, weight and age
- Sampling was WOR but we are treating it as WR for analyses
  - Effect: slight increase in variance estimates

Example: NSSH

33

- The estimator of mean length at age is based on the 1-stage HH estimator

$$\hat{\mu}_L = \left\{ \frac{\sum_{i=1}^{n} \frac{l_{ij}}{p_i}}{\sum_{i=1}^{n} \frac{n_{ij}}{p_i}} \right\}_{j=a}^{A}$$

where

- $i = 1, \ldots, n$ where n is the number of sampled PSUs
- $j = a, \ldots, A$ where a is the minimum age and A is the maximum age of a fish
- l_{ij} = sum of the lengths for fish sampled in the i^{th} PSU at the j^{th} age
- n_{ij} = number of fish sampled for weight and length in the i^{th} PSU at the j^{th} age
- p_i = the selection probability for the i^{th} PSU assumed to be with replacement (WR)

Example: NSSH

34

- The variance of $\hat{\boldsymbol{\mu}}_L = \left\{ \frac{\sum_{i=1}^{n} \frac{l_{ij}}{p_i}}{\sum_{i=1}^{n} \frac{n_{ij}}{p_i}} \right\}_{j=a}^{A}$ is given by
  - Diagonal elements are the variances of the elements of $\hat{\boldsymbol{\mu}}_L$

$$\widehat{var}\left(\hat{\mu}_{L,j}\right) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{\hat{T}_{ij}}{p_i} - \hat{\mu}_j \frac{\hat{N}_{ij}}{p_i} \right)^2$$

  - Off-diagonal elements are the covariances $cov\left(\hat{\mu}_{L,j}, \hat{\mu}_{L,j'}\right) = 0$ since sampling is WR assuming a multinomial model

Some Results and Comparisons

N=71 PSUs Sampled Under PPSWR and Estimated Assuming PPSWR

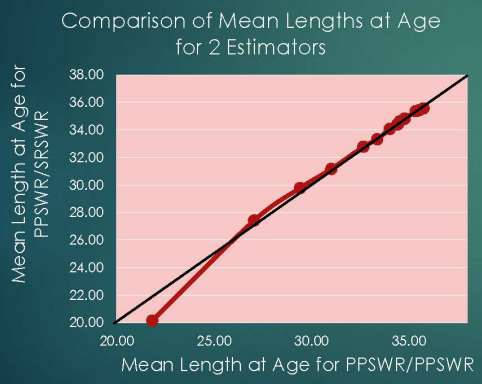
Age	Mean Length	SEM	CI Endpoints	
			5%	95%
1	na	na	na	na
2	21.83	2.49	19.75	25.00
3	27.06	1.13	24.97	28.18
4	29.42	0.49	28.51	30.18
5	31.02	0.24	30.60	31.41
6	32.67	0.14	32.40	32.94
7	33.38	0.12	33.15	33.55
8	34.02	0.12	33.83	34.21
9	34.41	0.13	34.17	34.61
10	34.45	0.08	34.33	34.58
11	34.55	0.15	34.31	34.82
12	34.80	0.08	34.67	34.92
13	35.36	0.22	34.99	35.70
14	35.53	0.22	35.20	35.91
15+	35.76	0.46	34.95	36.47

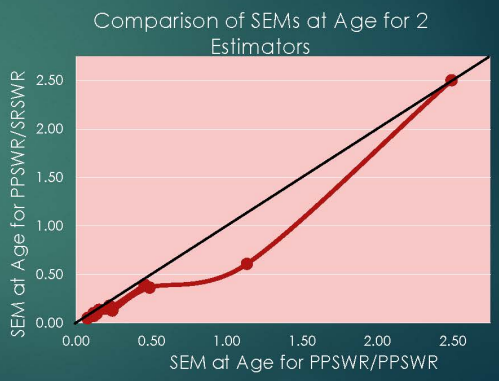
N=71 PSUs Sampled Under PPSWR and Estimated Assuming SRSWR

Age	Mean Length	SEM	CI Endpoints	
			5%	95%
1	na	na	na	na
2	20.15	2.50	19.75	25.00
3	27.42	0.61	26.32	28.14
4	29.78	0.37	29.12	30.33
5	31.17	0.13	30.96	31.38
6	32.78	0.11	32.60	32.96
7	33.33	0.07	33.21	33.45
8	34.07	0.10	33.90	34.23
9	34.39	0.09	34.23	34.54
10	34.42	0.05	34.34	34.50
11	34.64	0.13	34.42	34.87
12	34.81	0.05	34.73	34.90
13	35.36	0.18	35.00	35.59
14	35.43	0.15	35.20	35.68
15+	35.57	0.39	35.00	36.25

Comparison

36





Example: Herring in the Baltic

37

- Actual Sampling Design:
  - SRSWOR of PSU (Vessel×Trip)
  - SRSWR of fish in last fishing operation are sampled for length
  - Stratified SRSWOR of fish are subsampled for weight and age
    - Strata are length classes
    - Stratum sample size = min(observed number of fish in length class, 5)
- n = 62 fishing trips
- Estimate the Numbers at Age (NAA)

Example: Herring in the Baltic

38

- Estimation Approach
  - For each fishing trip (stage 1), a sample of fish (stage 2) were selected for length measurements
  - A stratified random subsample of these fish were further sampled for weight and age
- Step 1: for each trip a length-weight relationship was derived using

$$\ln(Weight_i) = \beta_0 + \beta_1 \ln(Length_i) + \varepsilon_i$$

where i is the identifier for the i^{th} fish measured for weight, , $i = 1, \ldots, N_w$ (N_w = number of fish weighed for the trip), and the error terms ε_i were assumed to be independently and identically distributed as normal with mean 0 and variance σ^2.

Example: Herring in the Baltic

39

- Step 2: the fitted regression ("trip-level LWR") was used to calculate a predicted weight (in kg) for each fish measured for length in trip t using

$$\widehat{Weight}_j = \exp\left(\hat{\beta}_0 + \hat{\beta}_1 \ln\left(Length_j\right)\right)/1000$$

where $j = 1, \ldots, N_t$, N_t = number of fish measured for length on the trip, and the ^ over a quantity indicates that it is the estimated value of that quantity based on the regression analysis.

- Step 3: Estimate the numbers at length for the full trip weight based on the proportions of observed fish at length within the sample. The estimated numbers at length for the trip was calculated for each of the observed length classes in the sample as

Example: Herring in the Baltic

40

- Step 3: Estimate the numbers at length for the full trip weight based on the proportions of observed fish at length within the sample as

$$\widehat{NL}_l = \left\{\frac{f_l \times \widehat{Weight}_l}{\widehat{SW}} \times TW\right\}_{l=1}^{LC}$$

where $\widehat{NL}_l$ is the predicted number of fish in l^{th} length class, $l = 1, \ldots, LC$, for the trip,

f_l is the number of fish sampled in length class l in the sample,

$\widehat{Weight}_l$ is the predicted weight for length class l in the sample,

$\widehat{SW}$ is the predicted sample weight,

TW is the total trip weight, and

LC is the number of distinct length classes observed on the trip.

Example: Herring in the Baltic

41

- Step 4: The predicted matrix of numbers at age by length class for the t^{th} trip was obtained by row-wise multiplying the elements of the length vector $\mathbf{NL}_t$ with the associated rows of the ALK matrix to obtain, for each length class, the number of fish in each age class

- Step 5: The numbers at age for trip t, $\mathbf{NA}_t$, were calculated as the sum of the numbers at age in each length class. The vector of the predicted numbers at age for trip t, $\mathbf{NA}_t = \{\widehat{NA}_{0,t}, \widehat{NA}_{1,t}, \ldots, \widehat{NA}_{13,t}\}$ where 13 is the largest age observed in the original 62 trips were summed over all trips to get an estimate of the total numbers at age over the 62 trips,

$$\mathbf{NAA}_{obs} = \sum_{t}^{T=62} \mathbf{NA}_t$$

- where $\mathbf{NAA}_{obs}$ is the vector of estimated total numbers at age for the observed dataset of 62 trips.

Example: Herring in the Baltic

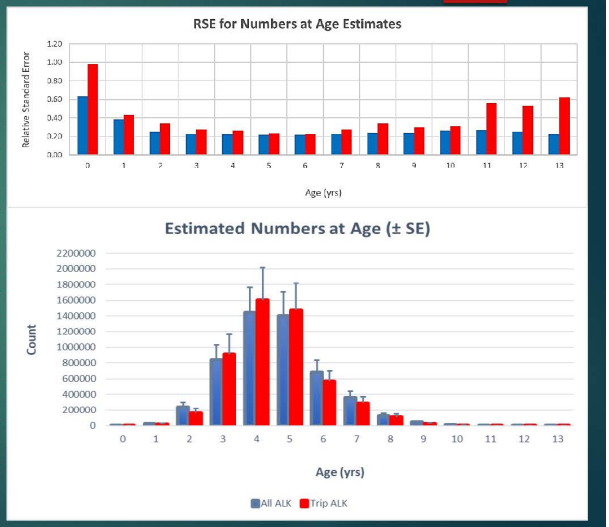
42

- To obtain estimates of the variances, standard errors (SEs), and relative standard errors (RSE = SE/mean) for each of the parameters, $B = 500$ bootstraps were performed. For each bootstrap, 62 trips were randomly selected with replacement from the pseudo-population of the original 62 trips. For each randomly selected trip, N_t fish lengths were randomly selected with replacement where N_t is the original number of fish sampled for length.

Example: Herring in the Baltic

43

Trip ALK and Trip LWR				
Age	Original Estimate	Bootstrap Mean	Bootstrap Std Err	RSE
0	2546	2099	2495.12	0.98
1	21420	24170	9201.95	0.43
2	164111	181320	55763.96	0.34
3	913234	992078	253074.39	0.28
4	1600759	1823373	418282.12	0.26
5	1472385	1554701	345039.59	0.23
6	571543	613846	128669.82	0.23
7	289272	289804	80553.69	0.28
8	111248	113435	37932.46	0.34
9	29307	33004	8886.26	0.30
10	8896	8980	2742.88	0.31
11	1277	1385	719.01	0.56
12	607	660	323.12	0.53
13	273	218	169.45	0.62
TOTAL	5186878	5639073		
MEAN/TRIP	83659	90953	18675	0.205



Example: Herring in the Baltic

44

Variance-covariance matrix for the numbers at age estimators (/10^6)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	6	11	80	5	-70	-29	-33	-12	-9	1	-1	0	0	0
1	11	75	293	397	1221	751	266	26	-14	19	0	0	0	0
2	80	293	2907	4059	6447	3004	1253	651	-89	75	-6	3	0	-1
3	5	397	4059	63852	56786	68027	26174	16935	4110	392	141	4	2	-4
4	-70	1221	6447	56786	179386	84863	42388	19787	6268	1669	616	52	11	-6
5	-29	751	3004	68027	84863	121000	38624	22555	9072	1359	427	28	2	-7
6	-33	266	1253	26174	42388	38624	18744	9922	3009	550	157	13	-1	-1
7	-12	26	651	16935	19787	22555	9922	8079	2044	227	87	1	0	-1
8	-9	-14	-89	4110	6268	9072	3009	2044	1698	138	42	2	-1	0
9	1	19	75	392	1669	1359	550	227	138	106	7	5	0	0
10	-1	0	-6	141	616	427	157	87	42	7	11	0	0	0
11	0	0	3	4	52	28	13	1	2	5	0	1	0	0
12	0	0	0	2	11	2	-1	0	-1	0	0	0	0	0
13	0	0	-1	-4	-6	-7	-1	-1	0	0	0	0	0	0