Strengthening Regional cooperation in the area of fisheries data collection – MARE/2016/22

Socio-economic data collection for fisheries, aquaculture and the processing industry

Work Package 2: Harmonization of methodologies for sampling design and estimation methods for fleet and aquaculture economic data collection

D.2.1: Handbook on sampling design and estimation methods for economic data collection in fisheries statistics

Partners involved:

LUKE,

NISEA

May 2019

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

FINAL

May 2019

Preface

Handbook on sampling design and estimation methods for economic data collection in fisheries statistics was produced under the EU funded project SECFISH: Socio-economic data collection for fisheries, aquaculture and the processing industry (EU Call for Proposals Mare 2016/22: Strengthening regional cooperation in the area of fisheries data collection). The Work Package 2 is aimed at harmonizing the methodologies of sampling design and estimation methods by providing a practical manual based on the general theory of probability sampling. The handbook can be used by the Member States as supporting guidelines in economic data production.

The handbook explains the general principles of probability sampling and essential requirements for a good quality survey plan, and covers the basic sampling techniques. Description of each design will be accompanied by the explanation of appropriate methods of estimation, as well as, uncertainty assessment leading to a well-based coefficient of variation.

The handbook has been produced by a team of contributors including Juha Heikkinen, Jarmo Mikkola, Heidi Pokki and Jarno Virtanen of Natural Resources Institute Finland, Evelina Sabatella of NISEA, and Risto Lehtonen of University of Helsinki (main author).

Contents

1 Introduction

Summary. The handbook focuses on practical sampling and estimation methods for fixed and finite populations of identifiable units (elements) under the conventional design-based framework. This framework is common in survey statistics in general and also in fisheries statistics. We discuss and demonstrate tools for the planning and implementation of sampling designs and estimation designs that would be statistically valid for proper inference and technically manageable for a given fisheries survey.

Auxiliary information on the population plays a crucial role. By introducing suitable auxiliary information in the sampling and estimation procedures, statistical efficiency and cost efficiency can be managed and improved in a controlled way. Data on auxiliary variables are assumed available, either at the unit level in the sampling frame (for certain sampling designs) or as aggregates taken from reliable sources, such as official statistics (for certain estimation designs). Statistical models are used as assisting tools when appropriate.

Traditional methods for element sampling are discussed: simple random sampling, systematic sampling, PPS sampling and stratified sampling. Cluster sampling and multi-stage designs are not treated in detail. For estimation of population parameters we discuss traditional methods including Horvitz-Thompson or expansion estimators and model-free calibration techniques as well as commonly used model-assisted methods, such as ratio and regression estimation and post-stratification. For the treatment of missing data we discuss imputation methods for item nonresponse and reweighting methods for unit nonresponse. Two case studies are presented, one for Italy and the other for Finland. The case studies represent different but manageable approaches for sampling and estimation in fisheries statistics.

The methods are illustrated with extensive worked examples under a realistic synthetic population by using sampling and estimation procedures for samples of different sizes. The results are evaluated with small simulation experiments. General methodological conclusions are provided as well as brief guidelines for practical application.

Computation tools (SAS and R) are briefly summarized in the annexes. SAS and R codes and data sets for worked examples in Chapters 3, 4 and 5 are made available together with the handbook on data collection website.

2 General concepts

2.1. Census versus sample survey

The general term *survey* refers to a query that is used to collect data for making inferences on a population of interest. The data collection method can vary. Sometimes the survey might be directed to entire population (*census survey*), but in most cases the data are collected from a share of the population (*sample survey*). Often the population information is obtained from a register (*register-based survey)*. Survey might also be executed in the internet (*web survey)*. In survey practice, combinations of different approaches are often used. For example, in a combined sample survey and register survey, a part of data are collected with a sample survey and additional information is taken from registers. This option is becoming increasingly common in fisheries statistics.

In most cases, sufficient resources are not available for a complete census. A carefully designed sample study would provide results that are accurate enough for practical purposes. In order to be able to generalize the sample results to the population, the sampling and estimation procedures should be based on well-established statistical methodologies. Survey sampling methods provide tools for cost-efficient and manageable ways to execute the sampling and estimation procedures. This handbook presents several approaches and methods as well as practical applications for surveys in fisheries statistics.

2.2 Target population and sampling frame

Target population contains all the units or elements that we are interested in, whereas the *sampling population* includes only those units that could actually be drawn into a sample. It often happens that all of the units of the target population cannot be reached, for example due to missing contact information. *Sampling frame*, therefore, contains only those units of the sampling population that can possibly be drawn into a sample. The sampling frame is said to have *over-coverage* when it contains units that do not belong to the target population. The opposite case, *under-coverage,* is probably more common, referring to the case where the frame does not contain all intended target population units. Both quality deficiencies of the sampling frame can cause biased results and therefore require careful examination and cleaning if necessary, before the implementation of the sampling operations.

The sampling frame consists of *identifiable* units that are attached with unique *labels*, for example the identification code of a registered fishing vessel or the PIN of a person. ID codes allow population units to be sampled and contacted for data collection. By using identification codes, information can be extracted from registers and other sources and merged with records of the sampling frame, to be used in sampling and estimation procedures.

Important additional variables of the sampling frame are *technical variables* Z that are related to the sampling method of the survey. These include variables determining the probability of population unit to be sampled and stratum and cluster membership identifiers, and variables carrying information on sizes of population elements for unequal probability sampling methods.

In fisheries statistics, a sample of fishing vessels is often drawn directly from the sampling frame that covers the intended population of vessels. Formally, a frame population is denoted $U = \{1, \dots, k, \dots, N\}$, it has N identifiable elements. In this handbook, the frame population SIMPOP consists of $N = 120$ vessels. SIMPOP is an artificially generated population but realistic enough for illustrating the methods of the handbook.

The population of vessels may be readily grouped into naturally existing sets called *clusters*. For example, a fisheries enterprise can manage several fishing vessels. A possible sampling scenario is to draw first a sample of enterprises or clusters from a sampling frame of enterprises. Then, all eligible vessels of the sample enterprises may constitute the element sample, leading to one-stage cluster sampling. For sampling of vessels, an element-level sampling frame is not needed. The element frame is needed if samples of vessels are to be drawn from the sampled enterprises, leading to two-stage cluster sampling. Another possible scenario is to first draw a sample of harbors from the population of target harbors and take all eligible vessels from the sample harbors in the element sample. Sampling frame for drawing a sample of harbors is needed.

Cluster sampling typically weakens statistical efficiency relative to element-level sampling, because clusters tend to be internally homogeneous with respect to the phenomenon of interest. Sampling from element-level frames is thus advisable. This approach has been adopted in the handbook. In some situations, cluster sampling may be justified for cost-efficiency reasons.

2.3 Survey variables and population parameters

The information of primary interest attached to the units of target population is denoted with the values of *study* or *target variable* Y i.e. $\{y_1, \dots, y_k, \dots, y_N\}$. The values y_k are the unknown values of the target variable. The survey is carried out to obtain measurements for Y, or several target variables, for elements $k \in s$ of the sample s that has been drawn from the frame population. Assuming error-free measurement, we denote by y_k the sample values of Y.

In addition to the target variable Y and the technical variables Z, the survey may include information on *auxiliary variables* or *covariates* X. Auxiliary information refers to the information on the population that is not of primary interest in the survey but can be useful for efficient sampling and estimation procedures. In general, for the auxiliary variables to be useful their values should be available for the sampled units. Some methods require population or subpopulation totals of the auxiliary variables, while other methods require their values for all units in the population. In the latter case, it is advisable to include the auxiliary variables in the frame population.

The aim of survey is to estimate the unknown values of *population parameters,* which in general are functions of the population values y_k of the target variable. *Estimators* of population parameters are functions of the sample values of Y, the technical variables Z and the possible auxiliary variables X. Various types of estimators, i.e. computational algorithms, for the estimation of the population parameters of interest are discussed in the handbook.

In the handbook, we mainly consider estimation of *population total*, the sum of the values of the target variable over all units of the population, given by:

$$t = \sum_{i=1}^{N} y_k. \qquad (1)$$

Consider, for example, the population of all registered fishing vessels in a country, and let y_k be the value of landings of vessel k over a year, say. Then the parameter t is the total value of the landings in the country during the year. Estimation of totals over subpopulations (*domains*, e.g., certain type of vessels or fishing) will also be discussed.

Population totals are often more meaningful than *population means* $\bar{y} = t/N$. For example, the mean value of landings per vessel depends heavily on the distribution of the vessel size and fishing effort. In comparisons between countries, it might then be more relevant to compare the total value of landings divided by the total costs, rather than divided by the number of vessels.

2.4. Probability sampling and inference

In probability sampling, a.k.a. random sampling, each unit in the population has a known positive *inclusion probability* (probability to be selected into a sample) π_k. The probabilistic nature of random samples guarantees valid statistical inference i.e. the generalization of the results to the target population by computing standard errors and confidence intervals for the estimators. Random sampling must be separated from *non-probabilistic* methods such as quota sampling, where there is no basis for proper statistical inference. We discuss in the handbook exclusively methods for probability sampling.

The collection of the rules and techniques used in the selection of a sample is referred to as a *sampling scheme*. Under a sampling scheme the probability of selection $p(s)$ can be attached to each sample $s \subset U$ i.e. subset of the population. The function $p(\cdot)$ is formally called the *sampling design*.

In the classical *randomization* or *design-based* inference, the values of the variable of interest Y in the population are regarded as fixed but unknown quantities. The only source of randomness is the sampling design. Design-based properties such as design expectation and variance of an estimator are evaluated under hypothetical repeated sampling by a given sampling design from the fixed population. We will examine these properties empirically for some estimators of totals by small-scale design-based simulation experiments.

2.5 Estimation of population parameters

In sample surveys, the unknown value of the population parameter of interest, such as total, is estimated by using the observed sample values of target variable under the chosen sampling design and estimation design. *Estimation design* is characterized by the structure of an estimator, including the way how auxiliary information is incorporated in the estimation procedure. In the handbook, a combination of sampling design and estimation design is called *strategy*.

Point estimation. For population total $t = \sum_{i=1}^{N} y_k$, a common general purpose estimation design is provided by the *Horvitz-Thompson estimator* (expansion estimator), given as

$$\hat{t}_{HT} = \sum_{k=1}^{n} w_k y_k = \sum_{k=1}^{n} y_k/\pi_k \,, \tag{2}$$

where *sampling weights* or *design weights* w_k are defined as inverses of inclusion probabilities, i.e. $w_k = 1/\pi_k$ for sample element k. In HT estimation, information on the sampling design is incorporated in the estimation procedure by sampling weights. *Calibration estimator* of total, given by

$$\hat{t}_{CAL} = \sum_{k=1}^{n} w_{CAL,k} y_k, \tag{3}$$

is another general purpose estimation design. In CAL estimation, information on sampling and estimation designs is incorporated in the estimation procedure by a combined element weight $w_{CAL,k} = w_k \times g_k$, where $w_k = 1/\pi_k$ is sampling weight and the sample-dependent weights g_k are specific for each calibration estimator. All model-assisted estimators for total discussed in the handbook can be expressed in the form (3).

HT estimator for population total is *design unbiased*: the design expectation of the estimator equals the true parameter value. All calibration estimators as well as model-assisted estimators of total considered in the handbook are *design consistent*; their design bias and variance tend to zero as the sample size increases. The most important of these estimators are *nearly design unbiased*, which is a favorable property of an estimator. The design bias of the estimator is an asymptotically insignificant contribution to its mean squared error (MSE). MSE is defined as the sum of design variance and squared bias of estimator.

Quality indicators. The degree of uncertainty attached to an estimated total is measured by *design variance* $V_{p(s)}(\hat{t})$ of an estimator $\hat{t}$ of a total, defined under a given sampling design and estimation design. Design variance is an unknown parameter and must be estimated from the sample. An estimator $\hat{v}_{p(s)}(\hat{t})$ of design variance of $\hat{t}$ also depends on the applied strategy. *Standard error* of $\hat{t}$ is defined as square root of the design variance and is estimated by

$$s.e(\hat{t}) = \sqrt{\hat{v}_{p(s)}(\hat{t})} \,. \tag{4}$$

Coefficient of variation of total estimate is defined as

$$cv(\hat{t}) = \frac{s.e(\hat{t})}{\hat{t}}, \tag{5}$$

often expressed as a percentage. Coefficients of variation are routinely used in official statistics when assessing the precision quality of estimates for publication.

Design effect of total estimator $\hat{t}$ is used in surveys to assess the efficiency of a strategy relative to a reference strategy, expressed as

$$DEFF_{p(s)}(\hat{t}) = \frac{V_{p(s)}(\hat{t})}{V_{SRS}(\hat{t}_{HT})}, \tag{6}$$

where $V_{p(s)}(\hat{t})$ is the design variance of estimator $\hat{t}$ under the *actual* sampling design and $V_{SRS}(\hat{t}_{HT})$ is the design variance of the HT estimator $\hat{t}_{HT}$ under the *reference* sampling design, usually simple random sampling without replacement. An estimator of $DEFF$ is constructed by using the sample counterparts of the design variances and can be written as

$$deff_{p(s)}(\hat{t}) = \frac{\hat{v}_{p(s)}(\hat{t})}{\hat{v}_{SRS}(\hat{t}_{HT})} \tag{7}$$

The estimator $\hat{t}$ of the total in the numerator variance expression may differ from that in the reference variance formula, as is the case for calibration and model-assisted estimators.

By definition, if the $deff$ is smaller than one, the actual strategy is more efficient than the reference strategy SRSWOR_HT, where sampling is with simple random sampling without replacement and estimation relies on the HT estimator. If $deff = 1$ then the actual and reference strategies are equally efficient. In cluster sampling, design effects are usually larger than one. One of the main aims of the handbook is to introduce sampling and estimation designs that attain improved efficiency over the SRSWOR_HT strategy.

2.6 Estimation for population subgroups

Estimates are often required for important population subgroups such as regional areas in a country or different vessel or fishing types. It is advisable to define the most important subpopulations as *strata* in the sampling design by using *stratified sampling* (see Section 3.6). Domains that are defined as strata are called *planned domains* and they are considered as independent subpopulations, and a separate sample is drawn from each of them. Sample sizes n_d in planned domains are usually fixed by the sampling design. Domain estimates and their associated quality indicators can be readily obtained by methods of the handbook applied separately for each subpopulation.

Subpopulations of interest that are not specified in advance but emerge after sampling and data collection are called *unplanned domains* and are denoted $U_d, d = 1, \dots, D$, where D is the number of domains. Unplanned domains are usually non-overlapping subgroups of the population not related to the sampling design. A single n element sample has been drawn, and sample sizes n_d for domains are not controlled by the sampling design but are random quantities such that $\sum_{d=1}^{D} n_d = n$, the overall sample size. Sample sizes in some domains can be small (even zero) and special techniques of *small area estimation* may be needed.

The random nature of domain sample sizes affects inference. Formally, there are two different approaches for inference for unplanned domains. In an *unconditional approach*, inference is based on hypothetical repeated sampling with sampling design $p(s)$ such that the overall sample s of n elements is allowed to distribute randomly over domain samples $s_d \subset s, d = 1, \dots, D$. Thus, all possible domain sample configurations are considered when averaging over variations in domain sample size, including configurations that did not occur. In the *conditional approach*, the procedure is conditional given the observed configuration of the n element sample s into domain samples s_d. Thus, only samples whose domain sample sizes correspond to the observed domain sample sizes are considered.

The inferential approach together with the chosen sampling and estimation designs affects the estimation. Typically, variances of total estimators under the unconditional approach tend to be larger than those of the conditional approach. The situation is similar as in post-stratification.

In practice, point and variance estimators for unplanned domains under the unconditional approach can be constructed by using *extended domain variables* defined as $y_{dk} = y_k$ if $k \in U_d$ and zero otherwise. The population total in domain d can thus be expressed as $t_d = \sum_{i=1}^{N} y_{dk}, d = 1, \dots, D$. For example, Horvitz-Thompson estimator (2) of domain total t_d for domain d takes the form

$$\hat{t}_{dHT} = \sum_{k=1}^{n} w_k y_{dk} = \sum_{k=1}^{n} y_{dk}/\pi_k \,. \qquad (8)$$

For variance estimation of $\hat{t}_{dHT}$ under the unconditional approach, the extended domain variable values y_{dk} are inserted in variance expressions instead of the original values y_k. In the conditional approach, where domain sample sizes are considered fixed and the domains are treated as independent subpopulations, the original values y_k are used in variance formulas. Instead of HT estimation, various model-assisted and calibration methods can be used, such as ratio estimation and post-stratification.

Design-based estimation for domains is discussed for example in Lehtonen & Veijanen (2009). Hidiroglou & Patak (2004) compared the conditional and unconditional approaches for various estimators of totals for unplanned domains. Some of the pioneering authors in the area, e.g. Durbin (1969) and Holt and Smith (1979), as well as more recent contributors (e.g. Särndal et al. 1992) favoured the conditional approach of inference for unplanned domains.

We demonstrate in sections 3.3.5 and 4.2.4 the various approaches for the estimation of domain totals and their design variances in connection to simple random sampling, HT estimation and post-stratification by using tools available in SAS survey programs.

3 Basic sampling methods

3.1 Sampling designs

Basic sampling designs for surveys can be divided into methods for element sampling and methods for multi-stage sampling, where the latter group consists of combinations of element sampling methods. A selection of methods is listed and characterized in Table 3.1.

Table 3.1 Basic sampling designs for sampling from finite populations.

Basic sampling designs		
	Auxiliary data needed in sampling frame	**Level and source of auxiliary data**
A. Element sampling		
Equal probability sampling designs		
(1) Simple random sampling SRS	Element identification variable	Unit-level Sampling frame
(2) Systematic sampling SYS	Element identification variable Implicit stratification: Sorting variable	
Unequal probability sampling designs		
(3) Sampling with probability proportional to size PPS	Element identification variable Size measure variable for elements	Unit-level Sampling frame
(4) Balanced sampling	Element identification variable Balancing variables	
(5) Stratified sampling STR	Element /cluster identification variable Stratification variables (categorical)	Element sampling: Unit-level Cluster sampling: Cluster level Sampling frame
	Optimal and power allocation: Additional auxiliary information needed	
B. Multi-stage cluster sampling		
(6) One-stage cluster sampling with SRS or SYS	Cluster identification variable	Cluster level Sampling frame for clusters
(7) One-stage cluster sampling with PPS	Cluster identification variable Size measure for clusters	
(8) Stratified one-stage cluster sampling	Cluster identification variable Stratification variables for clusters	
(9) Stratified two-stage (multi-stage) cluster sampling	Cluster identification variable Stratification variables for clusters Element identification variable for sample clusters	Cluster level Sampling frame for clusters Sampling frame for elements in sample clusters

The basic sampling techniques in Table 3.1 part A constitute methods for drawing population elements into the sample. *Simple random sampling* (SRS sampling) and *systematic sampling* (SYS sampling) are *equal probability sampling methods*: the probability of population element to be included in the sample is the same for all population elements. *Probability proportional to size sampling* (PPS sampling) and *balanced sampling* are *unequal probability sampling methods*, where the inclusion probabilities can vary between elements. These four methods are used for obtaining probability samples from the target population of the sample survey.

In *stratified sampling* (STR sampling), population elements are first grouped into non-overlapping subpopulations called *strata*. The strata are independent subpopulations, and a sample of elements is drawn from each stratum by one of the methods (1)-(4) in Table 3.1. The number of sample elements (individual elements or groups of elements called *clusters*) drawn from each stratum is defined by *allocation methods*. Stratified sampling can thus be applied for sampling of individual elements or groups of elements or clusters.

Part B contains combinations of methods of part A of varying complexity, depending on the requirements of the survey. For example in stratified cluster sampling, the population of clusters is first stratified. A sample of clusters is then drawn from each stratum, and the individual elements are selected from the sample clusters. In the handbook we discuss element sampling designs of part A, methods (1) to (3) and stratified sampling (5), which constitute the most popular schemes in fisheries statistics practice.

3.2 SIMPOP population

Computational examples in the handbook are based on an artificial population SIMPOP containing complete records for $N = 120$ fishing vessels (elements, units) on $p = 20$ variables. Table 3.2 presents the list of variables in SIMPOP. All variables are of numeric type.

Variables in the SIMPOP represent typical variable types in fisheries statistics at reference year. Variable ID is unique vessel identification code obtained from a vessel register. Variable STR3 is stratum variable for stratified sampling. For stratification, every population vessel is assigned a value indicating stratum membership. STR3 has been constructed by dividing the 120 population vessels into three equal-sized groups based on variable GT (vessel tonnage), whose values are known for all population vessels. The role of variable DOM01 is different. While STR3 is used in the sampling phase for the grouping of the population vessels into strata, DOM01 is not related to the sampling design. DOM01 is used for grouping of population elements after drawing the sample and data collection. The variable will be used in estimation for population subgroups (domains) and in post-stratification. DOM01 indicates whether a vessel catches "expensive" fish (DOM01 = 1) or not (DOM01 = 0). ACTIVITY indicates whether a vessel has been active in the reference time period considered (ACTIVITY = 1) or not (ACTIVITY=0). Of the 120 population vessels, 100 are coded active and 20 non-active. The main share of computational examples in this section consider the set of active vessels.

Table 3.2. The list of variables in the SIMPOP data set.

Variables in Creation Order		
#	**Variable**	**Label**
1	ID	Unique identification code
2	STR3	Stratum variable (3 strata)
3	DOM01	Fishing type (domain variable with 2 classes)
4	LENGTH	Length of vessel (meters)
5	GT	Vessel tonnage (GT)
6	kW	Engine power (kW)
7	ACTIVITY	Vessel activity indicator (1=active, 0=otherwise)
8	DAS	Days at sea
9	GT_DAS	GT_Days
10	kW_DAS	kW_Days
11	CATCH	Catch (ton)
12	VALUE	Value of landings (Euro)

Variables in Creation Order		
#	**Variable**	**Label**
13	FUEL	Fuel costs
14	LABOR	Labour costs
15	OTHER_VAR	Other variable costs
16	REPAIR	Repair costs
17	OTHER_NONVAR	Other non-variable costs
18	TOTAL_COST	Total costs
19	GROSS_PROFIT	Gross profit

In the examples of this section, some variables in SIMPOP are treated as target variables (variables of interest) and some are auxiliary variables. The target variables are of main interest in a fishery survey. Data for the target variables are collected from a sample drawn for the survey and further, incorporated in the estimation procedures. There are 9 potential target variables in SIMPOP. Values of these variables are coded zero for non-active vessels. Examples of relevant target variables are CATCH, VALUE and TOTAL_COST.

Data for the auxiliary variables are often taken from national registers on fishery or other administrative data sources. Some auxiliary variables are used in the sampling phase and they must be included in the sampling frame. Examples are STR3 for stratified sampling and GT for size variable in PPS sampling. Some auxiliary variables are used the estimation procedures. Variables GT and DAS (if included in frame) are examples of auxiliary variables suitable for ratio and regression estimation. In the worked examples, we assume that data for the auxiliary variables are available as aggregate-level values or unit-level values of the auxiliary variables, depending on the requirements of the chosen method.

In our examples, the roles of some variables can change depending on the given statistical data infrastructure. For example, in some examples variable days at sea (DAS) is treated as auxiliary variable. In this case, data on DAS are available for all vessels in the population. In some cases, DAS is a target variable and then its measurements are assumed known for sample vessels only. The variable ACTIVITY is treated as auxiliary variable in most cases. Descriptive statistics on selected variables in the entire SIMPOP are presented in Table 3.3. The data set contains both active and inactive vessels.

Table 3.3 Descriptive statistics of selected variables in the SIMPOP data set (all vessels).

Variable	N	Mean	Total	Minimum	Maximum
CATCH	120	5200	624036	0	13391
VALUE	120	1622301	194676173	0	5278581
TOTAL_COST	120	1041983	125037964	0	3059456
GROSS_PROFIT	120	580318	69638209	-36978	2263956
DAS	120	152.56667	18308	0	250.00000

Variable	N	Mean	Total	Minimum	Maximum
GT	120	326.45750	39175	210.60000	444.60000
kW	120	831.52427	99783	426.46500	1332
GT_DAS	120	50326	6039087	0	103565
kW_DAS	120	128476	15417180	0	309024

In the table, the first four variables are potential target variables for our examples and the rest are candidates for auxiliary variables. For the non-active vessels, the population values of the four target variables and the variable DAS are coded zero.

We next concentrate on the population of active vessels, where measurements for the target variables are available. Let us first examine the relations between the target variables. Table 3.4 presents their pair-wise correlations. The target variables appear strongly correlated. Highest correlation (0.98) is for VALUE and GROSS_PROFIT and lowest (0.56) is for CATCH and GROSS_PROFIT.

Table 3.4. Correlation matrix of selected target variables.

Pearson Correlation Coefficients, N = 100				
	CATCH	**VALUE**	**TOTAL_COST**	**GROSS_PROFIT**
CATCH	1.00000	0.61567	0.64087	0.56149
VALUE	0.61567	1.00000	0.97758	0.97664
TOTAL_COST	0.64087	0.97758	1.00000	0.90948
GROSS_PROFIT	0.56149	0.97664	0.90948	1.00000

Scatter Plot Matrix of target variables is presented in Figure 3.1. The mutual relationships of the three target variables appear to be of linear type. For CATCH, there seems to be two groups of vessels separated by VALUE, TOTAL_COST and GROSS_PROFIT.

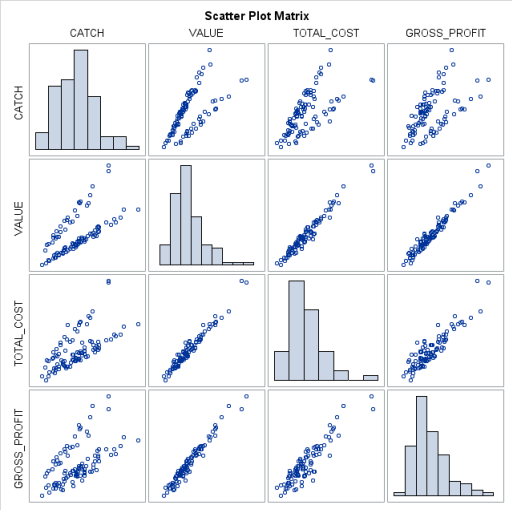


Figure 3.1 Scatter Plot Matrix of the three target variables (active vessels).

The availability of high-quality auxiliary data is a cornerstone for developing efficient estimation designs in fisheries statistics. It is thus important to put resources on collecting such data. For improved accuracy, it is beneficial if the auxiliary variables have strong relationship with the target variables. Correlation matrix of target variables with DAS, GT and kW is displayed in Table 3.5.

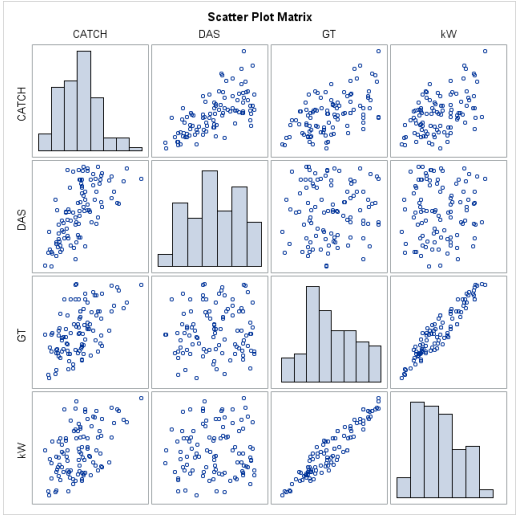
The table shows that CATCH correlates quite strongly with DAS, GT and kW. The corresponding correlations of TOTAL_COST are somewhat weaker and for GROSS_PROFIT even more weaker. Obviously, GT and kW correlate strongly, but the correlation of DAS with GT and kW is weak.

Table 3.5 Correlations of selected target variables with auxiliary variables (active vessels).

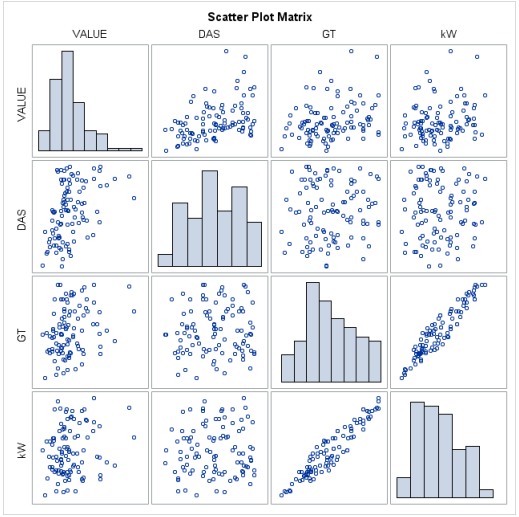
Pearson Correlation Coefficients, N = 100			
	DAS	**GT**	**kW**
CATCH	0.66039	0.55892	0.49882
VALUE	0.42809	0.27729	0.18079
TOTAL_COST	0.44040	0.42185	0.33924
GROSS_PROFIT	0.39574	0.11694	0.01071

The table shows that CATCH correlates quite strongly with DAS, GT and kW. The corresponding correlations of TOTAL_COST are somewhat weaker and for GROSS_PROFIT even more weaker. Obviously, GT and kW correlate strongly, but the correlation of DAS with GT and kW is weak.

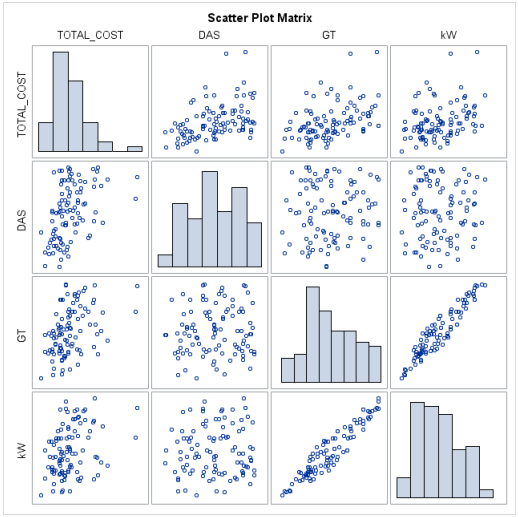
Figure 3.2 contains Scatter Plot Matrices of the four target variables with selected auxiliary variables, divided into four submatrices. Panels A, B, C and D indicate the association of each target variable with the three selected auxiliary variables, as well as the mutual relations between the three auxiliaries.



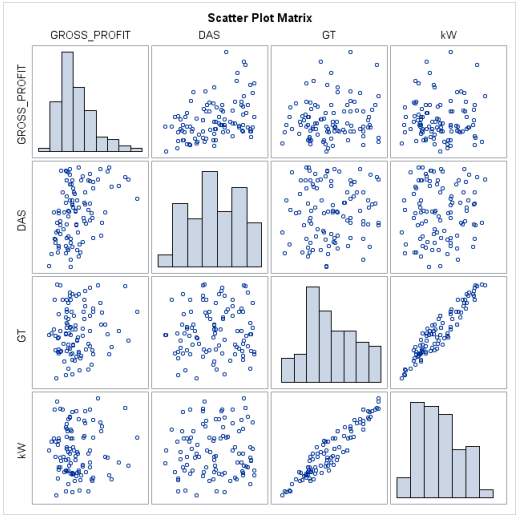
A. CATCH against DAS, GT and kW



B. VALUE against DAS, GT and kW



C. TOTAL_COST against DAS, GT and kW



D. GROSS_PROFIT against DAS, GT and kW

Figure 3.2 Scatter Plot Matrices of the four target variables with three auxiliary variables (active vessels).

Let us consider the derived variables GT_DAS and kW_DAS, which are constructed as products of GT with DAS and kW with DAS, respectively, for further illustration of relations between the target variables and auxiliary variables. Scatter Plot Matrix is in Figure 3.3.

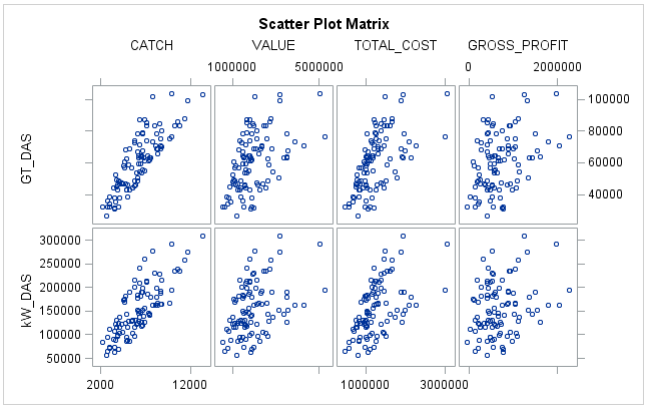


Figure 3.3 Scatter Plot Matrix of the four target variables with GT_DAS and kW_DAS (active vessels).

Pearson correlation coefficients of the variables are collected in Table 3.6. The new variables GT_DAS and kW_DAS indicate stronger relations to the target variables than the auxiliary variables DAS, GT and kW in the previous figure. The table shows that correlations of GT_DAS and kW_DAS with CATCH are substantial and pretty large with TOTAL_COST. GROSS_PROFIT seems to be less correlated to these auxiliary variables.

Table 3.6 Correlation of the four target variables with GT_DAS and kW_DAS (active vessels).

Pearson Correlation Coefficients, N = 100				
	CATCH	**VALUE**	**TOTAL_COST**	**GROSS_PROFIT**
GT_DAS	0.84440	0.50048	0.60103	0.37471
kW_DAS	0.79492	0.41083	0.54064	0.25935

Accuracy gains can be substantial if the relation between a target variable and an auxiliary variable is strong. Gains may remain minor if the relation is weak. A careful examination of the relations between the target variables and potential auxiliary variables is an important task in a fishery survey process. For stratified sampling

and PPS sampling, the relations can be studied from previous fishery surveys, for example. For calibration and model-assisted estimation with ratio or regression estimation, the collected survey data itself provides a source of information for the relations. We present in Chapter 6 estimation results for several target variables.

In our discussion this far, the variable DAS was treated as an auxiliary variable whose values are available at the unit (vessel) level in the sampling frame. This is possible in a number of fisheries where logbook data are available. In this case, DAS can be used both in the sampling phase or in the estimation phase. In other cases, data on DAS must be collected from a sample of active vessels. DAS cannot be used in the sampling phase but the use in the estimation phase is still possible, if aggregate-level data on DAS are available as population totals or means.

The population data set SIMPOP was constructed to contain complete records on both target variables and auxiliary variables, for all active vessels. We can thus compute numerical values for true population parameters, such as totals, and compare them with their sample-based estimates. This is important for pedagogical purposes.

In practice, however, the values of target variables are assumed known for the sample vessels only, and the population values are unknown. Still, values for auxiliary variables may be available in the sampling frame or at least as aggregates.

3.3 Simple random sampling

3.3.1 Background

Simple random sampling (SRS) suites for situations where useful auxiliary information is not available. In such cases it is reasonable to assign equal selection probability for each unit of the population. Simple random sampling is also a natural candidate for a reference method when comparing the efficiency of other sample selection methods. Furthermore, SRS is often integrated into more complex sampling procedures for the final randomization. Auxiliary information is, however, not utilized in the SRS sampling, even though it would be available.

3.3.2 Sample selection techniques

An equal selection probability is a common factor in the sample selection techniques of simple random sampling. In a population $U = \{1, \dots, k, \dots, N\}$ of N units, the probability of inclusion of element k in a n element simple random sample is $\pi_k = \pi = n/N$ for every population element $k \in U$. SRS designs are thus equal probability sampling designs.

There are three main sample selection techniques in SRS sampling: Bernoulli sampling, simple random sampling with replacement and simple random sampling without replacement. In *Bernoulli sampling*, a random number from uniform (0,1) distribution is drawn and attached to each element of the population. Then all the elements with random number smaller than a pre-fixed constant $\pi = n/N$ are drawn into the sample. In practice, sampling is carried out in a list-sequential manner applied to the sampling frame. Bernoulli sampling is a without-replacement type sampling technique. Because of the selection method, sample size in Bernoulli sampling is random with expected value $E(n_s) = N\pi$. In *conditional Bernoulli sampling* only samples of size n are accepted.

Replacement of drawn element after each random draw (*Simple random sampling with replacement*, SRSWR) guarantees that the inclusion probability remains equal for each draw. To draw a SRSWR sample of size n, the first element is drawn with probability $1/N$ and put back into the frame. The procedure is repeated n times to obtain the sample; the same unit can appear more than once in the sample. Samples from SRSWR are independent and the design variances of estimators are simpler than for without-replacement type designs. However, SRSWR is rarely used in sampling practice. The more common SRS method for practical purposes is *simple random sampling without replacement* (SRSWOR). Inclusion probability is still equal for each element in each separate draws, but the probability changes draw by draw as the draw progresses, because the number of elements in population frame decreases after each draw.

3.3.3 Estimation of parameters

The Horvitz-Thompson estimator (2) of the population total $t = \sum_{k=1}^{N} y_k$ under simple random sampling takes the form

$$\hat{t}_{HT} = \sum_{k=1}^{n} w_k y_k = N/n \times \sum_{k=1}^{n} y_k = N \times \hat{\bar{y}}_{HT}, \quad (9)$$

where the sampling weights are constant $w_k = N/n$, n is the sample size, and N is the population size. Population mean $\bar{Y} = t/N$ is estimated by the sample mean $\hat{\bar{y}}_{HT} = \frac{\hat{t}_{HT}}{N} = \sum_{k=1}^{n} y_k/n$.

Design variance of (9) for simple random sampling without replacement (SRSWOR) is given by

$$V_{SRSWOR}(\hat{t}_{HT}) = N^2 \left(1 - \frac{n}{N}\right) S^2/n$$

and the variance is estimated by

$$\hat{v}_{SRSWOR}(\hat{t}_{HT}) = N^2 \left(1 - \frac{n}{N}\right) \hat{s}^2/n \quad (10)$$

where $S^2 = \sum_{k=1}^{N}(y_k - \bar{Y})^2/(N-1)$ is the population variance of the target variable Y and $\hat{s}^2 = \sum_{k=1}^{n}(y_k - \hat{\bar{y}}_{HT})^2/(n-1)$ is the sample counterpart. The term $1 - \frac{n}{N}$ is the so-called *finite population correction factor* (FPC factor) that channels the effect of relative sample size (sampling fraction $f = n/N$) to the variance formulas.

Standard error and coefficient of variation for total estimate $\hat{t}_{HT}$ are estimated by formulas (4) and (5), respectively. For SRSWOR, the design effect is equal to one because sampling and estimation designs are the same in the numerator and denominator.

For simple random sampling with replacement (SRSWR), the design variance is $V_{SRSWR}(\hat{t}_{HT}) = N^2 \left(1 - \frac{1}{N}\right) S^2/n$, and the design effect is $DEFF_{SRSWR}(\hat{t}_{HT}) = \frac{V_{SRSWR}(\hat{t}_{HT})}{V_{SRSWOR}(\hat{t}_{HT})} = \frac{N-1}{N-n}$. Simple random sampling with replacement (SRSWR) is, therefore, less efficient than simple random sampling without replacement (SRSWOR), for sample sizes bigger than one ($n > 1$).

3.3.4 Worked example

Preliminaries. We consider here the population of $N = 100$ active vessels in SIMPOP. We execute the estimation of the population total of variable CATCH in the case where no auxiliary data are assumed, except the size N of the population (this piece of information is needed for variance estimation). We use the basic estimation strategy SRSWOR_HT, where the element sample is drawn by simple random sampling without replacement (SRSWOR), and the estimation relies on the Horvitz-Thompson (HT) estimator.

We demonstrate the effect of the sample size to variance, standard error, coefficient of variation and design effect estimates of the estimated total.

Sample selection. Our first SRSWOR sample size is $n = 5$ active vessels and thus, we draw a 5% sample from SIMPOP. The realized sample is listed in Table 3.7. SAMPLE1 represents one of the possible samples of size $n = 5$ active vessels that can be drawn with SRSWOR from SIMPOP. The sample has been drawn with the SAS procedure SURVEYSELECT. The variable SAMPLINGWEIGHT generated by the procedure is sampling method specific and is included in the sample data set by default.

Table 3.7 SAMPLE1 of $n = 5$ active vessels drawn from SIMPOP of $N = 100$ vessels.

Obs k	**ID**	**CATCH** y_k	**SamplingWeight** w_k
1	1	3541.44	20
2	44	4421.92	20
3	49	11355.97	20
4	55	6865.42	20
5	93	9942.19	20
Sum		36126.94	100

Estimation. Let us compute the estimates for CATCH from SAMPLE1 by using the computational formulas of Section 3.3.3. Because for SRSWOR the inclusion probabilities π_k are constants i.e. $\pi_k = \pi = 5/100 = 0.05$ for all active vessels in the population, the weights w_k for the sample vessels also are constants: $w_k = w = 1/\pi = 20$, and the sum of weights is 100 $(= N)$. The weights are needed in the construction of the HT estimator $\hat{t}_{HT}$ for CATCH total. By using the Horvitz-Thompson estimator (9) we obtain:

$$\hat{t}_{HT} = \sum_{k=1}^{5} w_k y_k = 20 \times \sum_{k=1}^{5} y_k = 20 \times 36126.94 = 722539.$$

It is noted that a HT estimator of a total is simply a sum of weighted sample observations of the target variable Y, where weights are inverses of the inclusion probabilities.

For statistical inference we need a variance estimate $\hat{v}_{SRSWOR}(\hat{t}_{HT})$ and standard error estimate $s.e(\hat{t}_{HT})$ of the total estimate$\hat{t}_{HT}$. By (10), variance is estimated by

$$\hat{v}_{SRSWOR}(\hat{t}_{HT}) = 100^2 \left(1 - \frac{5}{100}\right) \hat{s}^2/5 = 147823^2,$$

where sample variance of CATCH is $\hat{s}^2 = 11500800.86$. Standard error estimate is $s.e(\hat{t}_{HT}) = 147823$.

By using the estimated total and its $s.e$ we compute a two-sided 95% confidence interval for the estimated total. The interval is calculated as $\hat{t}_{HT} \pm s.e(\hat{t}_{HT}) \times t_{df,\alpha/2}$ where $t_{df,\alpha/2}$ is the chosen $100\left(1 - \frac{\alpha}{2}\right) = 97.5$ percentile point of the *t* distribution with $df = n - 1 = 4$ degrees of freedom and $\alpha = 0.5$. For lower limit we get $LCL(\hat{t}_{HT}) = 312117$ and for upper limit $UCL(\hat{t}_{HT}) = 1132960$. The interval is quite wide.

Coefficient of variation for $\hat{t}_{HT}$ is calculated by (5) as $cv(\hat{t}_{HT}) = \frac{s.e(\hat{t}_{HT})}{\hat{t}_{HT}} = \frac{147823}{722539} = 0.20$.

The design effect estimate of $\hat{t}_{HT}$ is $deff(\hat{t}_{HT}) = 1$ in the SRSWOR design. By using the SAS procedure SURVEYMEANS we obtain the results in Table 3.8.

Table 3.8 Estimated total, standard error and coefficient of variation for variable CATCH from SRSWOR sample SAMPLE1 of $n = 5$ vessels.

Variable	**True value**	**n**	**Sum of Weights**	**Total** $\hat{t}$	**Std Dev** $s.e(\hat{t})$	**95% CL**		**Coeff of Var** $cv(\hat{t})$
CATCH	624036	5	100.000000	722539	147823	312117.193	1132960.36	0.204588

The estimated total is $\hat{t}_{HT} = 722539$, standard error is $s.e(\hat{t}_{HT}) = 147823$ and coefficient of variation is $cv(\hat{t}_{HT}) = 0.204588$ i.e. 20%. The figures are the same as obtained by the computational formulas. The true parameter value is $t = 624036$ and for SAMPLE1 of *n* = 5 vessels, the 95% confidence interval would cover the true value. But the confidence interval is too wide for any practical purposes. The results seem not reliable enough.

We next draw a larger SRSWOR sample SAMPLE2 of size of $n = 20$ vessels. Estimates computed by SURVEYMEANS are in Table 3.9. Estimated standard error of total estimate is now much smaller than that from SAMPLE1 of $n = 5$ vessels. The estimated total is much closer to the true value, and the confidence interval is substantially narrower than for SAMPLE1. Not surprisingly, we obtain more precise estimation from a larger sample.

Table 3.9 Estimated total, standard error and coefficient of variation for variable CATCH from SRSWOR sample SAMPLE2 of $n = 20$ vessels.

Variable	True value	n	Sum of Weights	Total $\hat{t}$	Std Dev $s.e(\hat{t})$	95% CL		Coeff of Var $cv(\hat{t})$
CATCH	624036	20	100.000000	610603	54439	496661.885	724544.886	0.089156

Simulation experiment. We noted that the estimated total of CATCH from SAMPLE1 is far from the true value. This is because of the randomization mechanism underlying the sampling technique. To throw more light on this, we carry out a small pedagogic simulation experiment. We draw a reasonable number, let say $K = 100$ SRSWOR samples of small size $n = 5$ vessels from SIMPOP, compute the estimated total, standard error and coefficient of variation from each sample, and compute the mean of the statistics from the 100 samples. Then we do the same for a larger sample size $n = 20$. The results are in Table 3.10.

Table 3.10 Means of estimated totals, standard errors and coefficients of variation for CATCH from *K* = 100 simulated SRSWOR samples of sizes $n = 5$ and $n = 20$ vessels from SIMPOP.

Method	VarName	Replicates	Averages over simulations				
			SumWgt	n	Total $\hat{t}$	StdDev $s.e(\hat{t})$	CV $cv(\hat{t})$
SRSWOR	CATCH	100	100.000000	5	629966	91436	0.145160
SRSWOR	CATCH	100	100.000000	20	626895	44061	0.070264
True total					624036		

The following conclusions can be drawn. On average, the estimated totals tend to closely coincide with the true total. This holds for both sample sizes. The important property of design unbiasedness of the HT estimator is often appreciated in official statistics production. Official statistics tends to be a quite conservative affair and in that framework, people often want to stay on the safe side and try to avoid unpredictable design bias. Further, the average standard error and coefficient of variation decline when sample size increases. This means that for samples of size $n = 20$, the spread of total estimates computed from sample replicates are more condensed around the true total than those of samples of size $n = 5$ vessels.

The average coefficient of variation (cv) is 14% for sample size $n = 5$ and 7% for $n = 20$. This means that with four times larger sample the gain in efficiency is substantial. However, increasing the sample size for improved statistical efficiency is not necessarily optimal for cost efficiency. With a fixed sample size *n*, statistical efficiency can be improved over SRSWOR by a more effective sampling design, e.g. stratified sampling or PPS sampling, or by using auxiliary information in the estimation phase, for example with calibration or regression estimation.

3.3.5 Estimation for domains

Estimates are often requested for unplanned domains i.e. population subgroups that are not defined as strata in the sampling design. Principles for estimation for unplanned domains under the conditional and unconditional approaches was discussed in Section 2.5. We estimate the domain totals of CATCH for the SRSWOR sample SAMPLE2 of $n = 20$ vessels under strategy SRSWOR_HT. Strategy SRSWOR_RAT is applied for domain estimation in Section 4.2.4.

The domain variable DOM01 (type of fishing) indicates whether a vessel catches "expensive" fish (DOM01 = 1) or not (DOM01 = 0). DOM01 creates two unplanned domains whose sample sizes n_d are not controlled by the sampling design. The distribution of the data into the two domains is in the set-up below.

Domain d	Sample n_d	Sum of Weights $\widehat{N}_d$	Population N_d	Population totals of CATCH
0	12	60	70	457163
1	8	40	30	166873
Sum	20	100	100	624036

Estimate $\widehat{N}_d$ is the sum of sampling weights w_k in domain d $(d = 0,1)$ defined as the HT estimate of domain size N_d in population. Note that $\widehat{N}_d$ are not equal to the population counterparts N_d as would be the case if the domains were planned type domains.

Horvitz-Thompson estimator (2) of domain total t_d of CATCH for domain d can be expressed as $\hat{t}_{dHT} = \sum_{k \in s_d} w_k y_k = 5 \times \sum_{k \in s_d} y_k$, where notation $k \in s_d$ means summation over sample elements in domain sample s_d, and $d = 0$ for the first domain and $d = 1$ for the second domain. HT estimates for domain totals under conditional and unconditional approaches are:

Domain 0: $\hat{t}_{0HT} = \sum_{k \in s_0} w_k y_k = 5 \times \sum_{k \in s_0} y_k = 419536$

Domain 1: $\hat{t}_{1HT} = \sum_{k \in s_1} w_k y_k = 5 \times \sum_{k \in s_1} y_k = 191067$

The sum of domain estimates equals the total estimate for the entire population in Table 3.9, so the HT estimator is additive. Three scenarios are applied for variance estimation.

Scenario 1: Estimation under the conditional approach with known N_d. Domains are treated as independent subpopulations similarly as for planned domains i.e. strata. For variance estimation we use the estimator (13) of Section 3.5.3 separately for each domain d, given by

$$\hat{v}_{SRSWOR}(\hat{t}_{dHT}) = n_d \left(1 - \frac{n_d}{N_d}\right) \sum_{k \in s_d} (w_k y_k - \hat{t}_{dHT}/n_d)^2/(n_d - 1),$$

where $\hat{t}_{dHT} = \sum_{k \in s_d} w_k y_k$ is the HT estimator of domain total in domain d, $d = 0,1$. Original values y_k are used in the estimator. Variance estimates for domain totals are

Domain 0: $\hat{v}_{SRSWOR}(\hat{t}_{0HT}) = 12 \times \left(1 - \frac{12}{70}\right) \times 2580705499.1/(12 - 1) = 48298^2$

Domain 1: $\hat{v}_{SRSWOR}(\hat{t}_{1HT}) = 8 \times \left(1 - \frac{8}{30}\right) \times 349481459.16/(8 - 1) = 17114^2$

Scenario 2: Estimation under the conditional approach with unknown N_d. This situation is often met in practice. Original values y_k are again used. Variance estimator is

$$\hat{v}_{SRSWR}(\hat{t}_{dHT}) = n_d \sum_{k \in s_d} (w_k y_k - \hat{t}_{dHT}/n_d)^2/(n_d - 1),$$

Note that there is no fpc, contrary to Scenario 1. Variance estimates are:

Domain 0: $\hat{v}_{SRSWR}(\hat{t}_{0HT}) = 12 \times 2580705499.1/(12 - 1) = 53060^2$

Domain 1: $\hat{v}_{SRSWR}(\hat{t}_{1HT}) = 8 \times 349481459.16/(8 - 1) = 19985^2$

Standard errors increase relative to Scenario 1, because we did not have access to N_d.

Scenario 3. Estimation under the unconditional approach. Estimates for domains are computed using extended domain variables with values $y_{dk} = y_k$ if $k \in U_d$ and zero otherwise, $d = 0,1$, involving two extended domain variables y_{0k} and y_{1k}. Hence, the Horvitz-Thompson estimator (2) of domain total t_0 for domain 0 is $\hat{t}_{0HT} = \sum_{k=1}^{n} w_k y_{0k}$, and $\hat{t}_{1HT} = \sum_{k=1}^{n} w_k y_{1k}$, leading to same numerical estimates as for scenarios 1 and 2 under the conditional approach. Variance estimator for domain d is expressed as (Lehtonen & Veijanen 2009 p. 227):

$$\hat{v}_{SRSWOR}(\hat{t}_{dHT}) = n \left(1 - \frac{n}{N}\right) \sum_{k \in s} (w_k y_{dk} - \hat{t}_{dHT}/n)^2/(n - 1),$$

where $\hat{t}_{dHT} = \sum_{k \in s} w_k y_{dk}$, $d = 0,1$, is the HT estimator of domain total of the extended domain variable y_{dk} for the entire sample. Note that the sum extends over all elements in the sample and $y_{dk} = 0$ for elements outside domain d, but also these elements contribute to the variance estimate for the domain because $\hat{t}_{dHT}$ is nonzero. Variance estimates are:

Domain 0: $\hat{v}_{SRSOWR}(\hat{t}_{0HT}) = 20 \times 8447718032.6/(20-1) = 84344^2$

Domain 1: $\hat{v}_{SRSWOR}(\hat{t}_{1HT}) = 20 \times 3087490102.7/(20-1) = 50990^2$

Estimates were computed by SAS procedure SURVEYMEANS (Section A.3). For Scenario 1, we estimated separately for the two domains with `TOTAL=` option to define the domain sizes N_d in population for finite population corrections (fpc). For Scenario 2, fpc was not given and we used the `BY` statement for the entire sample, which invokes separate analyses for the two domains. For Scenario 3, estimates were computed with the `DOMAIN` statement for the entire sample, which accounts for the extra variance via the extended domain variables. Equal results Scenario 3 are obtained by the R `survey` function `svyby` (Section B.3.4).

Results for the three scenarios are presented in Table 3.11. The HT estimated CATCH totals are identical in all scenarios. The differences are in standard error estimates.

Table 3.11 Estimation of domain totals of CATCH with three scenarios for SAMPLE2 of $n = 20$ vessels under strategy SRSWOR_HT.

Domain d	**Variable**	**n** n_d	**Sum of Weights**	**Total** $\hat{t}_d$	**Std Dev** $s.e(\hat{t}_d)$	**95% CL**		**Coeff of Var** $cv(\hat{t}_d)$
Scenario 1. Conditional approach, known N_d								
0	CATCH	12	60	419536	48298	313232.887	525838.924	0.115122
1	CATCH	8	40	191067	17114	150598.627	231536.333	0.089572
Scenario 2. Conditional approach, unknown N_d								
0	CATCH	12	60	419536	53060	302752.639	536319.172	0.126472
1	CATCH	8	40	191067	19985	143810.041	238324.919	0.104597
Scenario 3. Unconditional approach								
0	CATCH	12	60	419536	84344	243002.412	596069.399	0.201041
1	CATCH	8	40	191067	50990	84343.946	297791.014	0.266870

In Scenario 1, even if a single sample has actually been drawn from the entire population, the domains are treated as if a separate sample would have been drawn from each sub-population i.e. domain. The domain sample sizes are regarded fixed and domain sizes in population were assumed known, as would be the case for planned domains or strata in stratified sampling. In Scenario 2, estimates were computed for the case where population domain sizes were unknown, the situation often encountered in domain estimation practice. Obviously, precision is weaker relative to Scenario 1, but not substantially. In Scenario 3, domains were treated as unplanned and the sample distribution over domains was taken uncontrolled. The observed sample sizes in domains are thus regarded as random variates suggesting the unconditional approach for variance estimation. The unconditional approach was implemented by using the extended domain variables technique. This approach was the most conservative.

The approach for variance estimation affects the precision of estimates. Standard errors and coefficients of variation are larger when accounting for the randomness of the domain sample sizes by the unconditional approach, when compared to the conditional approach. The HT estimates for domain totals are additive: their sum over the domains equals the HT estimate of the total estimated for the entire population. This property is often appreciated in official statistics. HT estimator does not involve auxiliary information. If domain totals are known, a Hajék type estimator $\hat{t}_{dHA} = \frac{N_d}{\hat{N}_d} \hat{t}_{dHT}$ that uses N_d as auxiliary information is often used as an alternative (see Section 4.2.4). However, Hajék type estimators are not additive in general but in special cases only (Hidiroglou & Patak 2004, Lehtonen & Veijanen 2009 p. 241).

Calibration and ratio and regression estimation (Chapter 4) are able to incorporate a variety of auxiliary variables and may improve precision over HT and Hajék estimation. Lehtonen & Veijanen (2009) provides a review of calibration and generalized regression estimation methods for the estimation of totals for planned and unplanned domains, including small domains (with small domain sample size). Variance estimators also are provided.

3.3.6 Guidelines

In fisheries statistics, simple random sampling may be selected as the ultimate sampling technique of elements in situations where useful auxiliary data are not available. SRS is often used for element sampling from the strata in stratified sampling designs. The most common SRS technique for practical purposes is simple random sampling without replacement (SRSWOR). SRSWOR is a natural choice, because the unfeasible occasion to draw the same unit two or more times in the sample is excluded.

Efficiency of estimation for simple random samples can be improved in the estimation phase. If aggregate-level data are available on an auxiliary variable that correlates with the target variable, ratio or regression estimation (Chapter 4) may be possible, assuming that unit-level measurements on the same auxiliary variable are available in the sample data set.

3.4 Systematic sampling

3.4.1 Background

Systematic sampling (SYS) is another equal probability sampling design where the inclusion probability is a constant for every population element, similarly as in simple random sampling. In a population $U = \{1, \dots, k, \dots, N\}$ of N units, the probability of inclusion in a n element systematic sample is $\pi_k = \pi = n/N$ for population element $k \in U$.

Auxiliary information does not play a role in standard application of systematic sampling. In *implicit stratification*, auxiliary information is sometimes used before sample selection. In this method, the population frame is sorted by one or several auxiliary variables that are assumed to correlate with the target variable. It should be noted that sorting of the population before systematic sampling can be harmful for the representativeness of the sample, if the sampling interval happens to coincide with a harmonic or periodic variation in the ordered population. Then, substantial parts of the population may not be represented in a systematic sample.

3.4.2 Sample selection techniques

Systematic sampling is a without-replacement type sampling technique. For a systematic sample with one random start, the sampling interval $q = N/n$ is set first. Assuming an integer q, each q^{th} element is selected in the sample. The first element is selected randomly from the q first frame elements or by taking a random integer from the interval $[1, N]$ for the first element and selecting the further elements in a closed loop over the entire frame with steps of length q. Statistical software use SYS algorithms with fractional intervals to provide exactly the specified sample size n.

For a fixed sorting order of the population, the number of different systematic samples with one random start is q i.e. the sampling interval. The selection probability is $p(s) = 1/q$ for sample s and the inclusion probability of element k is $\pi_k = 1/q = n/N$.

3.4.3 Estimation of parameters

In the estimation of population total and mean, formulas for SRSWOR can be used. There is no analytic estimator available for the design variance of an estimator of a total under systematic sampling. Therefore, approximations for the design variance are used. Assuming that units in the sampling frame are in random order relative to the variation of the target variable, the efficiency of systematic sampling is similar to the efficiency of simple random sampling without replacement. In this case, variance estimators of SRSWOR are often used in practice.

3.4.4 Worked example

Preliminaries. We continue working with the population of active vessels in SIMPOP. Our aim is to estimate the CATCH total and associated quality indicators under systematic sampling in the case where no auxiliary data are used. The estimation strategy is SYS_HT, where the sample is drawn by systematic sampling and the estimation relies on a Horvitz-Thompson estimator. Because inclusion probabilities in SYS are equal for all population elements, the weights for HT estimation are constants. The strategy SRSWOR_HT acts as the reference strategy

Estimation. As stated in 3.4.3, the estimation of population total proceeds as for the SRSWOR_HT strategy. Because no analytic estimator for the SYS design variance exists, the SRSWOR variance estimator (10) is often recommended for situations where the sorting order of the frame population is assumed independent on the variation of the target variable.

Simulation experiment. We want to examine whether it is justified to apply the SRSWOR variance and standard error formulas for SYS samples, when the SIMPOP population is sorted into random order. Because we need separate samples drawn with both methods, the assessment must be based on a simulation experiment. We proceeded as follows.

- Scenario 1: A random variate RANDOM was generated from uniform (0,1) distribution and assigned to SIMPOP. Then, SIMPOP was sorted by RANDOM, and $K = 100$ SYS and SRSWOR samples of $n = 5$ and $n = 20$ elements were drawn from the sorted population.
- Scenario 2: SIMPOP was sorted by GT, and $K = 100$ SYS samples of size 5 and 20 elements were drawn from the sorted population.
- For both scenarios, CATCH total, standard error and coefficient of variation were computed by SRSWOR formulas for each sample and estimates were averaged over the simulations.

Estimation results are in Table 3.12.

Table 3.12 Means of estimated totals, standard errors and coefficients of variation for CATCH from $K = 100$ simulated SYS and SRSWOR samples of sizes $n = 5$ and $n = 20$ from SIMPOP.

Obs	VarName	Replicates	Averages over simulations				
			SumWgt	n	Total $\hat{t}$	StdDev $s.e(\hat{t})$	CV $cv(\hat{t})$
Sample size n = 5							
Scenario 1: Population in random order							
SYS_HT	CATCH	100	100.000000	5	608791	84671	0.137494
SRSWOR_HT	CATCH	100	100.000000	5	633850	88704	0.139704
Scenario 2: Population sorted by GT							
SYS_HT	CATCH	100	100.000000	5	627867	96021	0.152833
Sample size n = 20							
Scenario 1: Population in random order							
SYS_HT	CATCH	100	100.000000	20	624636	43692	0.069728
SRSWOR_HT	CATCH	100	100.000000	20	631258	43574	0.068968
Scenario 2: Population sorted by GT							
SYS_HT	CATCH	100	100.000000	20	624281	44522	0.071634

Average coefficients of variation of SRSWOR estimates for the SRSWOR and SYS samples are quite close for both sample sizes. SRSWOR variance formula seems appropriate when the frame units are in random order. But for Scenario 2, where the sorting order of the population elements and the target variable are associated, the

SRSWOR variance estimator tends to produce somewhat larger coefficients of variation than in Scenario 1, for both sample sizes.

The results suggest warning against blind use of the SRS variance formulas for systematic samples. In uncertain situations it is advisable to examine the relation of the population sorting and the target variable or use alternative variance estimators.

3.4.5 Guidelines

In fisheries statistics, systematic sampling can be used instead of simple random sampling when appropriate. For example, systematic sampling is sometimes used in element sampling from frames that are first sorted by regional or related variables in order to have good geographical representation in the sample. If sorting is used, attention must be paid to the sorting order of elements in the population frame to avoid possible problems due to unfeasible sorting order.

In variance estimation, methods of simple random sampling can be used for samples drawn from randomly ordered sampling frames. In implicit stratification, the standard machinery of stratified sampling (see. Sect. 3.6) can be used. Other options are for example pseudo replication methods (jackknife, bootstrap) or the selection of replicated systematic samples (e.g. Lehtonen & Pahkinen 2004, Wolter 2007).

3.5 Sampling with probability proportional to size

3.5.1 Background

Sampling with probability proportional to size (PPS sampling) is an unequal probability sampling method, which is often used for random sampling in business statistics and elsewhere, where the sizes of sampling units vary significantly. If the values of size variable and target variable are closely related, the design variance of the estimator of total can be expected to be smaller than in equal probability designs.

In a population of N units, the probability of inclusion in a n element PPS sample is $\pi_k = n \times z_k/t_z$, where z_k is the value of the size variable Z for element $k \in U$, t_z/z_k is the relative size of element k and $t_z = \sum_{k=1}^{N} z_k$ is the known population total of the size variable. Sampling weights are given by $w_k = 1/\pi_k$. The sizes z_k are assumed known for each element k of the frame. The size variable should be chosen so that its variation resembles the variation of the variable of interest Y. The more the ratio y_k/z_k remains constant across the population, the more efficient the PPS sampling will be.

The inclusion probabilities should meet the requirement $\pi_k \leq 1$ for all k. When the size measure z_k is exceptionally large for one or several elements, it can happen that the inclusion probabilities become greater than one for those elements, that is, $nz_k/t_z > 1$. This situation can be met when working with skewed populations, e.g. in business surveys. In practice, separate strata called *certainty strata* are formed from these elements, and their inclusion probabilities are set $\pi_k = 1$ (i.e. they are drawn with certainty; see Lehtonen & Pahkinen 2004, p. 53).

3.5.2 Sample selection techniques

Similarly as in simple random sampling, a PPS sample can be drawn with replacement or without replacement. Computation of the inclusion probabilities is easier to manage under with-replacement type sampling, because the population remains unchanged after each draw. In without-replacement type PPS sampling, the population changes after each draw and the inclusion probabilities must be re-calculated for the remaining elements. The without-replacement type PPS complicates the estimation of design variances, because the joint (second-order) inclusion probabilities π_{kl} for the inclusion of both elements k and l in the sample are required. An exception is Poisson sampling, where $\pi_{kl} = \pi_k \times \pi_l$, which simplifies computation. This property also holds for PPS sampling with replacement.

Various versions of PPS sampling have been proposed in the literature and are available in computer software such as SAS, SPSS and R. Examples are *PPSWR* and *PPSWOR with cumulative total method* with replacement or without replacement, *systematic PPS sampling* and *Poisson sampling*. A popular method in fixed-size without replacement type PPS sampling is the Hanurav-Vijayan method (Hanurav 1967, Vijayan 1968).

In Poisson sampling, the inclusion probabilities $\pi_k = n \times z_k / t_z$ are first calculated for each population element. Independent random numbers $\varepsilon_k, k = 1, \ldots, N$ are then drawn from uniform (0,1) distribution and attached to elements k in the population. An element k is selected to the sample if $\varepsilon_k < \pi_k$. Similarly as in Bernoulli sampling (Section 3.3.2), the size of the resulting sample is random with expected value $E(n_s) = \sum_{k=1}^{N} \pi_k$. In *conditional Poisson sampling*, only samples of size n are accepted.

In fact, most basic sampling techniques are special cases of PPS sampling. For example, by setting the PPS size variable values $z_k = 1$ for all population elements in PPS_WOR sampling, estimates corresponding to SRSWOR sampling would be obtained.

Chaudhuri & Vos (1988) presents a unified approach for unequal probability sampling and a selection of various variants of PPS sampling schemes. Tillé (2006) introduces an inventory of new methods and algorithms for unequal probability sampling.

3.5.3 Estimation of parameters

Under PPSWOR, Horvitz-Thompson (HT) estimator of population total $t = \sum_{k=1}^{N} y_k$ of target variable *Y* is of the form (2) given by

$$\hat{t}_{HT} = \sum_{k=1}^{n} w_k y_k = \sum_{k=1}^{n} y_k / \pi_k,$$

where $w_k = 1/\pi_k$ are PPS sampling weights. Because inclusion probabilities are determined by element-specific size variable values, the weights can vary between sample elements.

A textbook variance estimator of $\hat{t}_{HT}$ is:

$$\hat{v}_{PPSWOR}(\hat{t}_{HT}) = \sum_{k=1}^{n} \sum_{l=1}^{n} (w_k w_l - w_{kl}) y_k y_l, \quad (11)$$

where $w_{kl} = 1/\pi_{kl}$. An alternative Sen-Yates-Grundy estimator for fixed-size samples is:

$$\hat{v}_{PPSWOR2}(\hat{t}_{HT}) = \sum_{k=1}^{n} \sum_{\substack{l=1 \\ l<k}}^{n} \left(\frac{w_{kl}}{w_k w_l} - 1 \right) (w_k y_k - w_l y_l)^2, \quad (12)$$

which is often preferred in practice. Sampling programs of standard software (SAS, R) are able to compute at least approximate joint inclusion probabilities for certain without-replacement type PPS sampling designs for not-too-large fixed-size samples. Examples are the basic PPSWOR and the PPS sampling methods of Sampford, Midzuno-Sen and Tillé. As an alternative, with-replacement type approximations for PPS variance estimation are implemented in some analysis programs. A somewhat conservative variance estimator is:

$$\hat{v}_{PPSWR}(\hat{t}_{HT}) = \frac{n(1-f)}{n-1} \sum_{k=1}^{n} (w_k y_k - \hat{t}_{HT}/n)^2, \quad (13)$$

where $f = n/N$ is sampling fraction. Estimator (13) assumes with-replacement PPS sampling (Lehtonen and Veijanen 2009 p. 227). For example, the SAS procedure SURVEYMEANS uses this variance estimator.

3.5.4 Worked example

Preliminaries. Sampling with probability proportional to size, i.e. PPS sampling, represents a traditional technique for sampling of elements whose sizes vary in some sense. Examples are samples of schools, establishments, regional areas and why not fishing vessels. In the sampling frame of PPS sampling, a continuous (or count) type auxiliary variable is required, which measures the size of population element, such as vessel tonnage, engine power etc. If the relation of target variable Y and size measure Z is strong, then PPS sampling may improve efficiency. We discuss PPS for element-level sampling designs.

We continue working with the population of active vessels. Our target variable is CATCH. We assume that we have access to a single auxiliary variable, for example GT (vessel tonnage) whose values are available for all population vessels in the sampling frame. The variable GT is promising: by Table 3.4, corr(CATCH,GT) = 0.56. GT will serve as the size variable in our PPSWOR sampling exercise. In addition, we will demonstrate PPS sampling with another, more powerful, auxiliary variable as the size variable and examine its effect to precision.

We adopt the estimation strategy PPSWOR_HT, where the sample is drawn from SIMPOP by PPSWOR and estimation relies on a Horvitz-Thompson (HT) estimator. We demonstrate the effect of the sample size *n* to

variance, standard error and coefficient of variation estimates of the estimated total of CATCH. We compare the results with our reference strategy SRSWOR_HT by computing the design effect estimate.

Sample selection. Our first sample size drawn by SURVEYSELECT is $n = 5$ active vessels for a 5% sample from SIMPOP. The realized sample SAMPLE3 is listed in Table 3.13. In addition to the variables ID, CATCH and weight variable SAMPLINGWEIGHT, the values of size variable GT are included. In PPS sampling, weights are inverses of the probabilities to be selected in the sample and are vessel specific, values depending on the value of GT. Therefore, the values of weights vary. PPS thus is an *unequal probability sampling technique*. Large vessels (measured in GT) get smaller weights than smaller vessels. In other words, probability of selection is larger for large vessels and smaller for small vessels. Note also that the sum of weights differ from the population size ($N = 100$). Sum of weights is sample specific and depends on the goodness of fit of the underlying implicit model $y_k = \beta x_k + \varepsilon_k$. If the model is approximately correct then the sum of weights will be close to N.

Table 3.13 PPSWOR sample SAMPLE3 of $n = 5$ active vessels drawn from SIMPOP of $N = 100$ vessels.

Obs k	**ID**	**CATCH** y_k	**GT** z_k	**SamplingWeight** w_k
1	65	3799.95	329.0	19.9978
2	89	6845.81	343.2	19.1704
3	27	6087.56	345.1	19.0649
4	53	7601.87	376.2	17.4888
5	94	10615.99	436.8	15.0625
Sum				**90.7843**

Estimation. Let us compute the estimates for CATCH total from SAMPLE3 by using the computational formulas of Section 3.5.2 and 3.5.3. By inserting the PPSWOR weights w_k and sample values of CATCH from Table 3.13 into the HT estimator (2) we obtain:

$$\hat{t}_{HT} = \sum_{k=1}^{5} w_k y_k = 616136.$$

We estimate the standard error $s.e(\hat{t}_{HT})$ by the square root of an approximate variance estimator (13). The estimator is based on the with replacement (WR) assumption and is often used in practice (e.g. SAS procedure SURVEYMEANS). This is a conservative estimator, because WR sampling tends to be less effective than WOR sampling.

By inserting the values of weights, CATCH and GT from Table 3.13 into the variance estimator (13) we get:

$$\hat{v}_{PPSWR}(\hat{t}_{HT}) = \frac{5\times(1-5/100)}{5-1}\sum_{k=1}^{n}(w_k y_k - 616136/5)^2 \; = 67055^2,$$

where $s.e(\hat{t}_{HT}) = \; 67055$. A two-sided 95% confidence interval for the estimated total is computed similarly as in Section 3.3.4:

Lower confidence limit: $LCL(\hat{t}_{HT}) = 429961$

Upper confidence limit: $UCL(\hat{t}_{HT}) = 802310$.

The confidence interval is much narrower than for the SRSWOR case. Coefficient of variation (5) for $\hat{t}_{HT}$ is calculated as:

$$cv(\hat{t}_{HT}) = \frac{s.e(\hat{t}_{HT})}{\hat{t}_{HT}} = 0.11.$$

Finally, we compute the design effect estimate (7) of $\hat{t}_{HT}$. SURVEYMEANS does not give deff estimates by default and we compute it separately.

$$deff(\hat{t}_{HT}) = \frac{\hat{v}_{PPSWR}(\hat{t}_{HT})}{\hat{v}_{SRSWOR}(\hat{t}_{HT})} = \frac{67055^2}{107952^2} = 0.39,$$

where $\hat{t}_{HT}$ variance estimate under the actual PPSWOR design for SAMPLE3 is in the numerator and the SRSWOR variance estimate for SAMPLE3 is in the denominator (note that numerical values of $\hat{t}_{HT}$ would be unequal because of different weighting). The PPSWOR_HT strategy clearly is more efficient than would be the SRSWOR_HT strategy for SAMPLE3.

Estimates are also computed by SURVEYMEANS and are displayed in Table 3.14.

Table 3.14 Estimated total, standard error and coefficient of variation for variable CATCH from PPSWOR sample SAMPLE3 of $n = 5$ vessels.

Variable	True value	n	Sum of Weights	Total $\hat{t}$	Std Dev $s.e(\hat{t})$	95% CL		Coeff of Var $cv(\hat{t})$
CATCH	624036	5	90.784295	616136	67055	429961.882	802310.945	0.108831

We next draw a larger sample of size $n = 20$ vessels. Estimates computed by SURVEYMEANS are in Table 3.15.

Table 3.15 Estimated total, standard error and coefficient of variation for variable CATCH from PPSWOR sample SAMPLE4 of $n = 20$ vessels.

Variable	True value	n	Sum of Weights	Total $\hat{t}$	Std Dev $s.e(\hat{t})$	95% CL		Coeff of Var $cv(\hat{t})$
CATCH	624036	20	99.434022	664942	40175	580855.134	749027.890	0.060418

Estimated standard error for total estimate from PPSWOR SAMPLE4 of $n = 20$ vessels is smaller than from SAMPLE3 $n = 5$ vessels. Coefficient of variation is smaller and confidence interval is narrower than for SAMPLE3. Estimation results with a larger PPS sample size are more reliable than for a smaller PPS sample.

Comparing with a SRSWOR sample of same size $n = 20$, design effect is calculated as:

$$deff(\hat{t}_{HT}) = \frac{\hat{v}_{PPSWR}(\hat{t}_{HT})}{\hat{v}_{SRSWOR}(\hat{t}_{HT})} = \frac{40175^2}{49158^2} = 0.67.$$

It can be shown that that SRSWOR can be considered as a special case of PPSWOR sampling. For example, by setting the size variable values $z_k = 1$ for all population elements in PPSWOR sampling, we would obtain numerically close estimation results as with SRSWOR

Simulation experiment. Let us demonstrate numerically some theoretical properties of PPSWOR sampling by drawing several PPSWOR samples from SIMPOP with GT as size variable and by examining the distribution of the estimated totals, se:s and cv:s. We draw $K = 100$ PPSWOR samples of small size $n = 5$ vessels and then with larger size $n = 20$ vessels from SIMPOP, compute the estimated total, s.e and cv from each sample. Finally, we compute the means of the statistics for the 100 samples.

Summary results are in Table 3.16. For comparison, we also show estimation results under PPSWOR sampling with GT_DAS as size variable. By Table 3.6, CATCH and GT_DAS are more strongly correlated than CATCH and GT: corr(CATCH,GT_DAS) = 0.84. We also include the results for SRSWOR sampling from Section 3.3.4.

Table 3.16 Means of estimated totals, standard errors and coefficients of variation for CATCH from $K = 100$ simulated SRSWOR and PPSWOR samples of sizes $n = 5$ and $n = 20$ vessels from SIMPOP.

Method	VarName	AuxVar	Replicates	Averages over simulations				
				SumWgt	n	Total $\hat{t}$	StdDev $s.e(\hat{t})$	CV $cv(\hat{t})$
Sample size *n* = 5								
1. SRSWOR	CATCH	none	100	100.000000	5	629966	91436	0.145160
2. PPSWOR	CATCH	GT	100	98.901065	5	619855	77979	0.127500
3. PPSWOR	CATCH	GT_DAS	100	98.622793	5	631954	45627	0.073189
Sample size *n* = 20								
4. SRSWOR	CATCH	none	100	100.000000	20	626895	44061	0.070264
5. PPSWOR	CATCH	GT	100	100.168244	20	625331	36331	0.058307
6. PPSWOR	CATCH	GT_DAS	100	100.044228	20	624245	22487	0.036093
True total	CATCH					624036		

Both SRSWOR and PPSWOR produce estimated totals that on average are close to the true total, for both sample sizes, confirming the design unbiasedness property for HT estimator under the equal-probability SRSWOR design and the unequal probability PPS sampling design. For both methods, the average standard error and coefficient of variation figures decline when sample size increases, as expected. For both sample sizes, coefficients of variation for PPSWOR samples are smaller than cv:s for SRSWOR, so PPS sampling clearly improves accuracy.

Estimation results under PPS with GT_DAS as size variable are most striking. When comparing standard errors or cv:s in Table 3.16 rows 3 and 4 we note the following. For PPS sampling with sample size $n = 5$ and GT_DAS as size variable, the same precision is obtained as under SRSWOR with sample size $n = 20$ i.e. four times larger sample size. It appears very cost effective to use PPS sampling in this case.

Statistical properties (design bias and accuracy) of PPS sampling and SRSWOR can be examined further by displaying the distributions of the estimates for both methods. Graphs under sample size *n* = 5 vessels are in Figure 3.4. For proper distributions we use *K* = 1000 simulated samples. A near symmetry of the distribution around the mean is beneficial for inference purposes. The mean of the estimates approximates the design expectation of the distribution. Design unbiasedness is attained if the mean is close to the true total. The variation of the estimates around the mean shows the precision behaviour of the strategy: the more condensed around the mean, the more effective strategy. It is clearly seen that both distributions are close to symmetry. Both methods indicate design unbiasedness, as expected. The variation of estimates is smaller for PPSWOR than for SRSWOR.

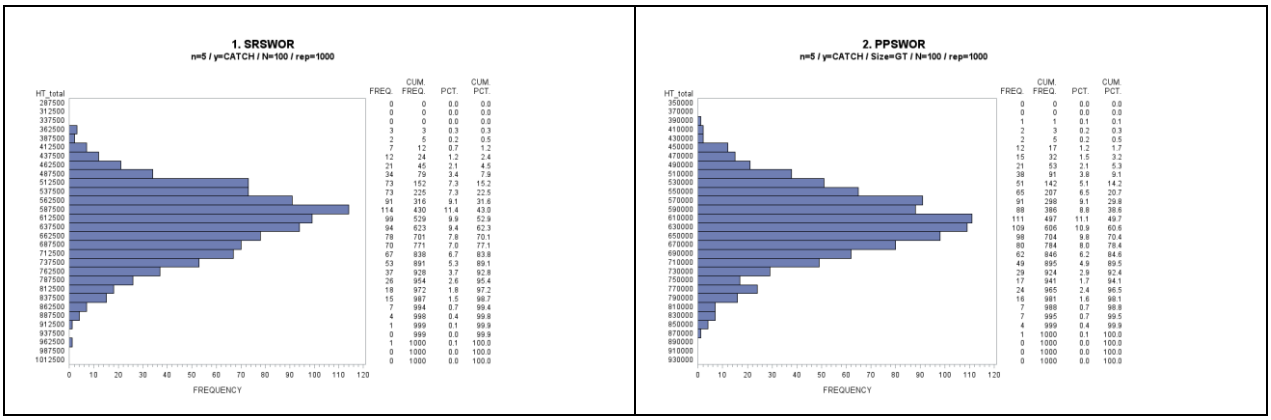
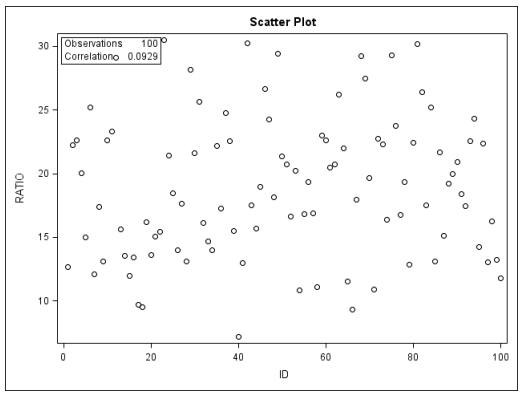
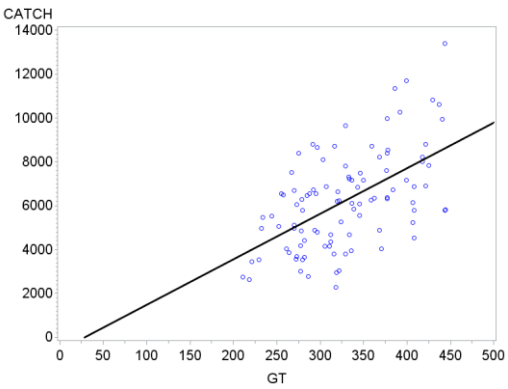


Figure 3.4. Distributions of total estimates from $K = 1000$ SRSWOR and PPSWOR samples.

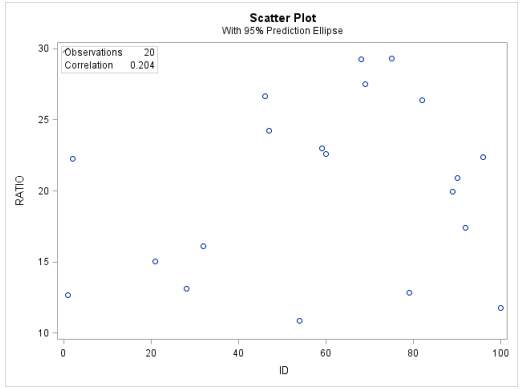
Examination of assumptions on PPS sampling. High correlation of target variable and size variable is good for PPS sampling to be effective. The correlation of CATCH and GT is reasonably high (0.56) in the population and for SAMPLE4, the correlation is 0.61. But high correlation alone is not enough for proper behaviour of PPSWOR. In addition, the ratio of CATCH and GT should be nearly constant over the population.



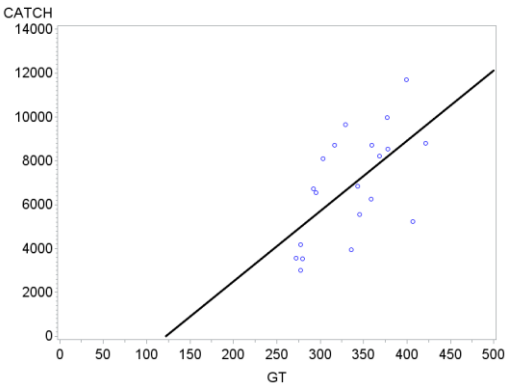
A: Scatter plot of ID with RATIO in SIMPOP



B: Scatter plot for CATCH with GT in SIMPOP



C: Scatter plot of ID with RATIO in SAMPLE4



D: Scatter plot for CATCH with GT in SAMPLE4

Figure 3.5 Scatter plots of RATIO with ID and CATCH with GT in SIMPOP and SAMPLE4.

The fitted regression line for the population (Panel B) goes close to the origin. These properties are favourable for good performance of PPS sampling. For SAMPLE4 of size $n = 20$, the situation in Panels C and D seems to be adequate enough for good performance of PPSWOR sampling in this exercise.

3.5.5 Guidelines

In PPS sampling, the auxiliary information is introduced in the sampling design. Values of the size variable must be available for all vessels in the sampling frame. It is important for good efficiency of estimation under PPS sampling to choose the size variable so that its variation resembles the variation of the target variable of interest.

PPS sampling is often used in descriptive surveys, where the focus is in a single important target variable or a few mutually correlated target variables, and reliable estimation is required for just these variables. If a powerful size variable is available in the sampling frame, the strategy can be optimized for efficient estimation, and PPS can be a reasonable choice for good cost-efficiency.

However, situations can be met in practice where a PPS sampling design is even worse in precision than SRSWOR, if the assumptions underlying PPS sampling are not met. It is necessary to examine the assumptions in each specific sampling situation, e.g. based on data from possible previous surveys.

Situations can occur in practice where the set of target variables consists of several diverge variables. A PPS sampling design cannot be optimized for all these variables, because a single size variable only can be introduced. We study in Chapter 4 how the precision can be improved by using model-assisted methods in the estimation phase under a simple sampling design.

3.6 Stratified sampling

3.6.1 Background

In *stratified sampling* (STR sampling), the population is divided into non-overlapping subpopulations by using one or several categorical *stratification variables*, whose values must be available for all population elements in the sampling frame. The subpopulations are called *strata* and they can be treated as separate populations in the sampling and estimation phases. Regional, demographic, socioeconomic, or other appropriate auxiliary information can be utilized in the stratification of the population elements, but strata can also be inherent in the data. For example, administrative areas can be used to guarantee exhaustive presentation of an entire country. Efficiency of estimation can improve relative to SRS sampling, if the strata are internally homogeneous with respect to the target variable.

3.6.2 Allocation and sample selection

Several sample allocation procedures have been proposed for the STR sampling in the literature, for example Lehtonen & Pahkinen (2004) and Lohr (2009). Some commonly used procedures are described briefly.

Proportional allocation is a reasonable and popular starting point as only the stratum sizes N_h are assumed to be known. The sampling fraction $n_h/N_h = n/N$ is constant for each stratum *h*, so that the share of the sample for each stratum is

$$n_{h,prop} = \frac{N_h}{N} \times n = W_h \times n,$$

where $W_h = \frac{N_h}{N}$ is stratum weight and *n* is the overall sample size. Sampling fraction n/N and therefore the inclusion probabilities are constant. Proportional allocation thus leads to an equal probability sampling design. This simplifies estimation, but can lead to minor improvements in statistical efficiency when compared with more advanced allocation schemes, if stratum variances in population vary greatly. If the strata are internally homogeneous with respect to the target variable, proportional allocation can improve precision relative to SRS sampling.

Neyman (optimal) allocation utilizes the stratum standard deviations of the variable of interest. Optimal allocation of sample elements to stratum h would then be

$$n_{h,Neyman} = n \times \frac{N_h S_h}{\sum_{h=1}^{H} N_h S_h},$$

where S_h is population standard deviation in stratum h. Standard deviations are usually unknown, but they can be approximated from earlier studies or other reliable source. Optimal allocation provides the most efficient allocation scheme for stratified sampling. This method is often used in repeated business surveys. The allocation

formula shows that more units are allocated to a large and internally heterogeneous stratum than for a small and homogeneous stratum. Neyman allocation is often used in optimizing the costs, if the unit costs of sampling vary in strata and the costs can be approximated.

Power allocation can be used if there are several small strata and precise estimates at all stratum levels are required. In addition to an approximation of the population coefficient of variation of the target variable, a known stratum-wise population total of an auxiliary variable can be introduced in the allocation procedure.

In **equal allocation**, the same number of elements is drawn from each stratum so that $\sum_{h=1}^{H} n_h = n$, the overall sample size. If the strata sizes N_h are unequal then equal allocation produces an unequal sampling design. Equal allocation is sometimes used in surveys to obtain a desired precision also for strata whose sizes are small. No auxiliary information is needed except the stratum sizes N_h.

Multivariate allocation methods have been proposed in the literature for the optimization of sample sizes for the population subgroups of interest (strata or domains) to attain a pre-specified precision of the estimates in multi-purpose sample surveys. In such surveys there are often a number of diverse target variables with a different variance $V_j, j = 1, \dots, J$. A precision constraint is first set to each of the variables and an allocation producing the minimal costs is then selected from the allocations that meet the constraints. Popular methods are the ones published by Bethel (1989) and Chromy (1987). Both are iterative methods and the Chromy method is often preferred in cases where the number of strata is large, because the convergence is expected to be faster. One of the computerized tools is the MAUSS-R software (Buglielli et al. 2013), which is used for the production of fisheries statistics in Italy (Section 8.1). Further, Benedetti et al. (2008) proposed an approach that combines stratification and sample allocation including the choice of stratifying variables, the number of class intervals for each variable, and the optimal Bethel allocation of the sample into the strata. More sophisticated methods are needed for skewed populations that are often encountered in environmental and business surveys, see e.g. Benedetti et al. (2010).

3.6.3 Estimation of parameters

In the estimation phase, the individual strata are considered as independent subpopulations. Stratum-wise parameters are estimated by using appropriate sampling weights and summed over the strata for estimates on the overall population parameters. A HT estimator for population total $t = \sum_{h=1}^{H} \sum_{k=1}^{N_h} y_{hk}$ thus is:

$$\hat{t}_{HT} = \sum_{h=1}^{H} \hat{t}_h = \sum_{h=1}^{H} \sum_{k=1}^{n_h} w_{hk} y_{hk}, \tag{14}$$

where $\hat{t}_h = \sum_{k=1}^{n_h} w_{hk} y_{hk}$ is an estimator of the total t_h of stratum h and $w_{hk} = 1/\pi_{hk}$ is the sampling weight for element k in stratum h, derived for the entire sample such that $\sum_{h=1}^{H} \sum_{k=1}^{n_h} w_{hk} = N$. For SRS sampling, for example, the total is estimated by

$$\hat{t}_{STR} = \sum_{h=1}^{H} \hat{t}_h = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{k=1}^{n_h} y_{hk} = \sum_{h=1}^{H} N_h \hat{\bar{y}}_h, \tag{15}$$

as the weights are $w_{hk} = \frac{N_h}{n_h}, h = 1, \dots, H$.

Due to the independence assumption, the variance estimator of the overall estimator $\hat{t}_{HT}$ is simply the sum of stratum variance estimators, given by:

$$\hat{v}_{STR}(\hat{t}_{HT}) = \sum_{h=1}^{H} \hat{v}(\hat{t}_h), \tag{16}$$

where $\hat{v}(\hat{t}_h)$ is variance estimator for $\hat{t}_h$ in stratum h. The stratum variance estimators depend on the element sampling technique and the type of the estimator of the total in stratum h. The variance formula indicates that variance estimate becomes small and estimation is efficient, if stratum samples are internally homogeneous. Allocation also affects the variance of the overall estimators since the stratum size has an effect on the stratum variance.

3.6.4 Worked example

Preliminaries. In stratified sampling (STR), the population elements are first grouped into non-overlapping strata by using a single or multiple auxiliary variables as the stratification variables. Typical stratification variables are regional and economic variables and variables describing the properties of population elements, such as type

of industry, turnover and staff size in business surveys. Stratification variables must be available in the sampling frame. Stratification is followed by sample allocation into strata using one of the allocation techniques. For the selection of a random sample from each stratum, one of the basic sampling techniques, simple random sampling, systematic sampling or PPS sampling, is used, the choice of technique depending on the type of the population element and the availability of auxiliary data in the sampling frame.

There are various objectives for stratified sampling in fisheries surveys. By stratification combined with an appropriate allocation scheme, it is possible to determine the number of vessels, to be drawn from each stratum so that all important parts of the vessel population will be properly represented in the sample. A sufficiently large sample size can be allocated for the strata that are of special interest and for rare subgroups to obtain precise enough estimation for these subgroups. Indeed, in surveys it is common to define the main subgroups of the population as strata for which accurate estimates are required.

We continue working with the population of active vessels. Stratification is made by the variable STR3. STR3 was created by dividing the variable GT (vessel tonnage) into three nearly equal-sized classes coded 1, 2 and 3. In Section 3.5.4 we used GT as a continuous size variable in PPS sampling. Now we use the same variable as a categorical variable for stratification purposes. This gives for us an option to compare the accuracy performance of PPS sampling and stratified sampling where basically, the same auxiliary information is used.

We use estimation strategies STR_SRSWOR_HT and STR_PPSWOR_HT, where the population is first stratified into three strata by the variable STR3. Then we fix the total sample size n and select the allocation scheme. We apply proportional allocation, where the stratum sample sizes are proportional to the stratum sizes in the population. This produces an equal probability sampling design.

Estimation in both strategies relies on a Horvitz-Thompson (HT) estimator. We demonstrate the effect of the sampling design within strata to standard error and coefficient of variation estimates of the estimated total of CATCH. We compare the results with our reference strategy SRSWOR_HT by computing the design effect estimates.

Sample selection. By using PROC SURVEYSELECT, we draw from SIMPOP the following stratified samples of $n = 20$ active vessels with STR3 as the stratification variable:

(a) SAMPLE5 by stratified SRSWOR
(b) SAMPLE6 by stratified PPSWOR with GT_DAS as the size variable.

We use proportional allocation for both cases. The realized stratified samples SAMPLE5 and SAMPLE6 are listed in Table 3.17.

Table 3.17 Stratified SRSWOR and PPSWOR samples SAMPLE5 and SAMPLE6 of n = 20 active vessels drawn from SIMPOP.

(a) Stratified SRSWOR sample

Obs k	ID	STR3	CATCH	Selection Prob	Sampling Weight
1	1	1	3541.44	0.18182	5.500
2	22	1	3538.14	0.18182	5.500
3	23	1	8402.75	0.18182	5.500
4	25	1	4978.66	0.18182	5.500
5	42	1	8811.94	0.18182	5.500
6	44	1	4421.92	0.18182	5.500
7	12	2	8644.48	0.21212	4.714
8	15	2	3786.06	0.21212	4.714
9	30	2	7208.94	0.21212	4.714
10	36	2	5855.21	0.21212	4.714
11	46	2	8100.05	0.21212	4.714
12	52	2	4888.34	0.21212	4.714
13	75	2	9652.44	0.21212	4.714
14	55	3	6865.42	0.20588	4.857
15	57	3	6364.10	0.20588	4.857
16	67	3	7160.06	0.20588	4.857
17	82	3	9959.59	0.20588	4.857
18	90	3	8803.08	0.20588	4.857
19	91	3	7823.12	0.20588	4.857
20	94	3	10615.99	0.20588	4.857
					100.000

(b) Stratified PPSWOR sample

Obs k	ID	STR3	CATCH	GT_DAS	Selection Prob	Sampling Weight
1	44	1	4421.92	46546.5	0.17298	5.7812
2	41	1	3651.90	52170.0	0.19387	5.1580
3	35	1	6046.40	53508.0	0.19885	5.0290
4	11	1	5458.75	56862.0	0.21131	4.7324
5	38	1	6288.66	64170.0	0.23847	4.1934
6	37	1	6682.50	67500.0	0.25084	3.9866
7	20	2	4158.35	38150.0	0.14405	6.9422
8	48	2	6107.27	48470.4	0.18301	5.4641
9	80	2	6879.04	61420.0	0.23191	4.3120
10	12	2	8644.48	68607.0	0.25905	3.8603
11	69	2	8709.47	75081.6	0.28349	3.5274
12	56	2	6185.86	78302.0	0.29565	3.3824
13	50	2	7179.00	83476.8	0.31519	3.1727
14	58	3	4519.01	57936.0	0.15776	6.3386
15	79	3	5227.51	68783.0	0.18730	5.3390
16	43	3	6359.22	71451.9	0.19457	5.1396
17	61	3	7173.60	85400.0	0.23255	4.3001
18	57	3	6364.10	87179.4	0.23740	4.2124
19	91	3	7823.12	101598.9	0.27666	3.6145
20	81	3	13391.04	103008.0	0.28050	3.5651
						92.0508

Estimation. In stratified sampling, estimation is carried out separately in each subpopulation or stratum, and estimates for the entire population are computed as sums of the stratum estimates.

CASE (a) STR_SRSWOR_HT. We estimate the total of CATCH by (15) and get:

$$\hat{t}_{HT} = \sum_{h=1}^{3} \sum_{k=1}^{n_h} w_{hk} y_{hk} = 691976,$$

where n_h is the number of sample vessels in stratum h and w_{hk} is the weight for element k in stratum h in Table 3.17 Part (a). For computing variance estimate of $\hat{t}_{HT}$ by (16) we compute the stratum-wise variance estimates using the SRSWOR variance estimator:

$$\hat{v}_{STR}(\hat{t}_h) = N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\hat{s}_h^2}{n_h},\ H = 1, \ldots, h,$$

where $\hat{s}_h^2$ is the sample variance of target variable in stratum *h*. Estimated stratum totals and variances are in Table 3.18.

Table 3.18 Stratum estimates for SAMPLE5.

Stratum *h*	**Variable**	*n*	**Total** $\hat{t}_h$	**Var of Total** $\hat{v}_h$
1	**CATCH**	6	185322	844469990
2	**CATCH**	7	226925	551281589
3	**CATCH**	7	279729	342431431
Sum		20	691976	1738183010

The overall variance estimate (16) for $\hat{t}_{HT}$ is obtained by summing up the stratum-wise variance estimates:

$$\hat{v}_{STR\_SRSWOR}(\hat{t}_{HT}) = \sum_{h=1}^{3} \hat{v}_{STR}(\hat{t}_h) = 1738183010 = 41692^2,$$

and $s.e(\hat{t}_{HT}) = 41692$. Variance formula shows that having internally homogeneous strata is beneficial for improved precision.

Coefficient of variation (5) is calculated as:

$$cv(\hat{t}_{HT}) = \frac{s.e(\hat{t}_{HT})}{\hat{t}_{HT}} = \frac{41692}{691976} = 0.060.$$

Design effect estimate (7) is computed as:

$$deff(\hat{t}_{HT}) = \frac{\hat{v}_{STR\_SRSWOR}(\hat{t}_{HT})}{\hat{v}_{SRSWOR}(\hat{t}_{HT})} = \frac{41692^2}{44300^2} = 0.88.$$

Coefficient of variation is quite small. The deff estimate indicates that stratification improves precision to some extent when compared to estimates obtained by assuming SAMPLE5 as a SRSWOR sample without stratification.

CASE (b) STR_PPSWOR_HT. We estimate the total of CATCH by the same formula as in (a) but with different weights that are taken from Table 3.17 Part (b):

$$\hat{t}_{HT} = \sum_{h=1}^{3} \sum_{k=1}^{n_h} w_{hk} y_{hk} = 576254.$$

Estimated total computed from SAMPLE6 happens to be much smaller than the estimate computed using the stratified SRSWOR sample. For computing variance estimates of stratum totals $\hat{t}_h$ we use the PPSWR variance estimator (13):

$$\hat{v}_{STR}(\hat{t}_h) = \frac{n_h(1-f_h)}{n_h - 1} \sum_{k=1}^{n_h} (w_{hk} y_{hk} - \hat{t}_h / n_h)^2.$$

For the overall variance estimate (16) we obtain:

$$\hat{v}_{STR\_PPSWR}(\hat{t}_{HT}) = \sum_{h=1}^{3} \hat{v}_h = 22282^2,$$

and $s.e(\hat{t}_{HT}) = 22282$. Coefficient of variation (5) is calculated as:

$$cv(\hat{t}_{HT}) = \frac{s.e(\hat{t}_{HT})}{\hat{t}_{HT}} = \frac{22282}{576254} = 0.039.$$

Design effect estimate (7) is computed as:

$$deff(\hat{t}_{HT}) = \frac{\hat{v}_{STR\_PPSWR}(\hat{t}_{HT})}{\hat{v}_{SRSWOR}(\hat{t}_{HT})} = \frac{22282^2}{42255^2} = 0.28,$$

indicating that the strategy STR_PPSWOR_HT is very efficient in this case and would improve estimation substantially when compared to a strategy SRSWOR_HT.

Estimation results for strategies (a) and (b) with PROC SURVEYMEANS are in Table 3.19. It can be observed that stratification and HT estimation in connection with PPSWOR sampling turns out to be substantially more efficient than stratification by HT estimation under SRSWOR, with the same stratification variable STR3 and proportional allocation in both cases.

Table 3.19 Estimated totals, standard errors and coefficients of variation for variable CATCH from (a) stratified SRSWOR sample SAMPLE5 and (b) stratified PPSWOR sample SAMPLE6 of $n = 20$ vessels.

Method	Variable	n	Sum of Weights	Total $\hat{t}$	Std Dev $s.e(\hat{t})$	95% CL		Coeff of Var $cv(\hat{t})$
(a) STR_SRSWOR	**CATCH**	20	100.000000	691976	41692	604014.144	779936.990	0.060250
(b) STR_PPSWOR	**CATCH**	20	92.050773	576254	22282	529243.006	623265.260	0.038667
True total	**CATCH**			624036				

Both stratified samples seem somewhat extreme because the estimated totals are far from the true total. Estimated total is much larger than true value for the STR_SRSWOR_HT strategy and much smaller for the STR_PPSWOR_HT strategy. It would be useful to examine closer the distributions of the HT estimates and their standard errors empirically for example by simulation experiments.

3.6.5 Guidelines

In fishery surveys, stratification of the population before sample selection is recommended for situations where sufficiently large sample sizes are required for the most important subgroups of the population for attaining a desired precision level for the estimates. In these situations, a non-proportional allocation scheme is often chosen, leading to an unequal probability type STR sampling design. Stratified sampling alone does not necessarily improve precision substantially. For improved precision, additional auxiliary information may be introduced in the sampling and estimation designs, such as PPS sampling of elements within the strata or regression estimation for the overall sample or separately in each stratum.

Additional auxiliary information can also be introduced in the allocation scheme, by using optimal (Neyman) allocation, which requires good approximations for the (unknown) stratum standard deviations of the target variable, or power allocation, where good approximations of the stratum-wise CV:s of the target variable are needed, in addition to known stratum totals of an auxiliary variable.

Stratification with SRS, systematic sampling or PPS sampling supplemented with a simple allocation scheme provides often a manageable sampling design for a fisheries survey. When feasible, it is advisable to consider the options for improving accuracy of estimates in the estimation phase. Model-assisted and calibration methods provide flexible methods this purpose.

It is recommended that the availability of suitable stratification variables in different administrative registers and related data sources are examined. If possible, these variables should be included in the sampling frame before sampling operations.

4 Model-assisted estimation and related methods

4.1 Estimation designs

The previous sections were devoted to sampling methods with special emphasis on the use of auxiliary information in the sampling design. In this chapter, we extend the discussion to estimation methods that use auxiliary information in the estimation design. These methods are applied in the analysis of the collected sample data set. There are many good reasons for the use of auxiliary information in the estimation phase. A typical descriptive survey can involve a variety of different target variables of interest. Because it is not possible to optimize the sampling design for all these variables, a compromise sampling design is often implemented. The compromise sampling design with the HT estimation strategy does not necessarily produce precise estimation for all variables of interest. Therefore, an estimation design is often adopted that guarantees the desired precision for population estimates and also for estimates for the important population subgroups.

For example, an equal probability sampling design can be applied for element sampling, possibly amended with stratification and non-proportional allocation. After sample selection, an estimation design is implemented that incorporates aggregate-level or unit-level auxiliary information and statistical modeling. Strategies of this type do not rely on the HT estimation but on more flexible *design-based model-assisted* methods and *calibration estimation*.

The framework of model-assisted methods comprises simple linear fixed-effects models up to more complex generalized linear mixed models, the model choice depending on the given statistical data infrastructure and the complexity of the estimation problem at hand, as well as the preferred statistical framework. This chapter covers the traditional model-assisted methods *ratio estimation*, *regression estimation* and *post-stratification*. Each of these methods involves an explicit model statement, based on a standard linear models framework for continuous target variables.

Examples of particular models underlying the traditional model-assisted estimators for population totals and means are:

a) Regression models of the form $y_k = \beta_0 + \beta_1 x_{1k} + \beta_2 x_{2k} + \cdots + \beta_p x_{pk} + \varepsilon_k$, where the covariates (auxiliary) variables are considered continuous, e.g. vessel tonnage GT, days at sea DAS, etc. These models act as assisting models in ratio and regression estimation. The estimated β-parameters and the auxiliary variables are used in the construction of a ratio or regression estimator.
b) ANOVA (Analysis of variance) models, where the explanatory variables are categorical, e.g. variable STR3 with three classes. These models are typical in post-stratification. Technically, models with categorical explanatory variables can be formulated as regression models if desired, with class membership indicator variables as the explanatory variables.

In the framework of *generalized regression (GREG) estimation,* the entire family of linear and generalized linear models can be applied. For example, linear ANCOVA (Analysis of covariance) models involving both continuous and categorical explanatory variables and their interaction terms can be implemented in a generalized regression estimator, and logistic models for a binary target variable (e.g. ACTIVITY) and a set of continuous and categorical explanatory variables can be incorporated in a logistic GREG estimator.

An important property of model-assisted methods is that estimators for totals discussed here remain (nearly) design unbiased irrespective of the correctness of the assisting model. The model affects efficiency: with a powerful model, precision will decline relative to HT estimation. A thorough presentation of model-assisted methods is in Särndal, Swensson and Wretman (1992).

In model-assisted estimation, the auxiliary data are incorporated in the estimation procedure by models. A *model-free* statistical framework is sometimes preferred leading to *model-free calibration techniques*. The approach was introduced in Deville and Särndal (1992). In model-free calibration (Särndal 2007), an explicit model statement is not required but the auxiliary information is incorporated in the estimation procedure via a weight variable. The methodology is often called *re-weighting*, since the original sampling weights are adjusted appropriately for new calibrated weights. The calibrated weights must satisfy certain conditions. By applying the calibrated weights to an auxiliary variable , the weighted sum of sample observations of the auxiliary variable must coincide with the known population total of the variable. The so-called *calibration equation* states:

$$\sum_{k=1}^{n} w_{CAL,k} x_k = \sum_{k=1}^{N} x_k = t_x, \qquad (17)$$

where $w_{CAL,k}$ is the new calibrated weight and x_k is the value of the auxiliary variable for element k. The known population total t_x of the auxiliary variable is the sole auxiliary information needed for calibration estimation of total for a target variable. The traditional model-assisted methods considered here also fulfil the calibration equation (17) and thus, they can be expressed as calibration estimators.

In calibration estimation, efficiency is expected to improve over HT estimation if the target variable correlates with the auxiliary variable. This can be seen by inspecting a simple variance approximation of a calibration estimator of a total. The calibration estimator for total is $\hat{t}_{CAL} = \sum_{k=1}^{n} w_{CAL,k} y_k$ and a simple estimator of the approximate design variance is

$$\hat{v}(\hat{t}_{CAL}) = \hat{v}(\hat{t}_{HT}) \times (1 - corr_{yx}^2). \quad (18)$$

Obviously, efficiency improves over HT estimation as soon as the correlation $corr_{yx}$ between variables Y and X is nonzero. The same property holds for the model-assisted estimators.

The main aim of Chapter 4 is to target how and to what extent the precision can be improved over the SRSWOR_HT strategy by making use of model-assisted methods as well as calibration methods under a simple random sampling SRSWOR design. These methods offer much flexibility when compared to strategies such as PPSWOR_HT, where a single auxiliary variable is applied. In stratified sampling and PPS sampling, the auxiliary data are needed at the unit level in the sampling frame, whereas in the traditional model-assisted methods, auxiliary data are needed at aggregate level, and unit-level values are only needed for the sample. Moreover, multiple auxiliary variables can be imposed in the assisting model, e.g. a regression model, offering an option to tailor the estimation design separately for each important target variable.

Basic estimation designs, target variable types, auxiliary variable requirements and assisting model types are summarized in Table 4.1. We concentrate in this chapter on the classical model-free calibration method (c) and the traditional model-assisted ratio and regression estimation and post-stratification methods (d), (e) and (f).

Table 4.1 Basic design-based estimation designs.

<table>
<tr><th>Estimation design</th><th>Target variable types</th><th>Auxiliary data requirements</th><th>Assisting models</th></tr>
<tr><td colspan="4">(1) Traditional model-free estimation</td></tr>
<tr><td>(a) Horvitz-Thompson type (HT)</td><td rowspan="3">Continuous, binary, count</td><td>None</td><td rowspan="3">None</td></tr>
<tr><td>(b) Hajék type (HA)</td><td>Population or domain size</td></tr>
<tr><td>(c) Model-free calibration (CAL)</td><td>Aggregate or domain level</td></tr>
<tr><td colspan="4">(2) Traditional model-assisted estimation</td></tr>
<tr><td>(d) Regression estimation (REG)</td><td rowspan="3">Continuous</td><td rowspan="3">Aggregate or domain level</td><td>Linear fixed-effects regression model</td></tr>
<tr><td>(e) Ratio estimation (RAT)</td><td>Linear fixed-effects regression model (no intercept)</td></tr>
<tr><td>(f) Post-stratification (POST)</td><td>ANOVA type linear fixed-effects model</td></tr>
<tr><td colspan="4">(3) Generalized regression (GREG) estimation and model-assisted calibration (MC)</td></tr>
<tr><td>(g) GREG family & MC family</td><td>Continuous, binary, count, categorical</td><td>Unit-level</td><td>Members of generalized linear models (GLM) family</td></tr>
</table>

4.2 Ratio and regression estimation and calibration

4.2.1 Background

In ratio and regression estimation, the auxiliary information is incorporated into the estimation procedure by using linear regression models with continuous target variable and a single continuous covariate (ratio estimation) or several covariates (regression estimation). For a ratio or regression estimator, the known population totals of the auxiliary variables are required, and the unit-level values are needed for the sample elements. If desired, the methods can be expressed in the form of calibration estimators. In this section we discuss ratio and regression estimation and calibration weighting. Post-stratification is treated in Section 4.3.

4.2.2 Sampling and estimation

Model-assisted and calibration methods are applicable under any sampling design, but a relatively simple sampling design is often adopted, and efforts for precision improvement are devoted to the estimation phase. We introduce the basic methods for the estimation of the population total and the derivation of the appropriate variance, standard error and coefficient of variation estimates within the worked example section of each method. In all strategies considered, the underlying sampling design is simple random sampling without replacement.

4.2.3 Worked example

Preliminaries. We continue working with the set of active vessels in SIMPOP. Our target variable is again CATCH. This selection allows us to compare the performance of the methods with methods that use (or, not use) auxiliary data in the sampling phase. We assume that we have access to data on continuous type auxiliary variables GT (vessel tonnage) and DAS (days at sea) and the binary variable DOM01 (type of fishing).

We study the estimation strategies SRSWOR_CAL, SRSWOR_RAT and SRSWOR_REG, where the sample is drawn from SIMPOP by SRSWOR and estimation relies on a ratio estimator or a regression estimator. The strategy SRSWOR_HT serves as a reference strategy. We compare the results with the reference strategy by computing standard error, coefficient of variation and design effect estimates. Different sample sizes are applied.

Sample selection. We use SRSWOR as the sampling design. Sample realizations are named SAMPLE7 and SAMPLE8, corresponding the SRSWOR samples SAMPLE1 and SAMPLE2 in Section 3.3.4. Both new samples are amended with the selected auxiliary variables. Our first sample of size $n = 5$ active vessels from SIMPOP is displayed in Table 4.2.

Table 4.2 SRSWOR sample SAMPLE7 of $n = 5$ active vessels drawn from SIMPOP of $N = 100$ vessels amended with sample values of auxiliary variables GT, DAS and DOM01.

Obs k	**ID**	**CATCH** y_k	**GT** x_{1k}	**DAS** x_{2k}	**DOM01** x_{3k}	**Sampling Weight** w_k
1	1	3541.44	280.0	136	0	20
2	44	4421.92	282.1	165	1	20
3	49	11355.97	386.1	228	0	20
4	55	6865.42	408.0	213	0	20
5	93	9942.19	440.7	235	1	20
Sum						**100**

The selected auxiliary variables are x_1 (variable GT), x_2 (variable DAS) and x_3 (binary variable DOM01). We assume that we have the population totals of these variables at our disposal. The totals are given in Table 4.3.

Table 4.3 Population totals of the auxiliary variables.

Obs	GT t_{x_1}	DAS t_{x_2}	DOM01 t_{x_3}
1	32896.4	18308	30

We need the population totals t_{x_1}, t_{x_2} and t_{x_3} of the auxiliary variables for the construction of the various calibration, ratio and regression estimators.

Calibration estimation. In calibration estimation for the total of a target variable, we do not need to postulate any underlying model. The calibration weights are obtained by computational operations directly on the target and auxiliary variables.

We examine calibration estimation applied to ratio estimation for the total of CATCH. Let us take GT (vessel tonnage) as the auxiliary variable. The sample is SAMPLE7. Building blocks for ratio estimation of CATCH total are collected in Table 4.4.

Table 4.4. Components needed for the construction of a calibration estimator for CATCH total.

Variable	Source	Component
CATCH	sample	$\hat{t}_{HT} = 722539$
GT	sample	$\hat{t}_{HTx_1} = 35938$
GT	population	$t_{x_1} = 32896.44$

The estimation of the total can be executed by model-free calibration using the so-called g weights. We construct calibration weights as a product of sampling weight and *g* weight:

$$w_{CAL,k} = w_k \times g_k = w_k \times \frac{t_{x_1}}{\hat{t}_{HTx_1}} = 20 \times 0.91537 = 18.3073 ,$$

where the g weights are computed as

$$g_k = \frac{t_{x_1}}{\hat{t}_{HTx_1}} = \frac{32896.44}{35938} = 0.91537$$

i.e. constant for all sample elements, and the sampling weights are $w_k = 1/\pi_k = 20$. First, we check the calibration property (17) by computing the calibrated total of the auxiliary variable GT and obtain $\sum_{k=1}^{n} w_k g_k x_k = \sum_{k=1}^{N} x_k = t_{x_1} = 32896.44$, see Table 4.4. The calibration property thus holds.

Using calibration, the ratio estimate of the total of CATCH with GT as the auxiliary variable is computed as

$$\hat{t}_{CAL} = \sum_{k=1}^{5} w_{CAL,k} y_k = 18.3073 \times 36126.94 = 661387.$$

For illustration, the components for calculating the ratio estimate with calibration are inserted in Table 4.5. The sum of the components over the sample produces the ratio estimate. We discuss variance estimate for $\hat{t}_{CAL}$ below in connection to ratio estimation.

Table 4.5 SRSWOR sample SAMPLE7 of $n = 5$ active vessels drawn from SIMPOP of $N = 100$ vessels amended with g weights, calibration weights and components for calibration estimation.

Obs k	**ID**	**CATCH** y_k	**GT** x_{1k}	**g Weight** g_k	**Sampling Weight** w_k	**Calibration Weight** $w_{CAL,k}$	**Components** $w_{CAL,k} \times y_k$
1	1	3541.44	280.0	0.91537	20	18.3073	64834.23
2	44	4421.92	282.1	0.91537	20	18.3073	80953.40
3	49	11355.97	386.1	0.91537	20	18.3073	207897.29
4	55	6865.42	408.0	0.91537	20	18.3073	125687.28
5	93	9942.19	440.7	0.91537	20	18.3073	182014.76
Sum		**36126.94**			**100**	**91.5365**	**661387**

Ratio estimation. Ratio estimator $\hat{t}_{RAT}$ of population total $t = \sum_{k=1}^{N} y_k$ of the target variable is traditionally constructed as

$$\hat{t}_{RAT} = \hat{t}_{HT} \times \frac{t_x}{\hat{t}_{HTx}}, \tag{19}$$

where $\hat{t}_{HT} = \sum_{k=1}^{n} w_k y_k$ is the HT estimator of the total of the target variable, t_x is the known population total and $\hat{t}_{HTx} = \sum_{k=1}^{n} w_k x_k$ is the HT estimator of the auxiliary variable, and $w_k = 1/\pi_k$ are the sampling weights. It is important to note that for ratio estimation we only need the population total of the auxiliary variable. Unit-level values of the auxiliary variable are only needed for the sample.

We use GT (vessel tonnage) as the auxiliary variable x_1 for ratio estimation of the total of CATCH. Building blocks for ratio estimation are again taken from Table 4.4. Ratio estimate for the total of CATCH is computed as:

$$\hat{t}_{RAT} = \hat{t}_{HT} \times \frac{t_{x_1}}{\hat{t}_{HTx_1}} = 722539 \times \frac{32896.4}{35938} = 722539 \times 0.91537 = 661387\,,$$

i.e. the same estimate as was obtained with calibration estimation.

We discuss briefly ratio estimation as a model-assisted estimation method. In ratio estimation, the assisting model is simple. The regression model is given by

$$y_k = \beta_1 x_{1k} + \varepsilon_k,$$

where the slope parameter β_1 for the auxiliary variable is the sole parameter to be estimated, and ε_k stands for residuals. Note that the model does not involve an intercept term; it is assumed that the regression line goes through the origin (recall a similar assumption in PPS sampling).

Ratio estimation can be proceeded as a special case of regression estimation by fitting the linear model $y_k = \beta_1 x_{1k} + \varepsilon_k$ and computing the ratio estimate by using the estimated β-parameter and the known population total of GT. In practice, the β-parameter is estimated by weighted least squares using sampling weights. For the current SRS case we obtain a slope estimate $\hat{\beta}_1 = \frac{\hat{t}_{HT}}{\hat{t}_{HTx_1}} = 20.1051$ that is equal to the estimated ratio $\hat{r}$. The ratio estimate of the total is now computed as

$$\hat{t}_{RAT} = t_{x_1} \times \hat{\beta}_1 = 32896.44 \times 20.1051 = 661387,$$

that is, numerically the same estimate as the previous ones.

Variance estimation. Variance estimation for the ratio estimator $\hat{t}_{RAT}$ can be carried out first by presenting the ratio estimator in the form

$$\hat{t}_{RAT} = t_{x_1} \frac{\hat{t}_{HT}}{\hat{t}_{HTx_1}} = t_{x_1} \times \hat{r},$$

where $\hat{r} = \frac{\hat{t}_{HT}}{\hat{t}_{HTx_1}}$, and writing the design variance as

$$V(\hat{t}_{RAT}) = V\left(t_{x_1}\hat{r}\right) = t_{x_1}^2 \times V(\hat{r}). \quad (20)$$

Various approximate estimators $\hat{v}(\hat{r})$ for the nonlinear estimator $\hat{r}$ of the ratio of two HT estimated totals and $\hat{v}(\hat{t}_{RAT})$ for the total $\hat{t}_{RAT}$ are available in the literature, e.g. Cochran (1963), Särndal et al. (1992), Lehtonen & Veijanen (2009), as well as in software documentation (e.g. SAS procedures SURVEYMEANS and SURVEYREG and R `survey` function `calibrate`). In addition to the linearization method, pseudoreplication methods are available by the software products, e.g. SAS SURVEY procedures. The methods include the jackknife technique and balanced half-samples method. Variance estimators $\hat{v}(\hat{r})$ for $\hat{r}$ in eq. (20) are implemented for example in the SAS procedure SURVEYMEANS. If the population total t_{x_1} of the auxiliary variable is available, it is straightforward to compute an estimate $\hat{v}(\hat{t}_{RAT})$.

Many of the approximate variance estimators methods for ratio estimator as well as regression and calibration estimators rely on the estimation of residual variance, where residuals are computed under the fitted model as $e_k = y_k - \hat{y}_k$. This approach is used for example in the SAS procedure SURVEYREG, For a ratio estimator the fitted values are $\hat{y}_k = \hat{\beta}_1 x_{1k}$.

A simple variance estimator for $\hat{t}_{RAT}$ based on a residual variance estimator is given by

$$\hat{v}_1(\hat{t}_{RAT}) = \frac{n(1-f)}{n-1}\left(\frac{n-1}{n-p}\right)\sum_{k=1}^{n}(w_k e_k - \hat{t}_{HTe}/n)^2, \quad (21)$$

where $\hat{t}_{HTe} = \sum_{k=1}^{n} w_k e_k$ is the HT estimator of residual total, $e_k = y_k - \hat{y}_k$ are residuals with fitted values $\hat{y}_k = \hat{\beta}_1 x_{1k}$ from the model, $\hat{\beta}_1 = \hat{r} = \frac{\hat{t}_{HT}}{\hat{t}_{HTx1}}$ is the estimated slope term, $f = \frac{n}{N}$ is the sampling fraction and p is the number of model parameters. A g weighted version is often preferred, given by

$$\hat{v}_2(\hat{t}_{RAT}) = \frac{n(1-f)}{n-1}\left(\frac{n-1}{n-p}\right)\sum_{k=1}^{n}(w_k g_k e_k - \hat{t}_{CALe}/n)^2, \quad (22)$$

where $g_k = \frac{t_{x_1}}{\hat{t}_{HTx_1}}$ are g weights and $\hat{t}_{CALe} = \sum_{k=1}^{n} w_k g_k e_k$ is the g-weighted residual total estimate. This type of variance estimator is used for example in the SAS procedure SURVEYREG. Variance estimators (21) and (22) are asymptotically equivalent, because the g weights tend to unity with increasing the sample size.

Variance estimator $\hat{v}_3(\hat{t}_{RAT})$ for $\hat{t}_{RAT}$ resembling the standard variance estimator for a ratio estimator of total (e.g. Lehtonen & Pahkinen 2004 p. 98) is often used in practice and is given by

$$\hat{v}_3(\hat{t}_{RAT}) = \frac{n(1-f)}{n-1}\sum_{k=1}^{n}(w_k g_k y_k - w_k g_k \hat{r} x_{1k})^2, \quad (23)$$

where $\hat{r} = \frac{\hat{t}_{HT}}{\hat{t}_{HTx_1}}$. The estimator also uses the g weights. The SAS procedure SURVEYMEANS (`RATIO` statement) does not compute (23) directly but computes an estimate for the ratio $\hat{r}$ given by

$$\hat{v}_3(\hat{r}) = 1/t_{x_1}^2 \times \hat{v}_3(\hat{t}_{RAT}).$$

We obtain:

$$\hat{v}_3(\hat{t}_{RAT}) = t_{x_1}^2 \times \hat{v}_3(\hat{r}) = 32896.44^2 \times 8.01513 = 93133^2.$$

Numerical results from equations (21) and (22) do not differ much from an estimate using the variance estimator (23).

With a powerful auxiliary variable, variance estimates $\hat{v}(\hat{t}_{RAT})$ can be substantially smaller than a HT variance estimate $\hat{v}(\hat{t}_{HT})$, because the variation between residuals will be smaller when compared with the variation of the original values of target variable.

To complete, we compute variance estimate $\hat{v}_3(\hat{r})$ and standard error estimate $s.e(\hat{r})$ by using the SURVEYMEANS procedure. Results are in Table 4.6. The estimate $se(\hat{t}_{RAT})$ also is included. The results (see also Section A:6) agree with results computed with the R `survey` function `calibrate` (Section B.7.2).

Table 4.6 Estimation of the ratio $\hat{r}$ for the SRSWOR sample SAMPLE7 of $n = 5$ elements computed with PROC SURVEYMEANS, amended with results for the ratio estimate $\hat{t}_{RAT}$.

Ratio Analysis					
Numerator	**Denominator**	**Ratio** $\hat{r}$	**Std Err** $s.e(\hat{r})$	**Total** $\hat{t}_{RAT}$	**StdErr** $se(\hat{t}_{RAT})$
CATCH	**GT**	20.105147	2.831090	661387	93133

Coefficient of variation (5) for $\hat{t}_{RAT}$ is calculated as:

$$cv(\hat{t}_{RAT}) = \frac{s.e(\hat{t}_{RAT})}{\hat{t}_{RAT}} = \frac{93133}{661387} = 0.14.$$

Design effect estimate (7) of $\hat{t}_{RAT}$ is:

$$deff(\hat{t}_{RAT}) = \frac{\hat{v}_{SRSWOR}(\hat{t}_{RAT})}{\hat{v}_{SRSWOR}(\hat{t}_{HT})} = \frac{93133^2}{147823^2} = 0.40.$$

It should be recognized that in the deff formula, the estimators for the total in the numerator and denominator are different. The SRSWOR variance estimator in the numerator is for the ratio estimator $\hat{t}_{RAT}$ and in the denominator, it is for the HT estimator $\hat{t}_{HT}$. Because the deff estimate is smaller than one, the SRSWOR_RAT strategy is more efficient than would be a SRSWOR_HT strategy for SAMPLE7.

We finally execute ratio estimation by the SAS procedure SURVEYREG, which is aimed to design-based regression modeling. The model $y_k = \beta_1 x_{1k} + \varepsilon_k$ is first fitted for the sample data set and the ratio estimate is obtained by the `ESTIMATE` statement. Results are in Table 4.7. The sample is the one displayed in Table 4.2. The results agree pretty closely with the previous ones. Differences to Table 4.6 results are caused by the slightly different computation algorithms in the SAS procedures. The differences vanish with large samples.

Ratio estimation involves biased estimation, except in the theoretical case where the intercept term β_0 of the regression model $y_k = \beta_0 + \beta_1 x_{1k} + \varepsilon_k$ is zero. The order of the bias is $1/n$, indicating that with a small sample size the bias can be substantial.

Table 4.7. Ratio estimation by the SAS procedure SURVEYREG for a SRSWOR sample SAMPLE7 of $n = 5$ elements.

a) Estimated β-parameter

Estimated Regression Coefficients				
Parameter	**Estimate** $\hat{\beta}_1$	**Standard Error** $se(\hat{\beta}_1)$	**t Value**	**Pr > |t|**
GT β_1	20.6761657	2.72832884	7.58	0.0016

b) Auxiliary information provided

Estimate Coefficients	
Effect	**Row1** t_{x_1}
GT	32896

c) Ratio estimate of CATCH total by `ESTIMATE` statement

Estimate		
Label	**Estimate** $\hat{t}_{RAT}$	**Standard Error** $s.e(\hat{t}_{RAT})$
CATCH total	680171	89752

Regression estimation. We apply regression estimation to the estimation of the total of target variable CATCH under a SRSWOR sample, thus the strategy is now SRSWOR_REG. The variable GT (vessel tonnage) acts first as the auxiliary variable. The sample data set remains as SAMPLE7 also for this analysis.

Regression estimator for a total uses a linear fixed-effects regression model as the assisting model. In the simplest case with a single auxiliary variable, the model is of the form:

$$y_k = \beta_0 + \beta_1 x_{1k} + \varepsilon_k, \tag{24}$$

where x_{1k} are values of the continuous auxiliary variable and ε_k are residuals. Note that the assisting model now involves an intercept term.

The model parameters intercept β_0 and slope β_1 are first estimated by weighted least squares with sampling weights. By inserting the HT estimate $\hat{t}_{HT}$ of CATCH and HT estimate $\hat{t}_{HTx_1}$ of GT together with the known population total t_{x_1} of GT (Table 4.4) as well as the estimated slope $\hat{\beta}_1$ into the textbook formulation of a regression estimator (Lehtonen & Pahkinen 2004 p. 97):

$$\hat{t}_{REG} = \hat{t}_{HT} + \hat{\beta}_1(t_{x_1} - \hat{t}_{HTx_1}), \tag{25}$$

we get an estimate $\hat{t}_{REG} = 722539 + 37.4647(32896.44 - 35938) = 608586$.

For variance estimation we use the textbook estimator for regression estimation (Lehtonen & Pahkinen 2004 p. 98) that is based on the linearization method, given by:

$$\hat{v}_{SRSWOR}(\hat{t}_{REG}) = N^2(1 - \tfrac{n}{N})\left(\tfrac{1}{n}\right)\left(\tfrac{n-1}{n-p}\right) \times \hat{s}^2_{e_{CAL}}, \tag{26}$$

where $\hat{s}^2_{e_{CAL}}$ is the sample variance of g weighted residuals $g_k e_k = g_k(y_k - \hat{y}_k)$ with fitted values $\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_{1k}$ from the model, and p is the number of model parameters. The residual variance estimator is given by $\hat{s}^2_{e_{CAL}} = \sum_{k=1}^{n}(g_k e_k - \bar{e})^2/(n-1)$ with $\bar{e} = \sum_{k=1}^{n} g_k e_k/n$, the mean of g-weighted residuals. Pseudoreplication methods can be used as an alternative.

We execute the estimation by the SAS procedure SURVEYREG using estimator (26). In Table 4.8, estimation results for model (1) are displayed in Part a), including the estimates of the β-parameters and standard errors. Part b) contains the auxiliary information t_{x_0} and t_{x_1} supplied, where variable x_0 refers to the intercept. Regression estimate $\hat{t}_{REG}$ for the total, with standard error estimate and confidence interval, is obtained by the SURVEYREG statement `ESTIMATE`. SAS results in tables 4.8 and 4.10 agree with R `survey` function `svyglm` results except for the constant $(n-1)/(n-p)$ of the SAS variance formula (26) (sections A.6 and B.7.3). Regression estimation is also illustrated in Section 8.2 (the case study for Finland).

Table 4.8. Regression estimation by the SAS procedure SURVEYREG for the SRSWOR sample SAMPLE7 of $n = 5$ elements.

a) Estimated β-parameters

Estimated Regression Coefficients				
Parameter	**Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept β_0	-6238.6791	2363.59006	-2.64	0.0576
GT β_1	37.4647	8.53363	4.39	0.0118

b) Auxiliary information provided

Estimate Coefficients	
Effect	**Row1**
Intercept t_{x_0}	100
GT t_{x_1}	32896

c) Regression estimate of CATCH total by `ESTIMATE` statement

Estimate					
Label	**Estimate** $\hat{t}_{REG}$	**Standard Error** $s.e(\hat{t}_{REG})$	**Alpha**	**Lower** $LCL(\hat{t}_{REG})$	**Upper** $UCL(\hat{t}_{REG})$
CATCH total	608586	78985	0.05	389288	827884

We next compute the coefficient of variation (5) and design effect estimate (7) for $\hat{t}_{REG}$:

- Coefficient of variation: $cv(\hat{t}_{REG}) = \frac{s.e(\hat{t}_{REG})}{\hat{t}_{REG}} = \frac{78985}{608586} = 0.13$
- Design effect estimate: : $deff(\hat{t}_{REG}) = \frac{\hat{v}_{SRSWOR}(\hat{t}_{REG})}{\hat{v}_{SRSWOR}(\hat{t}_{HT})} = \frac{78985^2}{147823^2} = 0.28$

Regression estimation with a single auxiliary variable GT appears effective for the CATCH total. Coefficient of variation is 13%, smaller than the SRSWOR_HT counterpart 20%. The deff estimate for the SRSWOR_REG strategy also indicates substantial improvement of statistical efficiency over the SRSWOR_HT strategy.

Extension of regression estimation for multiple auxiliary variables is straightforward. Let us take the variables GT, DAS and DOM01 in the model. The model is now of the form:

$$y_k = \beta_0 + \beta_1 x_{1k} + \beta_2 x_{2k} + \beta_3 x_{3k} + \varepsilon_k. \quad (27)$$

We fit model (27) for SAMPLE8 of n = 20 units. The multiple regression estimator is given by:

$$\hat{t}_{REG} = \hat{t}_{HT} + \hat{\beta}_1(t_{x_1} - \hat{t}_{HTx_1}) + \hat{\beta}_2(t_{x_2} - \hat{t}_{HTx_2}) + \hat{\beta}_3(t_{x_3} - \hat{t}_{HTx_3}). \quad (28)$$

Materials for computing $\hat{t}_{REG}$ with equation (28) under model (27) are summarized in Table 4.9. Estimated slopes are given in Part a) of SURVEYREG output presented in Table 4.10. We obtain:

$$\hat{t}_{REG} = 610603 + 20.5547(32896.44 - 31255) + 33.3466(18308 - 18680) - 545.3683(30 - 40) = 637401.$$

Table 4.9. Components needed for the construction of a regression estimator for CATCH total with three auxiliary variables under SAMPLE8 of $n = 20$ elements.

Variable	Source	Component
CATCH	sample	$\hat{t}_{HT} = 610603$
GT	sample	$\hat{t}_{HTx1} = 31255$
GT	population	$t_{x_1} = 32896.44$
DAS	sample	$\hat{t}_{HTx2} = 18680$
DAS	population	$t_{x_2} = 18308$
DOM01	sample	$\hat{t}_{HTx3} = 40$
DOM01	population	$t_{x_3} = 30$

Coefficient of variation is $cv(\hat{t}_{REG}) = \frac{31040}{637401} = 0.048$ and $deff(\hat{t}_{REG}) = \frac{31040^2}{54439^2} = 0.32$. The cv for SRSWOR_HT strategy was $cv(\hat{t}_{SRSWOR}) = 0.089$ indicating better efficiency for strategy SRSWOR_REG.

Table 4.10. Regression estimation for CATCH with GT, DAS and DOM01 as auxiliary variables by SAS procedure SURVEYREG for the SRSWOR sample SAMPLE8 of $n = 20$ elements.

a) Estimated β-parameters

Estimated Regression Coefficients				
Parameter	**Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept β_0	-6329.2340	1657.38173	-3.82	0.0012
GT β_1	20.5547	5.06384	4.06	0.0007
DAS β_2	33.3466	6.22707	5.36	<.0001
DOM01 β_3	-545.3683	561.62140	-0.97	0.3437

b) Auxiliary information provided

Estimate Coefficients	
Effect	**Row1**
Intercept t_{x_0}	100
GT t_{x_1}	32896
DAS t_{x_2}	18308
DOM01 t_{x_3}	30

c) Regression estimate of CATCH total by `ESTIMATE` statement

Estimate					
Label	**Estimate** $\hat{t}_{REG}$	**Standard Error** $s.e(\hat{t}_{REG})$	**Alpha**	**Lower** $LCL(\hat{t}_{REG})$	**Upper** $UCL(\hat{t}_{REG})$
CATCH total	637401	31040	0.05	572433	702369

Simulation experiment. For comparing the average capacity of strategy SRSWOR_REG with SRSWOR_HT we conduct a small pedagogic simulation experiment. We draw $K = 100$ SRSWOR samples of size $n = 20$ vessels from SIMPOP, compute the estimated total, s.e and cv for the regression estimator $\hat{t}_{REG}$ under the SRSWOR_REG strategy and the HT estimator $\hat{t}_{HT}$ under the SRSWOR_HT strategy from each sample, and compute the mean of the statistics over the 100 samples. Auxiliary variables are GT, DAS and DOM01. The results are in Table 4.11.

The average standard error of $\hat{t}_{REG}$ is smaller than that of $\hat{t}_{HT}$, leading to better efficiency for strategy SRSWOR_REG. This is further manifested by the coefficients of variation: average cv for REG estimator is 4% and average cv for HT estimator is 7%. The results also indicate the design unbiasedness of the regression estimator. It thus seems that the application of regression estimation makes sense.

Table 4.11 Means of estimated totals, standard errors and coefficients of variation for CATCH from $K = 100$ simulated SRSWOR samples of size $n = 20$ vessels from SIMPOP for strategies SRSWOR_HT and SRSWOR_REG with GT, DAS and DOM01 as auxiliary variables.

Strategy	**VarName**	**Replicates**	**Averages over simulations**				
			SumWgt	**n**	**Total** $\hat{t}$	**StdDev** $s.e(\hat{t})$	**CV** $cv(\hat{t})$
SRSWOR_REG	CATCH	100	100.000000	20	626600	26163	0.041837
SRSWOR_HT	CATCH	100	100.000000	20	626895	44061	0.070264
True total					624036		

4.2.4 Estimation for domains

We continue with the estimation for population subgroups or domains where the sample size in domains is not controlled by stratification but is a random variate; the domains are thus of *unplanned* type. In Section 3.3.4 we used the strategy SRSWOR_HT for the estimation for unplanned domains under the conditional approach (variances of estimators were computed conditionally on the observed domain sample sizes) and the unconditional approach (the randomness of domain sample sizes were accounted for by using the extended domain variables technique). We use here ratio estimation applied separately for the two domains, thus resembling the conditional approach for estimation for unplanned domains. The strategy adopted thus is SRSWOR_RAT.

The analysis is executed independently for the two domains by the SAS procedure SURVEYMEANS (`RATIO` statement). As the domain variable we use the two-category variable DOM01. For auxiliary information we use the known domain sizes $N_0 = 70$ and $N_1 = 30$ in population. CATCH is the variable of interest.

For domain ratio estimation of CATCH totals with DOM01 as the two-category domain variable we use the SRSWOR sample SAMPLE9 of size $n = 20$ vessels. The distribution of the sample over the domains is $n_0 = 12$ for the first domain and $n_1 = 8$ for the second, as presented in Table 4.13.

A domain-specific ratio estimator can be written as $\hat{t}_{dRAT} = \hat{r}_{dHT} \times N_d = \frac{\hat{t}_{dHT}}{\widehat{N}_d} N_d$ where $\hat{r}_{dHT} = \frac{\hat{t}_{dHT}}{\widehat{N}_d}$ is the HT estimated ratio, N_d is the known domain size in population and $\widehat{N}_d = \sum_{k \in s_d} w_k$ is the HT estimated domain size, and $\hat{t}_{dHT} = \sum_{k \in s_d} w_k y_k$. Here $\widehat{N}_0 = 60$ and $\widehat{N}_1 = 40$.

We obtain:

- DOM01=0: $\hat{t}_{0RAT} = \frac{\hat{t}_{0HT}}{\widehat{N}_0} N_0 = \frac{419535.95}{60} \times 70 = 489459$
- DOM01=1: $\hat{t}_{1RAT} = \frac{\hat{t}_{1HT}}{\widehat{N}_1} N_1 = \frac{191067.50}{40} \times 30 = 143301$

HT estimates $\hat{t}_{0HT}$ and $\hat{t}_{1HT}$ are from Table 3.11. For variance estimation we apply the standard SRSWOR variance formula of ratio estimator (Lehtonen & Pahkinen 2004 p. 93) separately for each domain, given by

$$\hat{v}_{SRSWOR}(\hat{t}_{dRAT}) = N_d^2 \left(1 - \frac{n_d}{N_d}\right) \left(\frac{1}{n_d}\right) \sum_{k \in s_d} (y_k - \hat{r}_{dHT} x_{dk})^2 / (n_d - 1),$$

where $x_{dk} = 1$ for $k \in s_d$ in this case. Here $\hat{r}_{0HT} = 6992.266$ and $\hat{r}_{1HT} = 4776.687$. We obtain:

- DOM01=0: $\hat{v}(\hat{t}_{0RAT}) = 70^2 \left(1 - \frac{12}{70}\right) \left(\frac{1}{12}\right) \times 103228219.96/(12 - 1) = 56348^2$
- DOM01=1: $\hat{v}(\hat{t}_{1RAT}) = 30^2 \left(1 - \frac{8}{30}\right) \left(\frac{1}{8}\right) \times 103228219.96/(8 - 1) = 12836^2$

Results are collected in Table 4.11. A comparison with Table 3.11 indicates that our results under SRSWOR_RAT strategy are close to those from strategy SRSWOR_HT for the conditional approach for variance estimation with using the known domain sizes in population.

Estimators of domain totals using known domain sizes as auxiliary information can also be derived as Hajék type estimators $\hat{t}_{dHA} = \frac{N_d}{\widehat{N}_d} \hat{t}_{dHT}$ giving the same numerical results as the ratio estimators in Table 4.12. Variance estimation for ratio and Hajék type estimators for domain totals is discussed in Lehtonen & Veijanen (2009) p. 241-242.

Table 4.12 Estimation of CATCH totals for two unplanned domains under strategy SRSWOR_RAT computed for SAMPLE9 of size $n = 20$ vessels.

DOMAIN d	**Variable**	**n**	**Sum of Weights**	**Total** $\hat{t}_{RAT,d}$	**Std Dev** $s.e(\hat{t}_{RAT,d})$	**Coeff of Var** $cv(\hat{t}_{RAT,d})$
0	**CATCH**	12	60.000000	489459	56348^2	0.11512
1	**CATCH**	8	40.000000	143301	12836^2	0.08957

4.3 Post-stratification

4.3.1 Background

Discrete or categorical auxiliary information can be used in stratification of a sample after it has been drawn (*post-stratification*). The idea in post-stratification, similarly as in stratified sampling (Section 3.6), is to make the estimation more efficient by selecting post-strata where the within-stratum variation of the variable of interest is smaller than the variation between the strata. Auxiliary information for post-stratification is often obtained from official statistics or some other reliable source. Post-stratification can also be used in the adjustment of unit nonresponse in surveys (Chapter 5).

4.3.2 Sampling and estimation

Similarly as for ratio and regression estimation, post-stratification is applicable under any sampling design, but a relatively simple sampling design is often adopted such as simple random sampling or stratified SRS.

Post-stratification estimator of a population total is given by

$$\hat{t}_{POST} = \sum_{c=1}^{C} \hat{t}_c = \sum_{c=1}^{C} \sum_{k=1}^{n_c} w_{POST,ck} y_k, \quad (29)$$

where $w_{POST,ck}$ are post-stratum weights for element k in post-stratum c derived for the entire sample. Post-stratum weights in (29) are

$$w_{POST,ck} = g_{ck} w_{ck},$$

where $g_{ck} = N_c / \widehat{N}_c$ and the denominator $\widehat{N}_c = \sum_{k=1}^{n_c} w_{ck}$ is the estimated post-stratum size, and $w_{ck} = 1/\pi_{ck}$ are the original sampling weights. For variance estimation, the post-strata may be regarded as unplanned domains (see sections 2.5, 3.3.4 and 4.2.4).

4.3.3 Worked example

Preliminaries. We continue working with the set of active vessels in SIMPOP and the target variable CATCH. Post-stratification is introduced as a calibration method in a simple and well manageable case. In the method, we assume an access to a single categorical auxiliary variable suitable for post-stratification. The binary variable DOM01 (fishing type) is chosen, which indicates whether a vessel catches "expensive" fish (DOM01 = 1) or not (DOM01 = 0). Two post-strata will be constructed under the given sampling design.

We study the estimation strategy SRSWOR_POST, where the sample is drawn from SIMPOP by SRSWOR and the estimation relies on post-stratification. The strategy SRSWOR_HT serves as the reference strategy. We compare the results with our reference strategy by computing coefficient of variation and design effect estimates.

Sample selection. The SRSWOR sample is SAMPLE9, corresponding to the SRSWOR sample SAMPLE8 in Section 4.3. The original sample contains the values of the ID variable, target variable CATCH, auxiliary variable DOM01 and the sampling weight. The sample data set has been amended with five derived variables that are used in constructing the post-stratification estimator. A two-class variable POST2 with value 1 (if DOM=0) and 2 (if DOM=1) has been created for post-stratum identification. To illustrate post-stratification by calibration we use a sample size $n = 20$ active vessels from SIMPOP. The complete data set is displayed in Table 4.12. The sample data set is sorted by the post-stratification variable.

Estimation. The binary auxiliary variable DOM01 is selected for post-stratification. We execute the estimation of the post-stratified estimate $\hat{t}_{POST}$ with the calibration technique, based on reweighting. For calibration we compute g weights and calibrated weights for the two post-strata. Data for g weights consists of the population distribution and weighted sample distribution of the variable POST2, given in the set-up below:

Variable	**Level** c	**n** n_c	N_c	$\widehat{N}_c$
POST2	**1**	12	70	60
	2	8	30	40

Using notation of Section 4.3.2, *g* weights are computed as $g_{ck} = \frac{N_c}{\widehat{N}_c}$, where N_c is known size of post-stratum c and $\widehat{N}_c = \sum_{k=1}^{n_c} w_{ck}$ is the HT estimate of $N_c, c = 1,2$, where $w_{ck} = 5$ is the SRSWOR sampling weight. We get for post-stratum 1: $g_{1k} = \frac{N_1}{\widehat{N}_1} = \frac{70}{60} = 1.16667$ and for post-stratum 2: $g_{2k} = \frac{N_2}{\widehat{N}_2} = \frac{30}{40} = 0.75$. The *g* weights are included in Table 4.13. Calibrated weights in Table 4.13 are computed as $w_{POST,ck} = w_{ck} \times g_{ck}$. The sum of the final column in the table provides the post-stratified (calibration) estimate $\hat{t}_{POST}$ of CATCH. The estimate is thus computed by (29) as

$$\hat{t}_{POST} = \sum_{c=1}^{2} \sum_{k=1}^{n_c} w_{POST,ck} y_k = 632759.$$

We execute post-stratification by the SAS procedure SURVEYMEANS using the `POSTSTRATA` statement. The procedure estimates totals, standard errors and cv:s by using the post-stratification (calibration) weights $w_{POST,ck}$ instead of the original sampling weights w_k. Variance estimate for $\hat{t}_{POST}$ is computed by equation:

$$\hat{v}(\hat{t}_{POST}) = n\left(1 - \frac{n}{N}\right) \sum_{c=1}^{C} \sum_{k=1}^{n_c} \left(w_{POST,ck} y_k - \hat{t}_{POST,c}/n_c\right)^2 / (n-1), \qquad (30)$$

where $\hat{t}_{POST,1} = 489459$ and $\hat{t}_{POST,2} = 143301$. We obtain:

$$\hat{v}(\hat{t}_{POST}) = 20 \times \left(1 - \frac{20}{100}\right) \times 3708808885.8/(20-1) = 55885.66^2.$$

Table 4.13 SRSWOR sample SAMPLE9 of $n = 20$ active vessels from SIMPOP of $N = 100$ vessels completed with sample values of auxiliary variable DOM01 and five derived variables.

Obs k	**ID**	**CATCH** y_k	**DOM01** x_{3k}	**POST2** c	**Sampling Weight** w_{ck}	**g Weight** g_{ck}	**Post-Weight** $w_{POST,ck}$	**Components** $w_{POST,ck} \times y_k$
1	1	3541.44	0	1	5	1.1667	5.833	20658.46
2	9	2752.96	0	1	5	1.1667	5.833	16058.98
3	29	7518.96	0	1	5	1.1667	5.833	43860.73
4	41	3651.90	0	1	5	1.1667	5.833	21302.81
5	47	8715.89	0	1	5	1.1667	5.833	50842.84
6	56	6185.86	0	1	5	1.1667	5.833	36084.29
7	63	10270.01	0	1	5	1.1667	5.833	59908.56
8	68	11693.89	0	1	5	1.1667	5.833	68214.55
9	69	8709.47	0	1	5	1.1667	5.833	50805.39
10	71	4031.71	0	1	5	1.1667	5.833	23518.38
11	78	6219.11	0	1	5	1.1667	5.833	36278.25
12	94	10615.99	0	1	5	1.1667	5.833	61926.79
13	7	2642.64	1	2	5	0.7500	3.750	9909.90
14	20	4158.35	1	2	5	0.7500	3.750	15593.81
15	22	3538.14	1	2	5	0.7500	3.750	13268.03

Obs k	ID	CATCH y_k	DOM01 x_{3k}	POST2 c	Sampling Weight w_{ck}	g Weight g_{ck}	Post-Weight $w_{POST,ck}$	Components $w_{POST,ck} \times y_k$
16	24	4962.48	1	2	5	0.7500	3.750	18609.30
17	34	4363.01	1	2	5	0.7500	3.750	16361.29
18	37	6682.50	1	2	5	0.7500	3.750	25059.38
19	51	6638.87	1	2	5	0.7500	3.750	24895.76
20	79	5227.51	1	2	5	0.7500	3.750	19603.16
		122120.68			**100**	**20**	**100.000**	**632760.63**

Results in Table 4.13 by the SAS procedure SURVEYMEANS agree with results from the R `survey` function `postStratify` (sections A.7 and B.8.2).

Post-stratification resembles stratified sampling, but there are certain differences. In stratified sampling, stratum sample sizes are fixed by sample allocation. In post-stratification, post-strata are created after sample selection and there is no underlying allocation scheme. Post-stratum sample sizes are not controlled by the sampling design but are random variates, similarly as for unplanned domains under the unconditional approach. In SURVEYMEANS the variance estimate is computed using the conditional approach given observed post-stratum sizes (i.e. assuming they are fixed quantities) and thus, the randomness of post-stratum sizes is not accounted for. This leads to somewhat liberal variance estimates at least in small samples, because the unconditional variance estimates would be larger (see details e.g. Lehtonen & Pahkinen 2004 pp. 89-92).

Estimated coefficient of variation (2) and design effect (7) for $\hat{t}_{POST}$ are the following.

- Coefficient of variation: $cv(\hat{t}_{POST}) = \frac{s.e(\hat{t}_{POST})}{\hat{t}_{POST}} = \frac{55885.66}{632759} = 0.088$
- Design effect estimate: $deff(\hat{t}_{POST}) = \frac{\hat{v}_{SRSWOR}(\hat{t}_{POST})}{\hat{v}_{SRSWOR}(\hat{t}_{HT})} = \frac{55885.66^2}{54439^2} = 1.05.$

Coefficient of variation estimate of 8.8% is reasonable for practical purposes. The deff estimate indicates that nearly equal efficiency would be obtained by strategies SRSWOR_POST and SRSWOR_HT. Results for strategy SRSWOR_POST computed with SURVEYMEANS are summarized in Table 4.13. Estimates for SRSWOR_HT of Section 3.3.4 are also given.

It can be seen that post-stratification does not improve relative precision much in this case. A certain benefit still remains. By post-stratification estimation, the sample distribution of DOM01 will coincide with that in the population. This property is called *coherence* and is appreciated in official statistics. It is often considered feasible that marginal distributions (or totals) of auxiliary variables in a survey reproduce the published official statistics of these variables. By using data in Table 4.14 we get $\hat{t}_{POST,x_3} = \sum_{c=1}^{2}\sum_{k=1}^{n_c} w_{POST,ck}x_{3k} = 3.750 \times 8 = t_{x_3} = 30$.

Table 4.14 Estimated totals, standard errors and coefficients of variation for CATCH under strategies SRSWOR_POST and SRSWOR_HT computed for SAMPLE9 of size $n = 20$ vessels.

Strategy	Variable	n	Sum of Weights	Total $\hat{t}$	Std Dev $s.e(\hat{t})$	95% CL		Coeff of Var $cv(\hat{t})$
SRSWOR_POST	CATCH	20	100.000000	632759	55889	515782.798	749735.535	0.088325
SRSWOR_HT	CATCH	20	100.000000	610603	54439	496661.885	724544.886	0.089156

4.4 Comparison of model-assisted estimates

We finally compare the efficiency of ratio and regression estimation and post-stratification in the estimation of CATCH total under the same SRSWOR sample of $n = 20$ elements that was used in Sections 3.3.4, 4.2.3 and 4.3.3. As auxiliary data we use the two-category post-stratification variable POST2 under strategy SRSWOR_POST, the continuous variable GT under SRSWOR_RAT and GT, DAS and DOM01 for SRSWOR_REG. Results are in Table 4.15.

Obviously, a clever use of auxiliary information in model-assisted estimation can improve substantially the efficiency of estimation when compared to HT estimation for a simple random sample. Regression estimation offers a flexible tool for efficient estimation with multiple auxiliary variables requiring minimal auxiliary information.

Table 4.15 Estimated totals, standard errors, coefficients of variation and design effects for CATCH under strategies SRSWOR_HT, SRSWOR_POST, SRSWOR_RAT and SRSWOR_REG, computed for SAMPLE9 of size $n = 20$ vessels.

Strategy	**n**	**Auxiliary Data**	**Total** $\hat{t}$	**Std Dev** $s.e(\hat{t})$	**Coeff of Variation** $cv(\hat{t})$	**Deff** $deff(\hat{t})$
SRSWOR_HT	20	none	610603	54439	0.089156	1.00
SRSWOR_POST	20	POST2	632759	55889	0.088325	1.05
SRSWOR_RAT	20	GT	654899	46310	0.070713	0.72
SRSWOR_REG	20	GT, DAS, DOM01	637401	31040	0.048698	0.32

5 Treatment of nonresponse

5.1 Introduction

Nonresponse is common in surveys where sample data are collected by direct data collection methods such as personal interviews and postal questionnaires. Missing data can sometimes also appear in administrative registers. In the course of data collection some information may be lost, the main reason being unit nonresponse. The number of records in the sample data set will be smaller than intended. If the sample size was not inflated beforehand by an anticipated nonresponse, the precision of estimates will be weaker than would be with the original sample size. The set-up below shows how missingness affects the analysis data set.

Type of missingness	**Effect on the analysis data set**
Unit nonresponse	The whole data record remains missing (or all items were rejected)
Item nonresponse	One of more items for one or more variables are missing/are rejected
Sub-unit or partial nonresponse	All data from one or more elements within a cluster are missing

Nonresponse causes *selection bias* in estimates if the responding and nonresponding sets of the sample differ with respect to the distribution of survey variables. Various methods have been proposed in the literature to adjust for selection bias. Major adjustment methods assume *ignorable nonresponse*, where *response mechanism*, i.e. an unknown stochastic process that generates response or non-response in survey, is independent on the target variable of interest when conditioning on one or several auxiliary variables. The selection bias can then be adjusted for by conditioning on the covariates, for example by inserting the auxiliary variables as covariates in a nonresponse model. Under a more serious *non-ignorable nonresponse*, selection bias does not vanish after conditioning on the covariates. This type of missingness is difficult to be handled. Sometimes an assumption of a *completely random missingness* is made, i.e. missingness does not correlate with the survey variables. Unfortunately, the assumption of no selection bias is rarely in effect in real world.

Some common methods for dealing with missingness in survey are introduced in this chapter. These include imputation methods (mean imputation, hot deck imputation and regression imputation) for adjusting for item nonresponse and reweighting methods for unit nonresponse adjustment, such as the response homogeneity (RHG) technique. In the worked example section (Section 5.5) we apply methods of Chapter 4 (regression estimation and post-stratification) for adjusting missingness in a survey. Methods dealing with nonresponse are discussed widely in the literature, for example Groves et al. (2002), Lehtonen & Pahkinen (2004), Särndal & Lundström (2005), Enders (2010) and Little & Rubin (2014).

5.2 Response mechanism

Various hypothetical response mechanisms have been suggested in the literature. Under a *missing completely at random* (MCAR) mechanism the probability of missingness is independent on the observed or missing data. This option is rarely in effect in real world. If the probability of response depends only on the observed data, missingness is said to follow a *missing at random* (MAR) mechanism. The MAR assumption is the most common in surveys, and many methods and programs for the adjustment of selection bias caused by the missingness are relying on this assumption. Thirdly, if the probability of nonresponse depends both on observed and missing data, response mechanism is defined as *not missing at random* (NMAR). From the three mechanisms the NMAR assumption is the most difficult to address as the missing values remain unknown (Heeringa et al. 2017).

5.3 Traditional nonresponse treatment methods

There are numerous traditional nonresponse treatment methods available and if the number of the missing values is relatively small or the response mechanism can be assumed ignorable the methods may provide a quick fix to the problem. Four traditional methods are described.

5.3.1 Case deletion methods

One of the simplest and probably the most popular traditional nonresponse treatment methods is the *case deletion methods* where records are removed from data matrix if they contain any missing information. In the *complete unit deletion* every unit with any missing item values is deleted before the analysis. This kind of deletion leads to a complete data set that is easy to analyze but may seriously distort the estimation results if the assumption of the MCAR mechanism is not in effect.

For *pairwise deletion method,* units with missing item information are not removed from the analysis data set but are 'dropped out' when these units cannot be used in a specific analysis. The number of observations, therefore, might change for example in computing several pairwise correlation coefficients and fitting regression models with different sets of covariates. If not otherwise specified, the pairwise deletion is the default treatment of missing items in most statistical software packages. Generally, the deletion also wastes data and leads to decreased statistical power. It cannot, therefore, be recommended unless the amount of missing data is trivially small (Enders 2010).

5.3.2 Mean imputation

One popular traditional method is *mean imputation* where missing values are replaced with some kind of mean value statistics. This approach is easy to understand but it can seriously distort the parameter estimates as the replaced values are 'forced' to the mean values. Intuitively it is clear that the use of the mean values decreases variance estimates.

5.3.3 Hot-deck imputation

Imputation is called *hot-deck imputation* when missing values are replaced with real observations of the same variable taken from *donors*. The method may be supplemented by drawing the donors from groups of similar observations. Hot-deck imputation does not necessarily underestimate the variance estimates as much as mean imputation but can produce biased estimates for the measures of association (e.g. correlations, regression estimates).

5.3.4 Regression imputation

In *regression imputation* a regression model is specified to predict the missing values by using the estimated model parameters and a set of covariates. Missing values are replaced with predicted values $\hat{y}_k = \boldsymbol{x}_k{}'\widehat{\boldsymbol{\beta}}$, where $\boldsymbol{x}_k$ is a vector of covariate values for unit k and$\widehat{\boldsymbol{\beta}}$ is the vector of estimated model parameters. If several analysis variables are imputed, different regression specification might be needed for each variable. Even though regression estimation is superior when compared to mean imputation, it can lead to overestimation of correlations.

Stochastic regression imputation is a modified version of the ordinary regression imputation. After the specification of regression model, normally distributed random numbers $(u_i\ )$ are generated and attached to the predicted values $\hat{y}_k$ giving modified predictions $\hat{y}_k^* = \boldsymbol{x}_k{}'\widehat{\boldsymbol{\beta}} + u_k$. When compared to the regression imputation, the extra step restores some of the original variation of the variable and leads to unbiased parameter estimates under the MAR response mechanism (Enders 2010).

5.4 Reweighting for unit nonresponse

For unit nonresponse, complete records may be missing. As the sample data set is smaller than intended, standard errors are increased. More importantly, estimation may be biased if the missingness is selective. Adjustment for selection bias can be done with reweighting methods by suitably modifying the original sampling weights.

Assuming estimated *response propensities* $\hat{\theta}_k; k \in s$, where s is the original sample of *n* units, modified sampling weights can be written as $w_{k,rw} = 1/(\pi_k\hat{\theta}_k)$. Estimated response propensities can be obtained for example by fitting a logistic regression model on a binary variable having value 1 for respondents and 0 for nonrespondents with covariates that explain the missingness and whose sample values are available both for respondents and nonrespondents.

Further, with a naïve assumption of constant response propensity θ_k for all population units, a reweighted Horvitz-Thompson estimator for total is given by

$$\hat{t}_{RHT} = \hat{\theta}^{-1} \sum_{k=1}^{n_{(r)}} w_k y_k,$$

where $n_{(r)}$ is the number of responding sample units and $\hat{\theta} = n_{(r)}/n$ is an estimator of the common response propensity. It is, however, preferable to model the structure of the response probabilities in greater detail. A straightforward modification is the *response homogeneity group method* (RHG method), where the population is divided into C response groups such that estimated response propensity $\hat{\theta}_c$ is assumed equal in response group c but can differ between groups. Propensities $\hat{\theta}_c$ are obtained in a similar manner as for $\hat{t}_{RHT}$. The RHG estimator for total is given by

$$\hat{t}_{RHG} = \sum_{c=1}^{C} \left(\hat{\theta}_c\right)^{-1} \sum_{k=1}^{n_{c(r)}} w_{ck} y_{ck},$$

where $\hat{\theta}_c = n_{c(r)}/n_c$ (Lehtonen & Pahkinen 2004).

5.5 Worked example

Preliminaries. We examine here the applicability of model-assisted and calibration methods for the adjustment for the possible bias due to nonresponse. Adjustment for nonresponse is discussed for the case where some measurements or entire records are missing for the target variable. If aggregate level or unit level auxiliary variables and their complete sample values are available, the model-assisted methods and calibration techniques of Chapter 4 can be used. Adjustment may be effective if the response mechanism is of *ignorable type* so that the response mechanism may correlate with the auxiliary variables but not with the target variable. We examine nonresponse adjustment with a single auxiliary variable. The methods can be readily extended to multiple auxiliaries case.

Sample selection. We assume that the original $n = 20$ element sample has been drawn by SRSWOR from SIMPOP. The realized sample SAMPLE10 is the one we had in Section 4.4.4. Target variable is CATCH, and the continuous variable GT and categorical variable POST5 have been taken from the sampling frame and merged with the sample data set. POST5 was created by dividing GT into five equally-sized classes. GT is aimed for regression estimation and POST5 is for post-stratification.

Because of unit nonresponse in the data collection phase, the sample data set is contaminated by nonresponse for the target variable CATCH. We generated nonresponse for CATCH in a controlled manner, under the Missing at Random (MAR) missingness mechanism. MAR refers to ignorable type missingness, where the nonresponse mechanism and the target variable are conditionally independent given the auxiliary variables or covariates.

The analysis data set of $n = 20$ elements includes complete records for two auxiliary variables GT and POST5 for all 20 elements. Measurements for of CATCH are missing for two records. The data set includes a missingness indicator variable with value $I_k = 1$ for respondents and $I_k = 0$ for nonrespondents. The sample data set is sorted by the variable POST5 and is displayed in Table 5.1.

Table 5.1 Analysis data set SAMPLE10 of $n = 20$ vessels and two vessels with missing data for target variable CATCH.

Obs k	**ID**	**I** I_k	**CATCH** y_k	**GT** x_{1k}	**POST5** x_{2k}	**Sampling Weight** w_k
1	7	1	2642.64	218.4	1	5
2	9	1	2752.96	210.6	1	5
3	22	1	3538.14	229.6	1	5
4	24	1	4962.48	232.0	1	5
5	29	1	7518.96	266.8	1	5
6	37	0	**...**	270.0	1	5
7	1	1	3541.44	280.0	2	5
8	20	1	4158.35	305.2	2	5
9	41	1	3651.90	282.0	2	5
10	34	1	4363.01	312.0	3	5
11	51	0	**...**	320.1	3	5
12	56	1	6185.86	319.6	3	5
13	69	1	8709.47	316.8	3	5
14	78	1	6219.11	321.9	3	5
15	47	1	8715.89	359.7	4	5
16	71	1	4031.71	370.8	4	5
17	63	1	10270.01	392.0	5	5
18	68	1	11693.89	399.6	5	5
19	79	1	5227.51	407.0	5	5
20	94	1	10615.99	436.8	5	5
Sum		**18**				**100**

... denotes a missing value of a variable for the record

Estimation. For adjusting for nonresponse in the sample data we apply model-assisted or reweighting methods; post-stratification and regression estimation were chosen. The aim is that the re-weighted estimate of the auxiliary variable POST5 or GT for the incomplete data set reproduces the known population distribution of POST5 or the population total of GT. A successful adjustment for nonresponse bias requires nonzero correlation between the auxiliary variable and the response mechanism. In addition to adjust for the

nonresponse, improvement of efficiency of estimation for CATCH is possible if the auxiliary variable correlates with CATCH. By the statements above, the methods introduced in Chapter 4 are promising.

We apply first the estimation strategy SRSWOR_POST, where the original sample has been drawn by SRSWOR and the adjustment for nonresponse relies on post-stratification. As a second alternative we apply SRSWOR_REG strategy by regression estimation with GT as the auxiliary variable. The strategy SRSWOR_HT serves as reference strategy. We compare the results with the reference strategy by computing coefficient of variation and design effect estimates.

The five-category variable POST5 is chosen for post-stratification. We execute the estimation of the post-stratified estimate $\hat{t}_{POST}$ with the SAS procedure SURVEYMEANS. In regression estimation we use the procedure SURVEYREG for the estimation of CATCH total by the estimator $\hat{t}_{REG}$. Results are summarized in Table 5.2. Totals estimated by SURVEYREG and R `survey` functions `postStratify` and `svyglm` are equal but standard errors differ somewhat because of slightly different variance estimators.

Table 5.2 Estimated totals, standard errors and coefficients of variation for CATCH under strategies SRSWOR_HT, SRSWOR_POST and SRSWOR_REG computed for the complete data set of size $n = 20$ vessels and incomplete data set of $n = 18$ vessels.

Strategy	**Variable**	**n**	**Sum of Weights** w_k	**Total** $\hat{t}$	**Std Dev** $s.e(\hat{t})$	**Coeff of Variation** $cv(\hat{t})$
a) Full sample estimates (no nonresponse)						
SRSWOR_HT	CATCH	20	100	610603	54439	0.089
b) Incomplete data, two missing values, no adjustment						
SRSWOR_HT	CATCH	18	90	543997	54466	0.100
c) Incomplete data, adjusted by post-stratification with POST5						
SRSWOR_POST	CATCH	18	90	564206	40894	0.072
d) Incomplete data, adjusted by regression estimation with GT						
SRSWOR_REG	CATCH	18	90	647368	50166	0.077

Regression estimation under strategy SRSWOR_REG (part d in the table) adjusted successfully the nonresponse bias in the HT based strategy SRSWOR_HT (part b) for the total estimate of CATCH. Post-stratification with variable POST5 (part c) was not successful in this case. Results on cv:s in parts b) and d) in the table indicate that the strategy SRSWOR_REG improved the efficiency of the estimation of the total of CATCH.

In this favourable situation we did know the response mechanism completely, because it was created by ourselves. We wanted to demonstrate the power of a nonresponse adjustment technique when having access to an auxiliary variable that correlates strongly with the target variable: corr(CATCH,GT) = 0.56 in the population. In practice, the process that creates missingness in a sample survey is unknown. Therefore, it is important in the preparation of the sampling frame to include various potential auxiliary variables in the frame and further, in the data preparation phase to search for potential auxiliary data from official statistics and other reliable sources. In both cases, the original sample data set must contain the values of the auxiliary variables that are planned to be used in the analysis phase.

Simulation experiment. We carried out a small simulation experiment in order to verify the capacity of the applied nonresponse adjustment method in the reduction of the bias due to nonresponse in the case considered. We generated unit nonresponse in the population data set SIMPOP with the following scenario. A response mechanism dependent on the auxiliary variable GT was defined such that the probability of non-response for

variable CATCH was higher for larger values of GT than for smaller values. GT was chosen because the values were known for all population vessels. Nonresponse within SIMPOP was generated by Poisson PPS sampling with two expected non-respondent cases in a Poisson sample of size $E(n_s) = 20$ elements. We then drew $K = 100$ independent SRSWOR samples of size $n = 20$ vessels from SIMPOP, computed the estimated total, standard error and coefficient of variation for the regression estimator $\hat{t}_{REG}$ of CATCH total under the SRSWOR_REG strategy and the HT estimator $\hat{t}_{HT}$ under the SRSWOR_HT strategy from each sample. Finally, the averages of the desired statistics over the 100 samples were computed. The results are in Table 5.2.

The average of the estimated CATCH totals computed over the 100 incomplete samples by HT estimation with no adjustment for nonresponse (case a) is much smaller than the true value. This indicates serious negative bias (too small value) for the HT total estimate because of the informative nonresponse i.e. the correlation between target variable CATCH and the response mechanism. Regression estimation with GT as the auxiliary variable (case b) shows that the method adjusted effectively the nonresponse bias: after adjustment the average of total estimates was close to the true value. When comparing with results for the full sample (case c), it is noted that the REG method also is efficient; coefficients of variation (7% and 6%) are close. This is due to the significant correlation between CATCH and GT.

In the simulation experiment we had a strong auxiliary variable GT at our disposal. The piece of auxiliary data incorporated in regression estimation was the known population total of GT. The adjustment by regression estimation appeared to reduce substantially the nonresponse bias that was present in the unadjusted HT estimate.

Table 5.2 Means of estimated totals, standard errors and coefficients of variation for CATCH from $K = 100$ simulated SRSWOR samples of size $n = 20$ contaminated by unit nonresponse.

Strategy	**VarName**	**Replicates**	**Averages over simulations**						
			SumWgt	**n Original**	**n Non-Missing**	**n Missing**	**Total** $\hat{t}$	**StdDev** $s.e(\hat{t})$	**CV** $cv(\hat{t})$
a) Unadjusted estimates									
SRSWOR_HT	CATCH	100	80.6	20	16.12	3.88	503622	41366	0.082553
b) Estimates adjusted for nonresponse									
SRSWOR_REG	CATCH	100	80.6	20	16.12	3.88	633428	43934	0.069640
c) Full sample estimates (no nonresponse)									
SRSWOR_REG	CATCH	100	100	20	20	0	625882	37931	0.060743
True total							624036		

6 Analysis of economic variables

6.1 Estimation strategies

This chapter concentrates on the analysis of selected economic variables under several estimation strategies for the SIMPOP population. Variables are VALUE, TOTAL_COSTS and LABOR. We discuss strategies where the auxiliary data are incorporated in the sampling design or, alternatively, in the estimation design. The main auxiliary variable is GT (vessel tonnage), whose values are taken from the sampling frame. Variables DAS (days at sea) and DOM01 (type of fishing) are additional auxiliary variables whose population totals are assumed available.

For VALUE, TOTAL_COSTS and LABOR, we operate with the SIMPOP population of $N = 100$ active vessels. We use the entire SIMPOP of $N = 120$ vessels when dealing with the variable ACTIVITY. Single-sample realizations of the strategies are considered first and are supplemented by small simulation experiments for multiple samples.

The following estimation strategies are applied.

Table 6.1 Strategies for the analysis of economic variables.

	Strategy	**Sampling design**	**Estimation design**
(1)	SRSWOR_HT	Simple random sampling without replacement	HT estimation
(2)	PPS_WOR_HT	PPS without replacement sampling GT as size variable	HT estimation
(3)	SRSWOR_RAT	SRSWOR	Ratio estimation GT as auxiliary variable
(4)	SRSWOR_REG	SRSWOR	Regression estimation GT as auxiliary variable
(5)	SRSWOR_REG	SRSWOR	Regression estimation GT, DAS and DOM01 as auxiliary variables

The first strategy is the reference strategy. In SRSWOR, inclusion probabilities are constants. PPS_SYS with GT as the size variable produces larger (measured in GT) inclusion probabilities for large vessels and smaller for small vessels.

Note that essentially, the same auxiliary information is supplied for strategies (2) to (4), but in different ways. In the PPS_WOR strategy (2), the auxiliary data are incorporated in the sampling design. A single size variable only is allowed. GT values are required for all population vessels.

Strategies (3), (4) and (5) rely on model-assisted ratio and regression methods under simple random sampling. In these methods, auxiliary data are incorporated in the estimation design; the sampling phase does not involve any auxiliary information. Strategies (3) and (4) use a single auxiliary variable (GT), whereas the strategy (5) uses three auxiliary variables: GT, DAS and DOM01. The important option of several auxiliary variables indicates the flexibility of model-assisted strategies. The population totals of these variables constitute the auxiliary data needed. Table 6.2 contains the auxiliary variable totals for the model-assisted methods.

Table 6.2 Auxiliary totals for model-assisted estimators.

GT t_{x_1}	**DAS** t_{x_2}	**DOM01** t_{x_3}
32896.4	18308	30

6.2 Variable VALUE

6.2.1 Study setting

The variable VALUE describes the total value of landings (in Euro) during the reference period. We examine the performance of strategies of Table 6.1 for the estimation of the population total of VALUE. Measurements for variable VALUE are obtained from samples of different size. Sample sizes $n = 5$ and $n = 20$ vessels are used first for the single sample cases and then for multiple samples generated by simulation experiments.

6.2.2 Efficiency comparison

Strategies of Table 6.1 are applied for SRSWOR and PPSWOR samples drawn from the SIMPOP population of active vessels. Table 6.3 presents results for the various strategies. Our main interest in the efficiency of each strategy, measured by coefficient of variation (cv = StdDev/Total) of an estimated total of VALUE.

Clear differences in efficiency are observed between the methods. The reference SRSWOR_HT strategy (1) shows largest coefficient of variation (cv), for both sample sizes, as expected. Incorporation of auxiliary information, either in the sampling design or in the estimation design, tends to improve efficiency. For samples of size $n = 5$, PPS_WOR_HT with $cv = 13.5\%$ shows best precision. For sample size $n = 20$, differences between strategies (2)-(4) are minor. Of these strategies, PPSWOR_HT is not anymore the best strategy; the model-assisted strategies (3) and (4) show better precision. In these strategies, a single auxiliary variable GT is used. The model-assisted regression strategy (5), SRSWOR_REG under a SRSWOR sample, incorporates the three auxiliary variable totals of Table 6.2 in the estimation procedure. This strategy attains efficiency of $cv = 3.3\%$, which is much smaller than cv:s of the other strategies.

Table 6.3 Estimation results for VALUE under five estimation strategies.

Variable	n	Sum of Weights	Total $\hat{t}$	Std Dev $s.e(\hat{t})$	95% CL		Coeff of Var $cv(\hat{t})$
Sample size *n* = 5							
(1) SRSWOR_HT	5	100.000000	255818777	68043670	66899263	444738291	0.265984
(2) PPSWOR_HT	5	104.874286	184743304	25004620	115319349	254167259	0.135348
(3) SRSWOR_RAT	5	100.000000	242130000	53522592	93528509	390730000	0.221050
(4) SRSWOR_REG	5	100.000000	211780000	51784330	68007133	355560000	0.244520
(5) SRSWOR_REG	5	100.000000	201460000	39171930	92697347	310210000	0.194440
Sample size *n* = 20							
(1) SRSWOR_HT	20	100.000000	201253849	17082556	165499648	237008050	0.084881
(2) PPSWOR_HT	20	101.573826	223968852	17827256	186655977	261281727	0.079597
(3) SRSWOR_RAT	20	100.000000	211120000	14885980	179960000	242280000	0.070509
(4) SRSWOR_REG	20	100.000000	210960000	15077603	179400000	242520000	0.071472
(5) SRSWOR_REG	20	100.000000	195200000	6383043	181840000	208560000	0.032700
True total			194676172				

6.2.3 Simulation experiments

Our results in this far are based on single sample realizations from SIMPOP. We next examine the behaviour of the strategies by a small simulation experiment. Table 6.4 contains average estimation results from $K = 100$ simulated samples for the five strategies, computed with PROC SURVEYSELECT, SURVEYMEANS and SURVEYREG.

Estimation results from simulation experiments in Table 6.4 show that the use of a single auxiliary variable GT for total estimation of VALUE in strategies (2), (3) and (4) do not improve efficiency over the reference strategy (1) that does not incorporate auxiliary information. Coefficients of variation for these methods are of similar size, in both sample sizes. Note that VALUE and GT are not strongly correlated: corr(VALUE,GT) = 0.28 in the population. Strategy SRSWOR_REG with three auxiliary variables GT, DAS and DOM01 produces smallest *cv*:s in both sample sizes: 14% for $n = 5$ and 4.7% for $n = 20$. This model-assisted strategy clearly outperforms the other strategies, in both sample sizes, confirming results from the single-sample experiments. The cost efficiency of the strategy is demonstrated by the fact that with SRSWOR_HT strategy, a sample size of 50 sample vessels would be required to attain the 4.5% efficiency of the SRSWOR_REG strategy with 20 vessels.

Table 6.4 Means of estimated totals, standard errors and coefficients of variation for five strategies for VALUE from $K = 100$ simulated samples of size $n = 5$ and $n = 20$ vessels from SIMPOP of $N = 100$ vessels.

	AuxVar	Replicates	Averages over simulations				
			SumWgt	n	Total	StdDev	CV
Sample size $n = 5$							
(1) SRSWOR_HT	none	100	100.000000	5	197242485	35762439	0.178504
(2) PPSWOR_HT	GT	100	98.901065	5	197469817	35691713	0.178674
(3) SRSWOR_RAT	GT	100	100.000000	5	194130000	34342342	0.174920
(4) SRSWOR_REG	GT	100	100.000000	5	195110000	38964736	0.198670
(5) SRSWOR_REG	GT, DAS, DOM01	100	100.000000	5	195200000	24544768	0.141860
Sample size $n = 20$							
(1) SRSWOR_HT	none	100	100.000000	20	196106848	17424625	0.088691
(2) PPSWOR_HT	GT	100	100.168244	20	196200908	17288292	0.087867
(3) SRSWOR_RAT	GT	100	100.000000	20	194270000	17198526	0.088436
(4) SRSWOR_REG	GT	100	100.000000	20	196070000	17389926	0.088600
(5) SRSWOR_REG	GT, DAS, DOM01	100	100.000000	20	196000000	9180696	0.046850
True total					194676172		

Conclusions. Over all methods in this exercise, regression estimation may be the best choice. The reasons are flexible tailoring for the purpose in the estimation phase, possibilities for improved efficiency over the other methods by using several auxiliary variables, and minimum requirements for the auxiliary variables, because the population totals of the variables only are needed.

Model -assisted methods ratio and regression estimation require an access to the auxiliary variable totals that are incorporated in the estimation procedure. These totals are often obtained from reliable sources, such as official statistics. In addition, the sample data set must contain the unit-level values of auxiliary variables. It is important that auxiliary variables and their counterparts in the sample data set are based on exactly the same definitions. Auxiliary variables are often readily available in the sampling frame. It is straightforward to obtain reliable population totals and the sample values of the variables in this case.

6.3 Variable TOTAL_COSTS

6.3.1 Study setting

The variable TOTAL_COSTS describes the total costs of fishing efforts (in Euro) during the reference period. We examine the performance of the strategies in Table 4.1 for the estimation of the population total of TOTAL_COSTS. Measurements for TOTAL_COSTS are obtained from samples of different size. Sample sizes $n = 5$ and $n = 20$ are used first for the single sample cases and then for multiple samples generated by simulation experiments.

6.3.2 Efficiency comparison

Strategies of Table 6.1 are applied for SRSWOR and PPSWOR samples drawn from the SIMPOP population of active vessels. Table 6.5 presents results for the various strategies. Our main interest in the efficiency of each strategy, measured by coefficient of variation (cv) of an estimated total of TOTAL_COST.

Table 6.5 Estimation results for TOTAL_COSTS under five estimation strategies.

Variable	n	Sum of Weights	Total $\hat{t}$	Std Dev $s.e(\hat{t})$	95% CL		Coeff of Var $cv(\hat{t})$
Sample size $n = 5$							
(1) SRSWOR_HT	5	100.000000	166160856	37627079	61691336.1	270630375	0.226450
(2) PPSWOR_HT	5	90.784295	134885883	19346484	81171432.1	188600335	0.143429
(3) SRSWOR_RAT	5	100.000000	156790000	27236195	81167924	232410000	0.173710
(4) SRSWOR_REG	5	100.000000	138920000	25121898	69165766	208660000	0.180840
(5) SRSWOR_REG	5	100.000000	130730000	15379475	88031200	173430000	0.117640
Sample size $n = 20$							
(1) SRSWOR_HT	20	100.000000	126235083	9062905	107266205	145203961	0.071794
(2) PPSWOR_HT	20	99.434022	131090256	9737997	110708393	151472119	0.074285
(3) SRSWOR_RAT	20	100.000000	132750000	7317553	117440000	148070000	0.055121
(4) SRSWOR_REG	20	100.000000	132730000	7433539	117170000	148290000	0.056006
(5) SRSWOR_REG	20	100.000000	124980000	2575597	119590000	130370000	0.020608
True total			125037964				

For TOTAL_COSTS, PPS sampling with GT as size variable and regression estimation with GT, DAS and DOM01 as the auxiliary variables in the assisting model for regression estimation show best efficiency for both sample sizes. The picture clarifies with samples of size $n = 20$ vessels. All three model-assisted estimators outperform the reference strategy as well as the PPS_WOR strategy. Strategy SRSWOR_REG with all three covariates attains best precision.

6.3.3 Simulation experiments

We examine the behaviour of the strategies by a small simulation experiment. Table 6.6 presents average estimation results from $K = 100$ simulated samples for the strategies.

For samples of size $n = 5$ vessels, the SRSWOR_REG strategy shows best efficiency; the differences between the other strategies are minor. The situation is the same for samples $n = 20$.

Table 6.6 Means of estimated totals, standard errors and coefficients of variation for five strategies for TOTAL_COST from $K = 100$ simulated samples of size $n = 5$ and $n = 20$ vessels from SIMPOP of $N = 100$ vessels.

	AuxVar	Replicates	Averages over simulations				
			SumWgt	n	Total	StdDev	CV
Sample size n = 5							
(1) SRSWOR_HT	none	100	100.000000	5	126782133	18020029	0.139668
(2) PPSWOR_HT	GT	100	98.901065	5	126375410	17016332	0.132446
(3) SRSWOR_RAT	GT	100	100.000000	5	125230000	16068656	0.126050
(4) SRSWOR_REG	GT	100	100.000000	5	124950000	18149866	0.143670
(5) SRSWOR_REG	GT, DAS, DOM01	100	100.000000	5	124540000	8687201	0.072160
Sample size n = 20							
(1) SRSWOR_HT	none	100	100.000000	20	125560582	8921162	0.070776
(2) PPSWOR_HT	GT	100	100.168244	20	125845732	8372180	0.066279
(3) SRSWOR_RAT	GT	100	100.000000	20	124980000	8331410	0.066364
(4) SRSWOR_REG	GT	100	100.000000	20	125470000	8380551	0.066570
(5) SRSWOR_REG	GT, DAS, DOM01	100	100.000000	20	125390000	3944847	0.031403
True total					125037964		

Conclusion. The picture for TOTAL_COST seems pretty similar with the variable VALUE. This is explained by the high correlation of the two variables (0.98) and by the fact that their correlations with GT and DAS are reasonable large (Table 3.5).

6.4 Variable LABOR

6.4.1 Study setting

The variable LABOR describes the total costs of labour force (in Euro) during the reference period. We examine the performance of the strategies in Table 6.1 for the estimation of the population total of LABOR. Measurements for LABOR are obtained from samples of different size. Sample sizes $n = 5$ and $n = 20$ are used first for the single sample cases and then for multiple samples generated by simulation experiments.

6.4.2 Efficiency comparison

Strategies of Table 6.1 are applied for SRSWOR and PPSWOR samples drawn from the SIMPOP population of active vessels. Table 6.7 presents results for the various strategies. Our main interest in the efficiency of each strategy, measured by coefficient of variation (cv) of an estimated total of LABOR.

Table 6.7 Estimation results for LABOR under five estimation strategies.

Variable	n	Sum of Weights	Total $\hat{t}$	Std Dev $s.e(\hat{t})$	95% CL		Coeff of Var $cv(\hat{t})$
Sample size $n = 5$							
(1) SRSWOR_HT	5	100.000000	51053745	14399614	11074006.5	91033482.6	0.282048
(2) PPSWOR_HT	5	90.784295	46421214	7762403	24869327.4	67973100.2	0.167217
(3) SRSWOR_RAT	5	100.000000	48239631	11691221	15779598.0	80699664.0	0.242360
(4) SRSWOR_REG	5	100.000000	42497390	11704792	9999678.0	74995103.0	0.275420
(5) SRSWOR_REG	5	100.000000	41559131	10142763	13398306.0	69719956.0	0.244060
Sample size $n = 20$							
(1) SRSWOR_HT	20	100.000000	42716843	3818528	34724572.8	50709114.1	0.089392
(2) PPSWOR_HT	20	99.434022	44588485	4444773	35285467.3	53891502.1	0.099684
(3) SRSWOR_RAT	20	100.000000	44761805	3407868	37629055.0	51894554.0	0.076133
(4) SRSWOR_REG	20	100.000000	44716084	3455566	37483501.0	51948666.0	0.077278
(5) SRSWOR_REG	20	100.000000	41207347	1765763	37511563.0	44903132.0	0.042851
True total			40914264				

For samples of size $n = 5$, the best strategy in efficiency is PPSWOR_HT, where the sample is drawn using PPS without replacement sampling with GT as the size variable. The model-assisted strategies ratio and regression estimation under SRSWOR sampling did not improve precision relative to the reference strategy SRSWOR_HT. For samples of size $n = 20$, the situation changes so that the model-assisted strategies outperform the reference strategy in efficiency, notably for regression estimation with three auxiliary variables.

6.4.3 Simulation experiments

We examine the behaviour of the strategies by a small simulation experiment. Table 6.6 presents average estimation results from $K = 100$ simulated samples for the strategies.

Table 6.8 Means of estimated totals, standard errors and coefficients of variation for five strategies for LABOR from $K = 100$ simulated samples of size $n = 5$ and $n = 20$ vessels from SIMPOP of $N = 100$ vessels.

	AuxVar	Replicates	Averages over simulations				
			SumWgt	n	Total	StdDev	CV
Sample size n = 5							
(1) SRSWOR_HT	none	100	100.000000	5	41376736	8224349	0.196836
(2) PPSWOR_HT	GT	100	98.901065	5	41529430	8227161	0.197577
(3) SRSWOR_RAT	GT	100	100.000000	5	40635354	8061915	0.198120
(4) SRSWOR_REG	GT	100	100.000000	5	41142233	9155431	0.222410
(5) SRSWOR_REG	GT, DAS, DOM01	100	100.000000	5	41303501	6904558	0.211300
Sample size n = 20							
(1) SRSWOR_HT	none	100	100.000000	20	41301541	3917383	0.094881
(2) PPSWOR_HT	GT	100	100.168244	20	41219210	3926993	0.095134
(3) SRSWOR_RAT	GT	100	100.000000	20	40787728	3922377	0.096401
(4) SRSWOR_REG	GT	100	100.000000	20	41315348	3974518	0.096323
(5) SRSWOR_REG	GT, DAS, DOM01	100	100.000000	20	41323075	2350461	0.057016
True total					40914264		

Conclusion. For samples of size $n = 5$ vessels , none of the methods that incorporate auxiliary information either in the sampling design with PPS sampling or with model-assisted methods in the estimation design do not improve precision over the reference strategy. For the last method with three auxiliaries in the model, the small sample size seems to become too small for reliable estimation because the estimates may become instable. The correlation of LABOR with GT is the weakest (0.22) among the target variables VALUE and TOTAL_COST and this might explain at least partly the results. For samples with size $n = 20$ the picture changes so that regression estimation with the three auxiliary variables GT, DAS and DOM01 appears most efficient with an average coefficient of variation of 5.7%, when compared with the other strategies of efficiency about 9.5%.

6.5 Variable ACTIVITY

6.5.1 Study setting

The binary variable ACTIVITY describes whether a vessel has been active in fishing (=1) or not (=0) during the reference period. In this exercise, ACTIVITY is taken as one of our target variables whose values have been measured from the sample vessels in the survey. We now operate with the entire SIMPOP population of N = 120 vessels. The following estimation strategies are applied.

Table 6.9 Strategies for the estimation of the total number of active vessels.

	Strategy	**Sampling design**	**Estimation design**
(1)	SRSWOR_HT	Simple random sampling without replacement SIMPOP N=120 sorted in random order	HT estimation
(2)	SYS_HT	Systematic sampling SIMPOP N=120 sorted by GT	HT estimation
(3)	PPS_SYS_HT	Systematic PPS sampling SIMPOP N=120 sorted in random order GT as size variable	HT estimation
(4)	SRSWOR_RAT	SRSWOR SIMPOP N=120 sorted in random order	Ratio estimation GT as auxiliary variable
(5)	SRSWOR_REG	SRSWOR SIMPOP N=120 sorted in random order	Regression estimation GT as auxiliary variable

The first strategy is the reference strategy. In strategies (2) and (3), the auxiliary data are incorporated in the sampling design. Strategies (3) and (4) rely on model-assisted ratio and regression methods under simple random sampling. In these methods, auxiliary data are used in the estimation phase; the sampling phase does not involve any auxiliary information.

In SRSWOR and SYS sampling, inclusion probabilities are constants. PPS_SYS with GT as the size variable produces larger (measured in GT) inclusion probabilities for large vessels and smaller for small vessels. Sample size n = 20 is used first for the single sample realizations and then, for multiple samples obtained by simulation.

6.5.2 Efficiency comparison

Results for ACTIVITY total from the single sample experiment are collected in Table 6.10. Strategy (1) is the reference strategy; no auxiliary data are used. Sorting the population frame by GT followed by systematic sampling in strategy (2) does not improve precision. Sorting is often used for good coverage over the population in a systematic sample. The realized samples in (1) and (2) are different, but the estimates are equal. This is because there are exactly two non-active observations in both samples. PPS_SYS_HT seems not to be a good choice in this case: coefficient of variation for (3) is larger than for the other strategies.

Model-assisted ratio and regression estimation strategies for the SRSWOR sample in strategy (1) are well competitive with the reference strategy. The benefit in (4) and (5) is that aggregate-level auxiliary data (population total of GT) only are needed, whereas in (2) and (3), unit-level auxiliary data are required. Ratio estimation SRSWOR_RAT gives best accuracy in this experiment.

Table 6.10 Estimation results for ACTIVITY under five different estimation strategies for samples of size $n = 20$ elements.

Strategy	n	Sum of Weights	Total $\hat{t}$	Std Dev $s.e(\hat{t})$	95% CL		Coeff of Var $cv(\hat{t})$
(1) SRSWOR_HT	20	120.000	108.00000	7.539370	92.2199165	123.780083	0.069809
(2) SYS_HT	20	120.000	108.00000	7.539370	92.2199165	123.780083	0.069809
(3) PPS_SYS_HT	20	122.687	104.12294	9.714304	83.7906707	124.455216	0.093296
(4) SRSWOR_RAT	20	120.000	105.16000	6.721000	91.0914000	119.230000	0.063913
(5) SRSWOR_REG	20	120.000	107.81000	7.696500	91.7039000	123.920000	0.071387
True total			100				

6.5.3 Simulation experiments

The results above only consider the single sample realizations from SIMPOP. Let us examine closer the behaviour of the strategies by a small simulation experiment. Table 6.11 contains average estimation results from $K = 100$ simulated samples for the five strategies, computed with PROC SURVEYSELECT, SURVEYMEANS and SURVEYREG.

Ratio estimation for a SRSWOR sample is of the best in efficiency and systematic PPS sampling is the worst. However, differences between the methods are minor. Note that essentially, the same auxiliary information was supplied for strategies (2) to (5), but in different ways. In (2) and (3), GT values are required for all population vessels, but in (4) and (5), population total of GT only is needed. This fundamental difference indicates the flexibility of model-assisted strategies.

Table 6.11 Means of estimated totals, standard errors and coefficients of variation for five strategies for ACTIVITY from $K = 100$ simulated samples of size $n = 20$ vessels from SIMPOP of $N = 120$ vessels.

	AuxVar	Replicates	Averages over simulations				
			SumWgt	n	Total	StdDev	CV
(1) SRSWOR_HT	none	100	120.000000	20	99.780000	8.851161	0.091158
(2) SYS_HT	none	100	120.000000	20	100.680000	8.695351	0.088989
(3) PPS_SYS_HT	GT	100	120.345357	20	100.955378	8.954131	0.093647
(4) SRSWOR_RAT	GT	100	120.000000	20	98.749600	8.556004	0.088489
(5) SRSWOR_REG	GT	100	120.000000	20	100.820000	8.830300	0.090218
True total					100		

Conclusion. Results for ACTIVITY do not follow the same pattern than those for the other target variables of this chapter. A reason might be a different correlation structure of ACTIVITY with the auxiliary variables (Table 6.12). It is noted that ACTIVITY correlates weakly only with all three auxiliary variables. The strongest correlation is with GT, the auxiliary variable used in PPS sampling and model-assisted methods. This example

thus demonstrates that it is important to invest efforts to search and test for the suitability of various auxiliary variables for a given estimation task.

Table 6.12. Correlation of ACTIVITY with auxiliary variables LENGTH, GT and kW (all vessels, $N = 120$).

Pearson Correlation Coefficients			
	LENGTH	**GT**	**kW**
ACTIVITY	0.10497	0.09685	0.08820

It should be noted that we are here working with a binary target variable. This is somewhat problematic from modeling point of view, because the assisting models in strategies (4) and (5) are based on a linear regression model. A logistic model would be a better justified model formulation for this case, suggesting the more general GREG (generalized regression) family estimators (see e.g. Lehtonen & Veijanen 2009). Numerically, however, the results might not change much because the mean (0.37) of ACTIVITY is not so far from 0.5.

6.6 Conclusions

Four important target variables were analysed under a variety of typical study settings in fisheries statistics. The aim was to examine to what extent it is possible to improve statistical efficiency of total estimates of the selected economic variables by using auxiliary information on the vessel population in the sampling and estimation phases. Simulation experiments were conducted to supplement the single-sample analyses.

We discussed strategies where the auxiliary data were incorporated in the sampling design or, alternatively, in the estimation design. The main auxiliary variable was GT (vessel tonnage), whose values were taken from the sampling frame. Variables DAS (days at sea) and DOM01 (type of fishing) were additional auxiliary variables whose population totals were assumed available. Strategies were SRS without replacement, systematic sampling and PPS without replacement sampling using GT as the size variable, where Horvitz-Thompson (HT) estimation design was used, and ratio and regression estimation design for a SRSWOR sample. In regression estimation, the case of three auxiliary variables was demonstrated in addition to the single covariate case.

Over all strategies applied for the target variable VALUE, regression estimation may be the best choice. The reasons are flexible tailoring for the purpose in the estimation phase, possibilities for improved efficiency over other methods by using several auxiliary variables, and minimum requirements for the auxiliary variables, because the population totals of the variables only are needed as auxiliary data. Auxiliary variable totals that are needed in ratio and regression estimation are often obtained from reliable sources, such as official statistics. In addition, the sample data set must contain the unit-level values of auxiliary variables. It is important that auxiliary variables and their counterparts in the sample data set are based on exactly the same definitions. Auxiliary variables are often readily available in the sampling frame, and it is straightforward to obtain reliable population totals and the sample values of the variables in this case. If the auxiliary variables are obtained from different sources, it is important to examine the quality of sources in order to avoid the possible bias in estimates because of measurement errors.

The picture for TOTAL_COST seems pretty similar with the variable VALUE. This is explained by the high correlation of the two variables (0.98) and by the fact that their correlations with GT and DAS are reasonable large (Table 3.5).

For the target variable LABOR, with samples of size $n = 5$ vessels , none of the methods that incorporate auxiliary information either in the sampling design with PPS sampling or with model-assisted methods in the estimation design do not improve precision over the reference strategy. For the last method with three auxiliaries in the model, the small sample size seems to become too small for reliable estimation because the estimates may become instable. The correlation of LABOR with GT is the weakest (0.22) among the target variables VALUE and TOTAL_COST and this might explain at least partly the results. For samples with size $n = 20$ the picture changes so that regression estimation with the three auxiliary variables GT, DAS and DOM01 appears most efficient with an average coefficient of variation of 5.7%, when compared with the other strategies of efficiency about 9.5%.

For the economic variables analysed here it seems that model-assisted methods, regression estimation with a multivariate model in particular, might offer a reasonable choice with respect to efficiency of estimation, when compared with the other methods.

Results for ACTIVITY do not follow the same pattern than those for the economic target variables of this chapter. A reason might be a different correlation structure of ACTIVITY with the auxiliary variables (Table 6.12). It is noted that ACTIVITY correlates weakly only with the auxiliary variables LENGTH, GT and kW. The strongest correlation is with GT, the auxiliary variable used in PPS sampling and model-assisted methods. Moreover, as a binary variable ACTIVITY might require a different model formulation than the linear model for the model-assisted methods. This example also demonstrates that it is important to invest efforts to search and test for the suitability of various auxiliary variables for a given estimation task.

7 General conclusions

The aim of the handbook was to introduce modern survey analysis methodology for the purposes of fisheries statistics and to examine to what extent it is possible to improve statistical efficiency of total estimates by using auxiliary information and tools of statistical modeling. Efficiency was measured by coefficient of variation of total estimate. Coefficient of variation is defined as the ratio of standard error estimate of the total with the total estimate itself. The measure is scale-free and suits well for comparing total estimates. Simulation experiments were conducted to supplement the single-sample analyses.

Our conceptual framework was build on the concepts of sampling design and estimation design, defining the estimation strategy for the survey. We discussed two types of estimation strategies. In the first type, the auxiliary information was introduced in the sampling design by probability proportional to size (PPS) sampling or stratified sampling, and the estimation design relied on the traditional expansion or HT estimation. Auxiliary variables for these strategies such as PPS size variable (e.g. vessel tonnage) and stratification variables had to be available in the frame population. This type of strategies are common in fisheries statistics. Their strengths are long tradition in official statistics, firm theoretical basis, technical simplicity in the sampling and estimation phases, availability of reliable software, and improved statistical efficiency over the reference strategy for variables for which the sampling design was optimized. We demonstrated in chapters 3 and 6 the ways to use auxiliary information in the sampling phase. Simple random sampling was used as the reference strategy. Because methods based on sample weighting are used in the estimation phase, the approach is easily implemented in the production work.

Weaknesses of the strategy are the possible lack of unit-level auxiliary variables in the sampling frame that are powerful enough for efficiency improvement for the desired set of target variables in PPS or stratified sampling, the possibility of low benefit in efficiency for some target variables, and a risk of method failure for some target variables, e.g. under PPS sampling. In stratification, multiple stratification variables can be used, but in PPS sampling, a restriction is that a single auxiliary variable only can be used as size variable. In this approach, the main investment of efforts is in the sampling phase of the survey, when constructing a high quality frame population rich enough of variables for sampling purposes.

The second type of strategies comprise methods that use the auxiliary information in the estimation phase, under a given sampling design. This set includes several traditional design-based model-assisted estimators, which incorporate the auxiliary data in the estimation procedure of a total. In the handbook, we focused on ratio and regression estimation and post-stratification, both based on linear fixed-effects regression models. Requirements for auxiliary information are different to the first type of strategies. In model-assisted methods, the population or sub-population (domain) totals of the auxiliary variables, or population distributions of categorical auxiliary variables, are required, and their unit-level measurements are needed in the sample data set. Typically, the sampling design is a compromise design involving for example simple random sampling without replacement, possibly supplemented with stratification of the population and an appropriate allocation scheme. Stratification can be applied for example for compromise sample allocation schemes that meet precision requirements for domains both with small and large sample sizes. This type of strategies would provide useful options in fisheries statistics. Benefits of the approach are flexibility so that estimation designs can be tailored efficient for a set of diverse target variables if desired, the use of multivariate assisting models with several auxiliary variables, and the fact that aggregate auxiliary data only are needed for the model-assisted estimators. We demonstrated these properties in chapters 4 and 6.

It is important that the auxiliary variables and their sample counterparts are based on exactly the same definitions. This is possible if the data source is the same for both the auxiliary and sample data, for example a statistical register. If the auxiliary variables are obtained from different sources, it is important to examine the quality of sources in order to avoid the possible bias in estimates because of different measurement and possible measurement errors. The tailoring approach might increase staff expertise requirements, because good capabilities for statistical modelling are important. This might not be a problem if high-level statisticians are available in the agency. In this approach, the main investment of efforts is in the estimation phase of the survey. To ensure success in the estimation phase, care must be taken to have access to either high quality aggregate auxiliary data or (even better), to have rich selection of auxiliary variables readily available in the frame population data set, taken from statistical register and other reliable sources.

Supplementing the tailoring approach, a compromise estimation strategy is often adopted in routine official statistics production work. Because model-assisted estimators discussed in the handbook can be expressed as calibration estimators, an overall strategy for production purposes can be introduced by creating multi-purpose calibrated weights for a large set of target variables of a survey. A reasonable set of several auxiliary variables can then be included in the calibration machinery. The calibrated weights applied to the sample values of the auxiliary variables reproduce the known population (or domain) totals or distributions. This property of *coherence* is often appreciated in official statistics. A calibrated estimator for the total of a given target variable will be more precise than an estimator for the reference SRS strategy, if some of the auxiliary variables in the calibration apparatus correlate with the target variable. This property can also be used in adjusting for the possible selection bias because of unit nonresponse, if some of the auxiliaries correlate with the target variable. This was demonstrated in Chapter 5. Because the methodology is based on weighting (with multipurpose calibration weights) and thus resembles the first type weighting strategies, the staff requirements also are similar. Weaknesses of this approach might be the lack of suitable and powerful auxiliary variables for efficiency improvement for a large set of target variables, overly complicated model formulation and over-fitting, and the possible lack of careful model diagnostics.

The handbook also offers materials for considering sample size determination for a survey. The budget restrictions and the adopted sampling and estimation strategies set the framework for sample size optimization. With clever use of auxiliary information in the sampling and estimation phases, it is possible to attain the precision requirements with a smaller sample size, when compared with a strategy that relies solely on simple random sampling and related estimation, leading to improved cost efficiency.

Survey quality is a complex phenomenon relating to all stages of the survey process. In the context of the European Statistical System, a number of quality criteria have been defined (see e.g. https://ec.europa.eu/eurostat/web/quality/). The framework of total quality management provides a useful approach for assessing the overall quality of a survey. Biemer and Lyberg (2003) define the goal of survey quality management as finding a balance between different error components so that the total survey error is as small as possible while considering the costs of improvements in different stages and the size of the budget. For the total survey error framework in practice see e.g. Biemer et al. (2017).

In the handbook we have concentrated on the measurement and improvement of *accuracy* of the survey results, which is one of the most important quality criteria. In this context, the quality assessment and improvement of the sources of sample data and auxiliary data is crucial. This aspect is becoming increasingly important in the era of diverse, sometimes of poor quality, data sources becoming available and used for official statistics.

8 Case studies

8.1 Italy

8.1.1 Introduction

The Italian case study represents a designed application of multivariate allocation in stratified PPS sampling and estimation for population subgroups.

The sample unit is the single vessel and this unit is selected from the Vessel Register, which also represents the frame population.

The sampling is of a stratified nature in that the fishing vessels of the fleet are divided into homogenous groups based on suitable variables and independent samples are taken from each of these strata (see Section 3.6).

The strata guarantee, as far as possible, that the vessels are homogeneous in terms of productive characteristics and socio-economic structure. For this reason, the criterion for delineating the strata as homogeneously as possible is based on the following three variables:

- Stratification variable 1: geographical (e.g. FAO Geographical Sub Areas)
- Stratification variable 2: technical (e.g. prevalent fishing technique)
- Stratification variable 3: dimensional (e.g. length of vessel)

Stratification variables 1 and 3 are available in the sampling frame. Information on the prevalent fishing activity (stratification variable 2) come from field surveys carried out periodically since the implementation of the DCF and updated every quarter. In fact, more than 70% of the Italian fishing-vessel licences allow the use of more than one fishing system. The existence or otherwise of actual polyvalent activity have to be verified through analysis of information on logbooks and field interviews. This survey involves all the vessels in the fleet register, including those less than 12 meters.

8.1.2 Multivariate allocation of sampling units

The multivariate allocation method is implemented in the MAUSS-R software developed at ISTAT as described in https://www.istat.it/it/files/2011/02/user_and_methodological_manual.pdf

The optimum sample number per stratum is defined according to Bethel's procedure (1989), the vessels are selected using PPS methodology (Probability Proportional to Size) by applying the algorithm of Hanurav-Vijayan. Bethel's procedure (1989) is a mathematical algorithm to achieve the optimum sample allocation in a multivariate sample survey. Bethel and Hanurav-Vijayan PPS methods are reported and explained in sections 3.5 and 3.6 of the main handbook text.

A numerical example on the application of MAUSS-R for a specific target variable is reported in the following tables:

Table 8.1 – Input file_1 for MAUSS-R

- The first column (stratum) includes the codes for the stratum, defined as explained before: DM1 (geographical subarea), DM2 (administrative region), DM3 (prevalent fishing technique, that in this case is "dredgers – DRB"), DOM4 (vessel length classes), DOM5-DOM7 different aggregations of stratification variables.
- N = number of units in the frame population
- M1 and M2 = average sample values for the variables "fuel costs" (M1) and "labour costs" (M2)
- S1 and S2 = standard deviation of the sampling values for the variables "fuel costs" (S1) and "labour costs" (S2)
- Cost = fieldwork costs in the stratum (cost per each interview). The variable cost of each stratum is set equal to one because there is no difference in cost between the different strata
- Cens = 0, that is all the strata should be sampled (0 to be sampled, 1 otherwise).

Table 8.2 - Input file_2 for MAUSS-R

The second file contains the constraints on sampling errors. In this specific example, we set a constraint of 20% for both variables at the domain level DOM5 (GSA+fishing technique+vessel lengt class) and a constraint of 4% for both variables at the domain level DOM7 (total segment at national level, that is actually the segmentation required by EUMAP[1]).

Table 8.3 – Output file for MAUSS-R

The system produces as output the sample size per stratum as reported in Table 3. Using this tool the user is able to make the necessary adjustments to achieve the desired sample size or, conversely, to achieve the desired expected precision on target estimates.

The manual to use MAUSS-R tool is available at the following web page:

https://www.istat.it/en/methods-and-tools/methods-and-it-tools/design/design-tools/mauss-r#Softwareanddocumentation-2.

[1] Commission Implementing Decision n. 2016/1251 adopting a multiannual Union programme for the collection, management and use of data in the fisheries and aquaculture sectors for the period 2017-2019

Table 8.1 – Input file_1 for MAUSS-R

STRATUM	DOM1	DOM2	DOM3	DOM4	DOM5	DOM6	DOM7	N	M1	S1	M2	S2	Cost	Cens
1	9	LAZIO	DRB	VL1218	9DRBVL1218	DRBVL1218	1	24	2666	2131	14725	7555	1	0
30	10	CAMPANIA	DRB	VL1218	10DRBVL1218	DRBVL1218	1	14	6356	2279	12379	6157	1	0
65	17	ABRUZZO	DRB	VL1218	17DRBVL1218	DRBVL1218	1	82	2307	713	13587	3907	1	0
66	17	ABRUZZO	DRB	VL1218	17DRBVL1218	DRBVL1218	1	20	2307	713	13587	3907	1	0
73	17	E.ROMAGNA	DRB	VL1218	17DRBVL1218	DRBVL1218	1	18	5379	1369	21108	9582	1	0
74	17	E.ROMAGNA	DRB	VL1218	17DRBVL1218	DRBVL1218	1	36	7488	913	54880	5577	1	0
84	17	F.V.GIULIA	DRB	VL1218	17DRBVL1218	DRBVL1218	1	22	7930	851	14512	13494	1	0
85	17	F.V.GIULIA	DRB	VL1218	17DRBVL1218	DRBVL1218	1	20	7190	3	26205	1	1	0
92	17	MARCHE	DRB	VL1218	17DRBVL1218	DRBVL1218	1	74	8929	1039	31636	16912	1	0
93	17	MARCHE	DRB	VL1218	17DRBVL1218	DRBVL1218	1	23	5883	923	21626	2242	1	0
94	17	MARCHE	DRB	VL1218	17DRBVL1218	DRBVL1218	1	65	8269	1568	57668	13264	1	0
95	17	MARCHE	DRB	VL1218	17DRBVL1218	DRBVL1218	1	58	13270	317	29015	1022	1	0
104	17	MOLISE	DRB	VL1218	17DRBVL1218	DRBVL1218	1	10	4263	410	7922	477	1	0
110	17	VENETO	DRB	VL1218	17DRBVL1218	DRBVL1218	1	57	8150	824	28281	126	1	0
111	17	VENETO	DRB	VL1218	17DRBVL1218	DRBVL1218	1	24	6865	665	24594	5791	1	0
112	17	VENETO	DRB	VL1218	17DRBVL1218	DRBVL1218	1	49	5857	2	15916	8	1	0
113	17	VENETO	DRB	VL1218	17DRBVL1218	DRBVL1218	1	34	7212	19	29925	11	1	0
123	18	PUGLIA Nord	DRB	VL1218	18DRBVL1218	DRBVL1218	1	25	2416	361	18977	2268	1	0
124	18	PUGLIA Nord	DRB	VL1218	18DRBVL1218	DRBVL1218	1	50	7655	2940	12348	8326	1	0

Table 8.2 - Input file_2 for MAUSS-R

DOM	CV1	CV2
DOM1	1	1
DOM2	1	1
DOM3	1	1
DOM4	1	1
DOM5	0.2	0.2
DOM6	1	1
DOM7	0.04	0.04

Table 8.3 – Output file for MAUSS-R

STRATUM	DOM1	DOM2	DOM3	DOM4	n
1	9	LAZIO	DRB	VL1218	10
30	10	CAMPANIA	DRB	VL1218	5
65	17	ABRUZZO	DRB	VL1218	3
66	17	ABRUZZO	DRB	VL1218	2
73	17	E.ROMAGNA	DRB	VL1218	2
74	17	E.ROMAGNA	DRB	VL1218	2
84	17	F.V.GIULIA	DRB	VL1218	2
85	17	F.V.GIULIA	DRB	VL1218	2
92	17	MARCHE	DRB	VL1218	8
93	17	MARCHE	DRB	VL1218	2
94	17	MARCHE	DRB	VL1218	6
95	17	MARCHE	DRB	VL1218	2
104	17	MOLISE	DRB	VL1218	3
110	17	VENETO	DRB	VL1218	2
111	17	VENETO	DRB	VL1218	2
112	17	VENETO	DRB	VL1218	2
113	17	VENETO	DRB	VL1218	2
123	18	PUGLIA Nord	DRB	VL1218	2
124	18	PUGLIA Nord	DRB	VL1218	4
					63

8.1.3 Random selection of sampling units

The sample survey is repeated every year applying a panel survey with a 20% turnover rate.

In the rotated samples the units to be observed are formed by replacing some statistical units in turn with randomly selected others. Often a rotation of the units in the sampling strategy is introduced with the purpose of limiting the cost of field work, reducing the amount of units to be identified before the survey. The organization and management of interviewers and data collection support tools can also benefit from the overlap with previous survey periods. The hypothesis behind this choice is that the maintenance of around 80% of the units in the sample from one year to another greatly facilitates the identification and location of the units with a consequent reduction in costs and the time required to collect the data.

Stratified random selection without unit substitution is performed by using the technique of permanent random numbers (PRN, Ohlsson 1995).

In the following text box, the R script used to randomly extract the sampling units is reported.

R script - Code for sample selection coordinated over time for longitudinal surveys

The *sampling* package is used

Line 4 sets the minimum number of units per stratum equal to 3 while in line 4 the rotation parameter is set at 20%

In line 6 I read the fleet file while in line 8 it is ordered with respect to the stratum, descending by size

In line 12 the sampling rate is calculated for each stratum

Lines 13 to 18 are used to calculate the actual turnover rate in order to respect the population numbers and the minimum required for each stratum

In row 21 the seed of generation of the random numbers (line 22) is kept fixed so that they are always reproduced the same (PRN)

In line 22 the ratio between random numbers and inclusion probabilities proportional to the overall length is calculated

In line 23 the archive is again sorted by stratum even with respect to this new indicator

From line 24 to line 30 the only units that correspond to the actual rotation rate are selected in the sample

Line 31 calculates a double entry table to check the actual number of units selected in each year and rotated from one year to another

The output file is produced in line 32

```
1     library(sampling)
2     rm(list=ls(all=TRUE))
3     # pongo pari a 5 il numero minimo battelli x strato, e tasso rotazione
4     minstr <- 3
5     rotate <- 0.2
6     flotta <- read.csv2("Flotta2018_v2.csv")
7     colnames(flotta)[4] <- "c17"
8     flotta <- flotta[order(flotta$Strato,-flotta$c17),]
9     strnum <- as.data.frame.matrix(table(flotta$Strato,flotta$c17))
10    colnames(strnum) <- c("outs","ins")
11    tots <- colSums(strnum)[1:2]
12    tassoc <- tots[2]/(sum(tots))
```

```
13              strnum$Strato <- as.numeric(rownames(strnum))
14              strnum$totN <- strnum$outs + strnum$ins
15              strnum$ins[strnum$ins == 0] <- ceiling(strnum$totN[strnum$ins == 0] * tassoc)
16              strnum$outs <- strnum$totN - strnum$ins
17              strnum$totn <- pmin(pmax(strnum$ins,minstr),strnum$totN)
18              strnum$rot  <- pmin(ceiling(strnum$totn * rotate),strnum$totN-strnum$totn)
19              strnum$totlft <- tapply(flotta$LFT,flotta$Strato,FUN=sum)
20              flotta2 <- merge(flotta,strnum[,c(3,5:7)],by = "Strato")
21              set.seed(160964)
22              flotta2$p <- runif(nrow(flotta2))/(flotta2$LFT/flotta2$totlft)
23              flotta2 <- flotta2[order(flotta2$Strato,-flotta2$c17,flotta2$p),]
24              flotta2$count <- 1
25              for (i in 2:nrow(flotta2))
26              {
27                ifelse(flotta2$Strato[i] - flotta2$Strato[i-1] == 0,flotta2$count[i] <- flotta2$count[i-1] + 1,1)
28              }
29              flotta2$c18 <- 0
30              flotta2$c18[(flotta2$count > flotta2$rot) & (flotta2$count <= flotta2$totn + flotta2$rot)] <-
1
31              table(flotta2$c17,flotta2$c18)
32              write.csv2(flotta2[,c(1:3,10)], file="campione18.csv", quote=F)
```

8.1.4 Estimation of the totals of interest by Horvitz-Thompson estimators and Sen-Yates-Grundy variance estimators

To obtain an estimate of totals per stratum, the Horvitz-Thompson estimator is used, while the Sen-Yates-Grundy formula is used to estimate the relative sampling error. Detailed explanations of these methods are reported in Chapter 3 of the Handbook.

In this section, an application of the estimation procedure is presented. In the Italian survey, the size variable is the Length Overall (LFT) of the vessel. In the following table, the calculation of HT estimators for a single stratum is presented. The raising factor is calculated as: LFT/(lft*n).

ID_vessel	Stratum_code	lft_vessel	n_sample size	LFT_tot stratum	N_population	Raising factor
4345	1072	26.9	6	371.06	15	2.299009
5148	1072	26.55	6	371.06	15	2.329316
7075	1072	30.16	6	371.06	15	2.050508
18561	1072	23.92	6	371.06	15	2.585424
27472	1072	28.4	6	371.06	15	2.177582
17247	1072	26.12	6	371.06	15	2.367662

In the following text box the R script used to produce the final estimates according to the HT estimators in the PPS survey is reported.

R script - Estimation of the totals

```
library(survey)
#setwd("~/Documents/Nisea")
setwd("C:/Users/…")
# reading  csv files
datic <- read.table("sample values.csv",header=TRUE,sep=";",dec=",")
vinco <- read.table("constraints.csv",header=TRUE,sep=";",dec=",")
popol <- read.table("population.csv",header=TRUE,sep=";",dec=",")
# create strvin with the indication of strata on the constraints
strvin <- merge(vinco,strati2,by="Strato",all = TRUE)
# definition of the sample design
disegno  <- svydesign(ids = ~ 1,strata = ~Strato,data = datic,pps="brewer",fpc=~pinc)
# estimations
stitot0 <-
svytotal(~FuelCost+Labour+Maint+Commerc+OtherFix+OtherVar+OtherRev+Invest+Subs+Depr+Debts+FuelCons+Crew,disegno,deff=TRUE)
stistr0 <-
svyby(~FuelCost+Labour+Maint+Commerc+OtherFix+OtherVar+OtherRev+Invest+Subs+Depr+Debts+FuelCons+Crew,~Strato,disegno,svytotal)
write.table(stistr0,file="stistr0.csv",sep=";",dec=",")
```

In the following text box the R script used to produce the Sen-Yates-Grundy variance estimators is reported.

R script - Sen-Yates-Grundy variance estimators

```
#tempdir="C:/Users/….."
library(data.table)
campionari=fread(paste(tempdir,'campionari.csv',sep="/") )
pr_i=fread(paste(tempdir,'pr_i.csv',sep="/") )
pr_ij=fread(paste(tempdir,'pr_ij.csv',sep="/") )
strati=fread(paste(tempdir,'strati.csv',sep="/") )
setkey(campionari, batcod)
setkey(pr_i, batcod)
# check if there are batcod in samples not present in pr_i, in case the procedure returns a data.table with the indications of the errors
if ( sum(campionari[,unique(batcod)] %in% pr_i[,batcod])  != nrow(campionari[,.N,by=batcod])  ) {


  cv=data.table( batcod_in_campionari_not_in_pr_i=setdiff(campionari[,unique(batcod)], pr_i[,batcod]) )


} else {


  strati_cod_variable_unique = pr_i[ campionari[,.(batcod, cod_variable)], .(batcod,cod_variable,strato)]
[,.N,keyby=.(strato,cod_variable)][,N:=NULL]


  pr_i_temp=pr_i[,list(batcod.x=batcod,pr_i.x=pr_i)]
  setkey(pr_i_temp, batcod.x)
  setkey(pr_ij, batcod.x)
  pr_ij = pr_i_temp[pr_ij]

  pr_i_temp=pr_i[,list(batcod.y=batcod,pr_i.y=pr_i)]
  setkey(pr_i_temp, batcod.y)
  setkey(pr_ij, batcod.y)
  pr_ij = pr_i_temp[pr_ij]

  setkey(pr_ij, strato)
  setkey(strati_cod_variable_unique, strato)

  pr_ij = strati_cod_variable_unique[pr_ij,allow.cartesian=TRUE, nomatch=0]

  camp_temp=campionari[,list(batcod.x=batcod, cod_variable, values.x=values)]
  setkey(camp_temp, batcod.x, cod_variable)
```

```
 setkey(pr_ij, batcod.x, cod_variable)
 pr_ij = camp_temp[pr_ij]

 camp_temp=campionari[,list(batcod.y=batcod, cod_variable, values.y=values)]
 setkey(camp_temp, batcod.y, cod_variable)
 setkey(pr_ij, batcod.y, cod_variable)
 pr_ij = camp_temp[pr_ij]
 pr_ij[is.na(values.x), values.x:=0]
 pr_ij[is.na(values.y), values.y:=0]

 campionari = pr_i[campionari, .(batcod,cod_variable,values,strato,pr_i)]
 setkey(campionari, strato,cod_variable)
 tot=campionari[,list( tot_values=sum(values/pr_i)), by=.(strato,cod_variable) ]
 setkey(pr_ij, strato,cod_variable)
 var=pr_ij[,list( var_values= sum( (pr_i.x * pr_i.y / pr_xy - 1) * (values.x/pr_i.x - values.y/pr_i.y)^2 )),
by=.(strato,cod_variable)]

 cv_strato=var[tot][var_values>=0]
 cv_strato[, cv:=ifelse(tot_values==0,0,sqrt(var_values)/tot_values)]



 # this part is executed only if there are var <0. In this case, the cv calculation is performed with ccs:
 if (nrow(var[var_values<0])!=0)   {

   str_cod_variable_per_ccs=var[var_values<0,.(strato, cod_variable)]

   setkey(str_cod_variable_per_ccs, strato)
   setkey(pr_i,strato )
   pr_i=pr_i[str_cod_variable_per_ccs]
   camp_temp=campionari[,list(batcod,cod_variable,values)]
   setkey(camp_temp, batcod,cod_variable)
   setkey(pr_i, batcod,cod_variable)
   pr_i=camp_temp[pr_i]
   pr_i[is.na(values), values:=0]

   cv_strato_ccs=pr_i[,list(m=mean(values), s2=var(values)), keyby=.(strato,cod_variable)]
   setkey(strati, strato)
   cv_strato_ccs=strati[cv_strato_ccs]
```

```
   cv_strato_ccs=cv_strato_ccs[,.(strato,cod_variable, tot_values= m * N, var_values=(N^2 * (1-n/N)/n) * s2) ]
   cv_strato_ccs[,cv:=ifelse(tot_values==0, 0, sqrt(var_values)/(tot_values) ) ]
   cv_strato_ccs = cv_strato_ccs[,.(strato,cod_variable,cv)]
   setkey(cv_strato_ccs, strato,cod_variable)
   cv_strato_ccs=tot[cv_strato_ccs,.(strato,cod_variable,tot_values,cv)]
   cv_strato_ccs = cv_strato_ccs[, var_values:=(cv*tot_values)^2][,names(cv_strato), with=F]
   cv_strato=rbindlist( list(cv_strato, cv_strato_ccs) )


  }


  cv_totale=cv_strato[,.(strato=0,tot_values=sum(tot_values), var_values=sum( (tot_values*cv)^2 ) ),by='cod_variable']
  cv_totale[tot_values>0,cv:=sqrt(var_values)/tot_values]
  cv_totale[is.na(cv), cv:=0]


  cv=rbindlist( list(cv_strato[,.(cod_variable, strato, tot=tot_values ,cv)], cv_totale[,.(cod_variable, strato, tot=tot_values ,cv)] ) )
  cv[,(c('tot','cv')):=list(round(tot,2),round(cv,5))]
  setorder(cv, strato,cod_variable)


}
filename=paste(tempdir,"cv_pps.csv",sep="/")
write.table(x = cv, file = filename,quote = F,sep = ";",na = "",row.names = F )
```

Key references

Key references to the multivariate allocation method:

- Buglielli, T., De Vitiis, C. and Barcaroli, G. (2013) MAUSS-R - Multivariate Allocation of Units in Sampling Surveys. User and Methodological Manual (version 1.1). ISTAT, Italy.
- Bethel J. (1989) Sample Allocation in Multivariate Surveys. Survey Methodology 15, 47-57.
- Chromy J. (1987) Design Optimization with Multiple Objectives. Proceedings of the Survey Research Methods Section, American Statistical Association, pp.194-199.

Key references to the Code for sample selection coordinated over time for longitudinal surveys

- Ohlsson E. (1995). Coordination of samples using permanent random numbers, In Cox BG, Binder DA, Nanjamma Chinnappa B, Christianson A, Colledge MJ, Kott PS (Eds.), Business Survey Methods, 153–169. New York: Wiley.

Key reference to the PPS sampling method (Hanurav-Vijayan):

- Chaudhuri, A. and Vos, J.W.E. (1988) Unified Theory and Strategies for Survey Sampling. Amsterdam: North-Holland. Sections 4.8 and 5.17.
- MAUSS-R web pages at: https://www.istat.it/en/methods-and-tools/methods-and-it-tools/design/design-tools/mauss-r#Softwareanddocumentation-2

8.2 Finland

Application of statistical methods in data collection: Regression estimation in Finnish data collection.

8.2.1 Introduction

Here we present the application of regression (and ratio) estimation in the Finnish trawler segment for estimating the cost and earnings variables. The trawler fleet consist of 53 vessels in 2017 are divided into three fleet segments. We present here the estimation for the TM1824 segment that consists of 13 vessels.

8.2.2 Data collection and sources

Economic data collection is based on hierarchical multi-stage survey that combines information from different data sources. Main sources are the central control register on commercial fishery (includes fishery catch data, fishing vessel register, first hand sales of quota species), and financial statement statistics, statistics on business subsidies and employment statistics from Statistic Finland (SF). An additional account surveys for trawlers conducted by Natural Resources Institute Finland (Luke).

Information on catches by species, value of landings by species, effort data and vessel capacity information is collected by vessel. This data is collected exhaustively for all vessels. Economic data is collected by fishing unit: company or fisherman (including family members). Financial statements data for fishing firms with income over a threshold level of around € 11 000 are obtained from the database of Statistics Finland (SF) on structural business and financial statement statistics.

Financial data gives a reliable estimate for profitability of the larger vessels, but the disaggregation of cost items does not follow that in regulation. Therefore data on the cost and earnings structure is collected with an additional account survey on trawlers every 3 year.

Luke compares landings statistics against the turnover data from Statistics Finland and from account survey. Ratio between turnover and value of landings per company is calculated to spot abnormalities. Due to the under-coverage in the structural business and financial statement statistics (compared to target population) the segment totals need to be estimated with regression estimation and additional cost structure analysis. Coefficients of variation and coverage rates are calculated for each variable and for each vessel segment. Regression output results are analyzed to check they are statistically valid.

8.2.3 Estimation procedures

Cost and earnings estimates for trawler segments are done by design-based and model assisted regression and ratio estimation using SAS.

1) First, the turnover and total income per segment are estimated with regression using PROC SURVEYREG of SAS and using the total value of catch as explanatory (auxiliary) variable. The actual syntax used in Finland is quite complex, so for demonstration purposes, more simple code is presented as follows:

```
title1 'Turnover from catchvalue';
proc surveyreg data=Table1 total=Totals;
   strata segment /list;
   model turnover = catch_value;
   weight Weight;
   estimate "Turnover in all classes under Model III"
            catch_value_1 catch_value_sum_1
               /e;
   ods output ParameterEstimates = MyParmEst_turnover;
run;
proc print data=MyParmEst_turnover;
run;
```

```
title1 'Total income from catchvalue';
proc surveyreg data=Table1 total=Totals;
   strata segment /list;
   model total_income = catch_value;
   weight Weight;
   estimate "Total income in all classes under Model III"
            catch_value_1 catch_value_sum_1
               /e;
   ods output ParameterEstimates = MyParmEst_totinc;
run;
proc print data=MyParmEst_totinc;
run;
```

2) Next, the total costs are estimated for total population per segments from the turnover as follows:

```
title1 'Total costs from turnover';
proc surveyreg data=Table2 total=Totals2;
   strata segment /list;
   model total_costs = turnover;
   weight Weight;
   estimate "Total costs in all classes under Model III"
            turnover_1 turnover_sum_1
               /e;
   ods output ParameterEstimates = MyParmEst_totcost;
run;
proc print data=MyParmEst_totcost;
run;
```

3) As third step, the average percentage share for each cost item from total costs in each vessels segment is calculated. For example, the percentage share for fuel costs is calculated with the following formula:
   Fuelcost_%=(sum of fuel costs in a vessel segment)/(sum of total costs in a vessel segment).
4) Finally, the cost variables are estimated as ratio estimates from the estimated total costs by multiplying the percentage share of each cost item with the total costs for each vessel segment as follows:
   Fuel_costs =Fuelcost_%*Total_Cost

8.2.4 Results

The results from the estimation of turnover for TM1824 are presented in Table 8.1. The Coefficient of variation for the segment turnover is 123328/4364399=0,028. Similar results for each regression (Turnover, Total income, Total costs) and each vessels segment are generated by SAS.

Table 8.1. SAS output of regression estimation of turnover for TM1824.

Total estimate for turnover

Estimate					
Label	Estimate	Standard Error	DF	t Value	Pr > |t|
Ivaihto luokassa 4 under Model II	4364399	123328	412	35.39	<.0001

The estimated regression coefficients are presented in Table 8.2. The regression function used in the estimation of turnover for TM1824 is also given.

Table 8.2. SAS output of estimated regression coefficients for Turnover.

Turnover=-24783.446+ 0.898*value of catch.

Estimated Regression Coefficients				
Parameter	Estimate	Standard Error	t Value	Pr > |t|
luokka6 p1	1526.263	1094.5903	1.39	0.1640
luokka6 p2	20931.954	4045.3862	5.17	<.0001
luokka6 t1	45638.838	23257.2236	1.96	0.0504
luokka6 t2	-24783.446	13676.6077	-1.81	0.0707
luokka6 t3	59219.478	17413.6183	3.40	0.0007
Arvo_1	1.285	0.1108	11.60	<.0001
Arvo_2	0.542	0.0862	6.29	<.0001
Arvo_3	0.810	0.0665	12.18	<.0001
Arvo_4	0.898	0.0377	23.80	<.0001
Arvo_5	0.888	0.0166	53.56	<.0001

Note: The degrees of freedom for the t tests is 412.

REFERENCES

Benedetti, R., Piersimoni, F., Bee, M. and Espa, G. (eds) (2010) *Agricultural Survey Methods*. Wiley.

Bethel, J. (1989) Sample allocation in multivariate surveys. *Survey Methodology* 15, 47-57.

Biemer, P.P., de Leeuw, E., Eckman, S., et al (Eds.) (2017) *Total Survey Error in Practice*. New York: John Wiley & Sons.

Biemer, P.P., and Lyberg, L.E. (2003) *Introduction to survey quality*. Hoboken, NJ: John Wiley & Sons.

Buglielli, T., De Vitiis, C. and Barcaroli, G. (2013) MAUSS-R - Multivariate Allocation of Units in Sampling Surveys. User and Methodological Manual (version 1.1). ISTAT, Italy.

Chaudhuri, A. and Vos, J.W.E. (1988) *Unified Theory and Strategies for Survey Sampling*. Amsterdam: North-Holland. Sections 4.8 and 5.17.

Chromy, J. (1987) Design optimization with multiple objectives. Proceedings of the Survey Research Methods Section, American Statistical Association, pp.194-199.

Cochran, W.G. (1963) *Sampling Techniques*. New York: John Wiley & Sons.

Deville, J.-C. and Särndal, C.-E. (1992) Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376–382.

Durbin, J. (1969) Inferential aspects of the randomness of sample size in survey sampling. In Johnson, N.L. and Smith, H. (Eds.) *New Developments in Survey Sampling*. New York: Wiley, 629-651.

Enders, C.K. (2010) *Applied missing data analysis*. New York: Guilford Press.

Groves, R.M., Dillman, D.A., Eltinge, J.L. and Little, J.A. (2001) *Survey Nonresponse*. New York: John Wiley & Sons.

Hanurav, T.V. (1967) Optimum utilization of auxiliary information: πps sampling of two units from a stratum. *Journal of the Royal Statistical Society, Series B*, 374–391.

Heeringa, S.G., West, B.T. and Berglund, P.A. (2017) *Applied Survey Data Analysis*, 2nd Edition. Chapman and Hall/CRC

Hidiroglou, M.A. and Patak, Z. (2004) Domain estimation using linear regression. *Survey Methodology*, 30, 67–78.

Holt, D. and Smith, T.M.F. (1979) Post-stratification. *Journal of the Royal Statistical Society, Series A*, 142, 33-46.

Horvitz, D.G. and Thompson, D.J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

Lehtonen, R. and Pahkinen, E. (2004) *Practical Methods for Design and Analysis of Complex Surveys*. Second Edition. Chichester: John Wiley & Sons.

Lehtonen, R. and Veijanen, A. (2009) Design-based methods of estimation for domains and small areas. In Rao, C.R. and Pfeffermann, D. (Eds.) *Handbook of Statistics, Vol. 29B. Sample Surveys. Inference and Analysis*. Amsterdam: Elsevier, 219–249.

Little, R. J., and Rubin, D. B. (2014) Statistical analysis with missing data. John Wiley & Sons.

Lohr, S. (2009) *Sampling: Design and Analysis*. 2nd ed. Brooks/Cole, Cengage Learning.

Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology,* 33, 99–119.

Särndal, C.-E., and Lundström, S. (2005) *Estimation in surveys with nonresponse*. New York: John Wiley & Sons.

Särndal, C.-E., Swensson B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. New York: Springer.

Tillé, Y. (2006) *Sampling Algorithms*. New York: Springer.

Vijayan, K. (1968) An exact πps sampling scheme: generalization of a method of Hanurav. *Journal of the Royal Statistical Society, Series B*, 556–566.

Wolter, K. (2007) *Introduction to Variance Estimation*. New York: Springer.

Appendices

Appendix A: SAS implementation of worked examples

A.1 SAS SURVEY procedures

We introduce briefly the basic SAS SURVEY procedures for sample selection from populations and the analysis of the drawn sample, and then we present the SAS codes (with selected results) that were used in the "Worked example" sections in Chapters 3 to 5 of the main text. The SAS version 9.4 was used in the examples. Up-to-date information on survey sampling and analysis features can be found at http://support.sas.com/rnd/app/stat/procedures/SurveyAnalysis.html. The SAS 9.4 procedures employed for the examples are the following.

PROC SURVEYSELECT: Sample selection from the sampling frame data set with a variety of equal and unequal probability sampling methods involving stratification (`STRATA` statement) and clustering (`CLUSTER` statement). In stratified sampling, proportional allocation, Neyman allocation and optimal allocation can be used. An element weight variable SAMPLINGWEIGHT is included in the output data set. Joint selection probabilities can be computed for some sampling designs. The replicated sampling option allows independent sampling from the frame by different sampling designs and output the samples to a SAS data set. The following basic sample selection techniques (see also Table 3.1) are included:

Equal probability sampling techniques:

- Simple random sampling without replacement and with replacement
- Systematic sampling
- Bernoulli sampling
- Balanced bootstrap sampling

Unequal probability sampling techniques:

- PPS sampling without replacement using the Hanurav-Vijayan algorithm or with the Brewer, Murthy or Sampford methods
- PPS sampling with replacement
- PPS systematic sampling
- Sequential PPS sampling with minimum replacement by the Chromy method
- Poisson PPS sampling

Up-to-date information on capabilities of PROC SURVEYSELECT can be obtained at: https://documentation.sas.com/?docsetId=statug&docsetTarget=statug_surveyselect_syntax01.htm&docsetVersion=15.1&locale=en

PROC SURVEYMEANS: Horvitz-Thompson (expansion) estimation of totals, means, medians and other descriptive statistics for populations and pre-defined subpopulations or strata (`BY` statement) and domains (`DOMAIN` statement). Estimation of ratios and post-stratification are also included. Variances for the estimated statistics can be estimated by the linearization or sample re-use methods (balanced repeated replications and the jackknife method). In domain estimation, the extra variation because of random domain sample sizes is accounted for by the extended domain variables technique (Lehtonen & Veijanen 2009 p. 223). The sampling design can be complex involving stratification, clustering, and unequal weighting.The following statistics, their standard errors, confidence intervals and coefficients of variation can be computed:

- Means and totals
- Proportions
- Quantiles
- Geometric means
- Ratios of two totals or means

Up-to-date information on capabilities of PROC SURVEYMEANS can be obtained at: https://documentation.sas.com/?docsetId=statug&docsetTarget=statug_surveymeans_toc.htm&docsetVersion=15.1&locale=en

PROC SURVEYREG: Design-based linear regression analysis, ANOVA and ANCOVA under stratified one-stage and multi-stage sampling designs with equal and unequal probability sampling. The estimation of totals and means by ratio and regression estimation, post-stratification and calibration methods can be performed for the entire survey population and pre-defined sub-populations or strata (`BY` statement) and domains (`DOMAIN` statement). The aggregate-level auxiliary information is incorporated in regression estimation by specifying suitable linear functions (`ESTIMATE` statement). The sampling design can be complex involving stratification, clustering, and unequal weighting.

Up-to-date information on capabilities of PROC SURVEYREG can be obtained at: https://documentation.sas.com/?docsetId=statug&docsetTarget=statug_surveyreg_toc.htm&docsetVersion=15.1&locale=en

PROC SURVEYIMPUTE can be used for imputation for item missingness (not used here): https://documentation.sas.com/?docsetId=statug&docsetVersion=15.1&docsetTarget=statug_surveyimpute_overview.htm&locale=en

A.2 Section 3.3.4: Simple random sampling example

Selection of a SRSWOR sample with PROC SURVEYSELECT

```
* CODE BOX 3.1: SRSWOR sampling from SIMPOP with PROC SURVEYSELECT;
/* Simple random sampling without replacement (SRSWOR) is requested by the SAS
option method=srs.

Option seed given by the user specifies the initial seed for random number
generation in SRSWOR. Here the initial seed is kept constant over the examples (to
be able to reproduce the samples and estimates). The requested sample size is n=5.
Sampling weights (inverses of inclusion probabilities) are included automatically in
the sample data set.
The drawn SAMPLE1 of n=5 elements in Table 3.7.*/

proc surveyselect data=pop out=sample1
sampsize=5 seed=98765 method=srs stats;
run;

/* NOTE: SAS programming language is not case sensitive (uppercase and lowercase
code is treated as equivalent)*/
```

Estimation of population total under SRSWOR_HT strategy with PROC SURVEYMEANS

```
* CODE BOX 3.2: Estimation of total of CATCH for SAMPLE1 by PROC SURVEYMEANS;
proc surveymeans data=sample1
sumwgt nobs sum std cvsum clsum total=100;
var CATCH;
weight SamplingWeight;
run;

/* Estimation results are in Table 3.8

Options for output control:
sumwgt: sum of weights
nobs: sample size
sum: total estimate
std: standard error for total
cvsum: coefficient of variation for total
clsum: 95% confidence limits for total;*/
```

A.3 Section 4.3.4: Domain estimation example

```
* CODE BOX 3.3: Domain estimation of totals of CATCH for domains DOM01=0 and DOM01=1
of SAMPLE2 of n=20 elements by PROC SURVEYMEANS;

* SCENARIO 1: Estimation under the conditional approach assuming known domain sizes
N=70 for DOM01=0 and N=30 for DOM01=1 in population;

data domain0;
set sample2;
where DOM01=0;
run;

proc surveymeans data=domain0 nobs sumwgt sum cvsum clsum total=70;
var CATCH;
weight SamplingWeight;
run;

* Computation for DOM01=1 similarly for data set domain1 of n=8 elements and setting
the option total=30;

* Output for DOM01=0 (Table 3.11 first row);

The SURVEYMEANS Procedure

            Data Summary
Number of Observations            12
Sum of Weights                    60

                                          Statistics

                                       Sum of
Variable  Label             N       Weights           Sum       Std Dev      95% CL for Sum
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
CATCH     CATCH            12     60.000000        419536         48298   313232.887 525838.924
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ

      Statistics

              Coeff of
             Variation
Variable       for Sum
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
CATCH         0.115122
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
```

```
* CODE BOX 3.4: Domain estimation of totals of CATCH for domains DOM01=0 and DOM01=1
of SAMPLE2 of n=20 elements by PROC SURVEYMEANS;

* SCENARIO 2: Estimation under the conditional approach assuming unknown domain
sizes in population and by using data set domain0 of n=12 from Code Box 3.3 (NOTE:
No total= option in SURVEYMEANS call);

proc surveymeans data=domain0 nobs sumwgt sum cvsum clsum;
var CATCH;
weight SamplingWeight; run;

* Computation for DOM01=1 similarly for data set domain1 of n=8 elements

* Output for DOM01=0 (Table 3.11 3rd row);

The SURVEYMEANS Procedure

            Data Summary

Number of Observations            12
Sum of Weights                    60


                                        Statistics

                                    Sum of
Variable  Label              N       Weights           Sum       Std Dev      95% CL for Sum
ffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
CATCH     CATCH              12     60.000000        419536         53060  302752.639 536319.172
ffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff

      Statistics

              Coeff of
             Variation
Variable       for Sum
fffffffffffffffffffff
CATCH         0.126472
fffffffffffffffffffff
```

```
* CODE BOX 3.5: Domain estimation of totals of CATCH for domains DOM01=0 and DOM01=1
of SAMPLE2 of n=20 elements by PROC SURVEYMEANS;

* SCENARIO 3: Estimation under the unconditional approach using extended domain
variables for data set sample2 of n=20 (NOTE: Analysis over the entire sample data
set);

proc surveymeans data=sample2 nobs sumwgt sum cvsum clsum total=100;
var CATCH;
domain DOM01;
weight SamplingWeight; run;

* Output (Table 3.11 last 2 rows);

The SURVEYMEANS Procedure

            Data Summary

Number of Observations            20
Sum of Weights                   100

                                        Statistics

                                    Sum of
Variable  Label              N       Weights           Sum       Std Dev      95% CL for Sum
ffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
CATCH     CATCH             20    100.000000        610603         54439  496661.885 724544.886
ffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff

                                  Domain Analysis: DOMAIN
                                                        Sum of
  DOMAIN    Variable    Label               N         Weights              Sum         Std Dev
ffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
       0    CATCH       CATCH              12       60.000000          419536           84344
       1    CATCH       CATCH               8       40.000000          191067           50990
ffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff

                  Domain Analysis: DOMAIN
                                                       Coeff of
                                                      Variation
  DOMAIN    Variable        95% CL for Sum             for Sum
ffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
       0    CATCH       243002.412 596069.399        0.201041
       1    CATCH        84343.946 297791.014        0.266870
ffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
```

A.4 Section 3.5.4: PPS sampling example

Selection of a PPSWOR sample with PROC SURVEYSELECT

```
* CODE BOX 3.6: PPSWOR sampling from SIMPOP with PROC SURVEYSELECT;
/* PPS sampling without replacement (PPSWOR) is requested by the SAS option
method=pps. The requested sample size is n=5.
The size variable is given by the size statement.

The drawn SAMPLE3 of n=5 elements in Table 3.13.*/

proc surveyselect data=simpop out=sample3 sampsize=5 seed=98765 method=pps stats;
size GT;
run;
```

Estimation of population total under PPSWOR_HT strategy with PROC SURVEYMEANS

```
* CODE BOX 3.7: Estimation with SAMPLE3 under PPSWOR sampling;

proc surveymeans data=sample3 sumwgt nobs sum std cvsum clsum total=100;
var CATCH;
weight SamplingWeight;
run;

* Estimation results are in Table 3.14;

/* NOTE: The only difference in SURVEYMEANS code for estimation under PPSWOR and
SRSWOR sampling is in the contents of the SAMPLINGWEIGHT variable.*/
```

A.5 Section 3.6.4: Stratified sampling example

```
*CODE BOX 3.8. SAS code for stratified sampling and HT estimation;
* Sampling: Sort SIMPOP by STR3 and save it as population data set POPS;
proc sort data=pop out=pops;
by STR3; run;

*(a) STR_SRSWOR sampling, n=20;
proc surveyselect data=pops out=sample5(keep=id str3 catch samplesize
      selectionprob samplingweight)
      sampsize=20 seed=98765 method=srs stats;
strata STR3 / alloc=proportional; * definition of strata and allocation;
run;

(b) STR_PPSWOR sampling, n=20;
proc surveyselect data=pops out=sample6(keep=id str3 catch samplesize
      selectionprob samplingweight)
      sampsize=20 seed=98765 method=pps stats;
strata STR3 / alloc=proportional;
size GT_DAS; run;

* Estimation;
* Input strata sizes N_h, h = 1,2,3 in population into data set STRATA;
data STRATA;
input STR3 _TOTAL_;
datalines;
1 33
2 33
3 34
;
run;

(a) STR_SRSWOR_HT with SAMPLE5;
proc surveymeans data=SAMPLE5 sumwgt nobs sum std cvsum clsum total=STRATA;
var CATCH;
strata STR3;
weight SamplingWeight; run;

(b) * STR_PPSWOR_HT with SAMPLE6;
* Replace SAMPLE5 by SAMPLE6 in (a);

* The samples are in Table 3.17 and estimation results are in Table 3.19;
```

A.6 Section 4.2.3: Ratio and regression estimation examples

```
* CODE BOX 4.1 Ratio estimation with SAMPLE7 of n=5 by PROC SURVEYMEANS;
* For ratio estimation for CATCH total with GT as auxiliary variable we first
estimate by equation (19) the ratio r=CATCH/GT between HT estimated totals of CATCH
and GT. We then compute the variance estimate for the ratio estimated total t(RAT)
by using the variance estimate v(r) of the estimated ratio r and the square of known
GT total in population;

* The SRSWOR sample SAMPLE7 of n=5 elements is shown in Table 4.2;
* Estimation results (computed below) are in Table 4.6;

proc surveymeans data=SAMPLE7 ratio total=100;
ratio CATCH/GT;
weight SamplingWeight;
run;

The SURVEYMEANS Procedure

            Data Summary

Number of Observations            5
Sum of Weights                  100

                  Ratio Analysis

Numerator Denominator        Ratio         Std Err
ffffffffffffffffffffffffffffffffffffffffffffffffff
CATCH     GT             20.105147       2.831090
ffffffffffffffffffffffffffffffffffffffffffffffffff


* Variance and standard error estimation of ratio estimated total of CATCH;
data a;
t_x=32896.4; * known population total of GT;
se_r=2.831090; * s.e of ratio r;
var_r=se_r**2; * variance of ratio r;
var_t_rat=t_x**2*var_r; variance of ratio estimated CATCH total;
se=sqrt(var_t_rat); * standard error of ratio estimated total;
run;
proc print data=a; run;

* Computed standard error estimate;

Obs      t_x        se_r      var_r       var_t_rat        se

 1     32896.4    2.83109    8.01507    8673694049.2    93132.67
```

```
* CODE BOX 4.2 Ratio estimation with SAMPLE7 of n=5 by PROC SURVEYREG;
* Ratio estimation for CATCH total with GT as auxiliary variable is executed as a
special case of regression estimation by fitting a linear model without an intercept
term (option noint) and using the estimate statement to supply the GT total;

* Estimation results are in Table 4.7;

proc surveyreg data=SAMPLE7 total=100;
model CATCH=GT / solution noint;
weight SamplingWeight;
estimate "CATCH total" GT 32896.4 / E; run;

* Output;

The SURVEYREG Procedure
Regression Analysis for Dependent Variable CATCH

            Data Summary
Number of Observations             5
Sum of Weights             100.00000
Weighted Mean of CATCH        7225.4
Weighted Sum of CATCH       722538.8


      Fit Statistics
R-square            0.9297

             Estimated Regression Coefficients

                               Standard
Parameter      Estimate         Error    t Value    Pr > |t|
GT           20.6761657    2.72832884       7.58      0.0016

    Estimate
  Coefficients

Effect      Row1
GT         32896

Estimate

                           Standard
Label          Estimate       Error       DF    t Value    Pr > |t|

Catch total      680171       89752        4       7.58      0.0016
```

CODE BOX 4.3 Regression estimation with SAMPLE7 of n=5 by PROC SURVEYREG;

```
* Regression estimation for CATCH total with GT as auxiliary variable is executed by
fitting a linear regression model (24) for SAMPLE7 and using the estimate statement
to supply the INTERCEPT total (=population size) and the GT total;

* Estimation results are in Table 4.8;

proc surveyreg data=SAMPLE7 total=100;
model CATCH=GT / solution;
weight SamplingWeight;
estimate "CATCH total" INTERCEPT 100 GT 32896.4 / CL E;
run;

* Output;
```

```
The SURVEYREG Procedure

Regression Analysis for Dependent Variable CATCH

            Data Summary
Number of Observations             5
Sum of Weights             100.00000
Weighted Mean of CATCH        7225.4
Weighted Sum of CATCH       722538.8

      Fit Statistics
R-square            0.6701

             Estimated Regression Coefficients
                              Standard
Parameter      Estimate          Error    t Value    Pr > |t|

Intercept    -6238.6791    2363.59006      -2.64      0.0576
GT              37.4647       8.53363       4.39      0.0118

     Estimate
   Coefficients

Effect         Row1

Intercept       100
GT            32896

Estimate

                          Standard
Label         Estimate      Error      DF   t Value   Pr > |t|    Alpha      Lower      Upper
CATCH total     608586      78985       4      7.71     0.0015     0.05     389288     827884
```

A.7 Section 4.3.3: Post-stratification example

```
* CODE BOX 4.4 Post-stratification with SAMPLE9 of n=20 by PROC SURVEYMEANS;
* The post-stratification variable POST2 is formed from variable DOM01 by changing
the class codes (if DOM01=0 then POST2=1, if DOM01=1 then POST2=2);
* Post-stratification is based on equation (29) and is executed with PROC
SURVEYMEANS by using the POSTSTRATA statement;

* Estimation results are in Table 4.14;

* Defining the distribution of POST2 in population;
data FREQ;
input POST2 _PSTOTAL_;
datalines;
1 70
2 30
; run;

proc surveymeans data=SAMPLE9 nobs sum cvsum sumwgt clsum total=100;
var CATCH;
poststrata POST2 / pstotal=FREQ outpswgt=PSWGT;
weight SamplingWeight;run;

* Output;

The SURVEYMEANS Procedure

            Data Summary
Number of Poststrata              2
Number of Observations           20
Sum of Weights                  100

                                          Statistics

                                     Sum of
Variable  Label            N       Weights           Sum       Std Dev      95% CL for Sum
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
CATCH     CATCH           20    100.000000        632759          55889  515782.798 749735.535
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ

              Coeff of
             Variation
Variable       for Sum
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
CATCH         0.088325
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
```

A.8 Section 5.5: Nonresponse adjustment example

```
* CODE BOX 5.1 Adjustment for non-response with SAMPLE9 of n=20 by post-
stratification and regression estimation;

* Estimation results are in Table 5.2;

* 1) Post-stratification with PROC SURVEYMEANS;
* Post-stratification variable POST5 is formed from variable GT by dividing its
values into 5 equally-sized classes;

data FREQ;
input POST5 _PSTOTAL_ ;
datalines;
1 20
2 20
3 21
4 19
5 20
;
run;

proc surveymeans data=SAMPLE9 sum total=100;
var CATCH;
poststrata POST5 / pstotal=FREQ outpswgt=PSWGT;
weight SamplingWeight;
run;

* 2) Regression estimation by PROC SURVEYREG;
* The original continuous variable GT is used as the auxiliary variable in the
linear regression model in regression estimation of CATCH total with the ESTIMATE
statement;

proc surveyreg data=SAMPLE9 total=100;
model CATCH=GT / solution ;
weight SamplingWeight;
estimate "CATCH total" INTERCEPT 100 GT 32896.4 / cl e;
run;
```

Appendix B: R-implementation of worked examples

Most of the analyses in the "Worked example" -sections of Chapters 3-5 in the main text were also implemented in the R environment (R Core Team 2018, version 3.4.4). The main workhorses were R packages `sampling` (Tillé and Matei 2016) and `survey` (Lumley 2019). Their use is described briefly in the beginning of this Appendix. Those descriptions are followed by vignettes of the worked examples. The code of the vignettes is available online at www.zzz.zzz.

The explanation of the code in the vignettes is very brief, and gets briefer towards the end. The main objective of the comments is to point out the corresponding parts of the main text, where the examples are discussed in more detail. Sometimes same operations are coded differently in different examples in order to illustrate different ways of doing things in R, that might be most convenient in different situations. Some code used to print the results was not included in the vignettes, but everything is available in the online R codes.

B1 `sampling`: R functions for sample selection

Package `sampling` is arguably the most extensive collection of R functions for implementing various sampling designs. It also contains functions for estimation and calibration, but `survey` package was chosen for those tasks because of its easier use.

Simple random sampling is obtained with R base function `sample`, but all other sampling designs are easier to implement with tailored sampling functions. In the examples below, we used functions `UPsystematic`, `UPtille`, and `strata` from `sampling` package to implement systematic, PPS, and stratified designs, respectively.

The first two functions, `UPsystematic` and `UPtille`, require only one argument, vector `pik` whose length is equal to the population size `N` and elements equal to the desired inclusion probabilities. For equal probability sampling, set `pik=rep(n/N, N)`, where `n` is the sample size. They return a binary vector of length `N` with value 1 indicating inclusion in the sample and 0 exclusion.

Use of function `strata` is a bit more complicated. It takes as its first argument a data frame `data` containing one row corresponding to each of the `N` elements in the sampling frame, and as second argument `stratanames` name(s) of the categorical variable(s) in `data` that are used for stratification. Formally, they can be numeric, character, or factors, but they should naturally contain several replications of each value. `data` must be sorted in ascending order by the columns given in the `stratanames` argument before applying the function. The third required argument of `strata`, `size`, is a vector that gives the stratum sample sizes in the order, in which the strata are given in `data`. The available sampling designs within strata, optional argument `method`, are `"srswor"` (the default), `"srswr"`, `"poisson"`, and `"systematic"`. The last two allow for unequal probability sampling within strata, in which case one further argument is required: `pik` , the inclusion probabilities or, more conveniently, a vector of values such that the inclusion probabilities are proportional to them within strata. Typically this would be one (auxiliary variable) column of `data`.

Differently from the other sample selection functions in package `sampling`, `strata` returns the indices (rather than indicators) of the frame elements chosen to the sample (column `ID_unit` in the output data frame, the values corresponding to row numbers in input data frame `data`). The output data frame also contains other information, most significantly the computed inclusion probabilities for the sampled elements (column `Prob`).

For all sample selection functions in package `sampling`, but especially for `strata`, `getdata` is a recommendable convenience function for extracting the measurements of the sampled units from the population data.

B2 `survey`: R functions for design-based estimation

Package `survey` offers a unified approach for implementing a wide collection of different estimation strategies to complex survey samples. The analysis always starts by specifying the sampling design through function `svydesign`. Its most important arguments are

`ids`: identification of clusters; for element sampling designs, give `ids=~1`; this is a required argument

`probs`: inclusion probabilities for unequal probability sampling; not necessary for equal probability sampling; argument `weights` is an alternative

`strata`: specification of stratifying variable(s) in `data`, if any

`fpc`: finite population correction; most convenient to specify by giving a vector of population (stratum) sizes; see the examples

`data`: data frame to look up variables in the other arguments (the sample)

The estimator of the design variance is also essentially specified in the call of `svydesign` (arguments `pps` and `variance`). The default is to use the with-replacement approximation (formula 13 in the main text) for PPS sampling.

In simpler strategies, where auxiliary information is not utilized in the estimation phase, the next step is to call one of the estimation functions in package `survey`. In the examples, we use exclusively function `svytotal` to estimate the population total, but a whole bunch of other functions is available for other needs (see `vignette('survey')`, for examples). Typically these functions only need the specification of the design (obtained with `svydesign`, or its extensions discussed below) and target variable(s).

Finally, the estimate, its standard error, confidence interval, and coefficient of variation can be extracted from the object returned by `svytotal` or its relatives using functions `print`, `coef`, `SE`, `confint`, and `cv` as illustrated in the examples.

Domain estimation is obtained with function `svyby` taking `svytotal` or a relative as one of its arguments. This is illustrated in the end of the simple random sampling example.

Model-assisted estimates are obtained with two further steps after the call of `svydesign`. First, either `svyratio` (ratio estimation) or `svyglm` (regression estimation) is called to fit the model, and then their `predict` method returns the population estimates.

Function `calibrate`, also illustrated in the ratio and regression estimation examples, produces an extended design object by reweighting and adding information to an object returned by `svydesign`. After that, functions like `svytotal` can be called with the design specified by `calibrate` in the same way as with the design returned by `svydesign`. Function `postStratify` works in a similar manner (see the post-stratification example).

B3 Section 3.3.4: Simple random sampling example

B3.1 Preliminaries

We use the population of active vessels in SIMPOP. R code in `preliminaries.r` includes reading `SIMPOP` data from the Excel file.

```
source('preliminaries.r')
library(survey)
pop <- subset(SIMPOP, ACTIVITY == 1)
N <- nrow(pop)
```

B3.2 Sample selection

SRSWOR samples can be selected with base R function `sample`. Given a vector of unique frame element id's and desired sample size as the first two arguments, `sample` returns the vector of n id's included in the sample.

```
n <- 5
s1 <- subset(pop, ID %in% sample(pop$ID, n), select=c(ID, CATCH))
```

```
 Obs ID    CATCH
   1 20 4158.350
   2 33 3871.413
   3 50 7179.005
   4 69 8709.466
   5 77 6314.279
```

We demonstrate estimation using the sample of 5 vessels listed in Table 3.7. rather than the sample drawn above.

```
n <- 5
SAMPLE1 <- subset(pop, ID %in% c(1,44,49,55,93))
SAMPLE1$SamplingWeight <- N/n
```

```
 Obs ID     CATCH SamplingWeight
   1  1  3541.440             20
   2 44  4421.918             20
   3 49 11355.973             20
   4 55  6865.416             20
   5 93  9942.192             20
 Sum    36126.939            100
```

B3.3 Estimation

To obtain HT estimator for CATCH total, we first specify the sampling design using function `svydesign` and then compute the estimator and its standard error using function `svytotal`, both from package `survey`.

```
des <- svydesign(
  ids=~1,
  fpc=rep(N, n),
  data=SAMPLE1)
res <- svytotal(~CATCH, des)
```

```
       total     SE
CATCH 722539 147823
```

Package `survey` also contains a specific method for generic R function `confint` to compute an appropriate confidence interval from the `svystat` object returned by `svytotal`, as well as, function `cv` to compute the coefficient of variation (c.f. Table 3.8).

```
confint(res,df=degf(des))
```

```
         2.5 %  97.5 %
CATCH 312117.2 1132960
```

```
cv(res)
```

```
          CATCH
CATCH 0.2045879
```

20 vessels included in the larger SAMPLE2 of Section 3.3.5 are listed in Table 4.12. Results of Table 3.9 are thus replicated as follows.

```
n <- 20
SAMPLE2.Obs.ID <- data.frame(
  Obs=1:n,
  ID=c(1, 9, 29, 41, 47, 56, 63, 68, 69, 71,78, 94, 7, 20, 22, 24, 34, 37, 51
, 79)
)
SAMPLE2 <- merge(SAMPLE2.Obs.ID, pop)
SAMPLE2$SamplingWeight <- N/n
des <- svydesign(
  ids=~1,
  fpc=rep(N, n),
  data=SAMPLE2)
res <- svytotal(~CATCH, des)
```

The following numbers are obtained from `res` with extractor functions `coef` (the estimate), `SE` (standard error), `confint`, and `cv`

```
  Total  s.e. lower 95% CL upper 95% CL       cv
 610603 54439     496661.9     724544.9 0.089156
```

B3.4 Estimation for domains

SAMPLE2 is also used to demonstrate domain estimation.

```
Nd <- as.numeric(table(pop$DOM01)) # population size of domains
nd <- as.numeric(table(SAMPLE2$DOM01)) # sample size of domains
print(data.frame(
  Domain = sort((unique(pop$DOM01))),
  Sample = nd,
  Population = Nd
), row.names=FALSE)
```

```
 Domain Sample Population
      0     12         70
      1      8         30
```

The unconditional analysis (Table 3.11, Scenario 3) is obtained with `survey` function `svyby`:

```
res_ucond <- svyby(~CATCH,~DOM01,des,svytotal,vartype=c('se','ci','cv'))
print(res_ucond, row.names=FALSE)
```

```
 DOM01    CATCH       se      ci_l     ci_u        cv
     0 419535.9 84343.75 254225.20 584846.6 0.2010406
     1 191067.5 50990.11  91128.69 291006.3 0.2668697
```

Confidence intervals produced directly by `svyby` are based on the normal distribution. In order to obtain confidence intervals based on the t-distribution (which is more appropriate for small samples), we need a separate call to `confint`:

```
res_ucond[,c('ci_l','ci_u')] <- confint(res_ucond, df=degf(des))
```

```
Scenario 2. Unconditional approach
```

```
 Domain  Total  s.e. lower 95% CL upper 95% CL       cv
      0 419536 84344    243002.41     596069.4 0.201041
      1 191067 50990     84343.95     297791.0 0.266870
```

B4 Section 3.4.4: Systematic sampling example

B4.1 Preliminaries

We use the population of active vessels in SIMPOP and sort it by GT

```
source('read_population.r')
library(sampling)
library(survey)
pop <- subset(SIMPOP, ACTIVITY == 1)
N <- nrow(pop)
pop <- pop[order(pop$GT),c('ID','CATCH','GT'),]
pop$newID <- 1:N # indicating the position in the ordered population
```

`First ten vessels in the sorted population`

```
 ID    CATCH    GT newID
  9 2752.963 210.6     1
  7 2642.640 218.4     2
 13 3453.840 221.4     3
 22 3538.136 229.6     4
 24 4962.480 232.0     5
 11 5458.752 234.0     6
 10 5529.550 244.4     7
  4 5055.050 252.5     8
 31 6538.224 255.2     9
  6 6481.075 257.4    10
```

B4.2 Sample selection

Function `UPsystematic` in package `sampling` can also do unequal probability systematic sampling with a vector of inclusion probabilities given as the first argument. Equal probability systematic sampling is, of course, obtained by giving equal inclusion probabilities. `UPsystematic` returns a vector of sample inclusion indicators 1 (included in the sample) or 0 (excluded).

```
n <- 20
s <- subset(pop,
  UPsystematic(rep(n/N,N))==1,
  select=c('newID','CATCH','GT'))
```

`First six vessels in the systematic sample`

```
 newID    CATCH    GT
     5 4962.480 232.0
    10 6481.075 257.4
    15 5115.130 269.7
    20 8402.750 275.5
    25 3541.440 280.0
    30 2776.712 286.2
```

B4.3 Estimation

HT estimator for CATCH total from equal probability systematic sample is calculated exactly as for SRSWOR (Section 3.3.5). The standard error estimate is approximate as discussed in Sections 3.4.3 - 3.4.5 of the main text.

```
des <- svydesign(
  ids=~1,
  fpc=rep(N, n),
```

```
  data=s)
svytotal(~CATCH, des)
```

```
       total    SE
CATCH 647994 35765
```

Demonstration of more appropriate estimators of standard error for systematic sampling is beyond the scope of these guidelines.

B5 Section 3.5.4 PPS sampling example

B5.1 Preliminaries

We use the population of active vessels in SIMPOP.

```
source('preliminaries.r')
library(sampling)
library(survey)
pop <- subset(SIMPOP, ACTIVITY == 1)
N <- nrow(pop)
```

B5.2 Sample selection

Although base R function `sample` allows argument `prob` ("a vector of probability weights for obtaining the elements of the vector being sampled"), it does *not* in general produce a PPS sample with inclusion probabilities equal to these probability weights. Package `sampling` contains several functions for proper PPS sampling, `UPtille` among others.

Function `UPtille` takes as its first argument a vector of desired inclusion probabilities - here we make them proportional to size variable GT - and returns a vector of sample inclusion indicators 1 (included in the sample) or 0 (excluded).

```
n <- 5
pop$pik <- inclusionprobabilities(pop$GT,n)
s3 <- pop[UPtille(pop$pik)==1,]
s3$SamplingWeight <- 1/s3$pik
```

```
 Obs ID     CATCH    GT                pik SamplingWeight
   1 13   3453.84 221.4 0.0336510986004548       29.71671
   2 18   3052.56   322 0.0489415255164699       20.43255
   3 32 5565.7728 345.1 0.0524525479991731       19.06485
   4 55  6865.416   408 0.0620128646295643       16.12569
   5 71 4031.7084 370.8 0.0563587505015746       17.74347
 Sum                                           103.08327
```

We demonstrate estimation using the sample of 5 vessels listed in Table 3.13. rather than the sample drawn above.

```
n <- 5
SAMPLE3.Obs.ID <- data.frame(
  Obs=1:n,
  ID=c(65, 89, 27, 53, 94)
)
SAMPLE3 <- merge(SAMPLE3.Obs.ID, pop)
SAMPLE3$SamplingWeight <- 1/SAMPLE3$pik
SAMPLE3 <- SAMPLE3[order(SAMPLE3$Obs),]
```

```
 Obs ID      CATCH    GT SamplingWeight
   1 65    3799.95   329       19.99781
   2 89  6845.8104 343.2       19.17040
```

```
   3 27   6087.564 345.1       19.06485
   4 53  7601.8734 376.2       17.48878
   5 94 10615.9872 436.8       15.06245
 Sum                           90.78430
```

B5.3 Estimation

To obtain HT estimator for CATCH total, we first specify the sampling design using function `svydesign` and then compute the estimator and its quality indicators using function `svytotal` and extractor functions `coef` (the estimated total), `SE` (standard error), `confint` (confidence intervals), and `cv` (coefficient of variation); cf. Table 3.14 in the main text. All these functions (or their methods for `svystat` objects produced by `svytotal`) are from package `survey`.

```
des <- svydesign(
  ids=~1,
  fpc=rep(N, n),
  weights=~SamplingWeight,
  data=SAMPLE3
)
res <- svytotal(~CATCH, des)
```

```
  Total  s.e. lower 95% CL upper 95% CL       cv
 616136 67055     429961.9     802310.9 0.108831
```

B6 Section 3.6.4 Stratified sampling example

B6.1 Preliminaries

We use the population of active vessels in SIMPOP and divide it to three nearly equal-sized strata (new variable STR3) according to the values of variable GT. For stratified sampling function `strata` in package `sampling`, the population must be sorted in ascending order by the stratifying variable. Within strata, we order by ID.

```
source('preliminaries.r')
library(sampling)
library(survey)
pop <- subset(SIMPOP, ACTIVITY == 1)
N <- nrow(pop)
pop$STR3 <- as.numeric(cut_number(pop$GT, 3, right=FALSE)) # the last argumen
t was needed to make the same division as in the main text
pop <- pop[order(pop$STR3, pop$ID),]
( Ns <- table(pop$STR3) ) # population sizes of strata
```

```
 1  2  3
33 33 34
```

B6.2 Sample selection

The required arguments to `sampling` function `strata` are the data frame containing the population, name of the stratifying variable in that data frame, and a vector of sample sizes with length equal to the number of unique values of the stratifying variable and with order corresponding to the order of the strata in the sorted population (see above). `"srswor"` is the default method, but it was included in the call to avoid unnecessary messages in the printed output.

`strata` adds the inclusion probabilities as variable `Prob` and function `getdata` offers a safe way to combine the actual data to the frame element indicators returned by `strata`.

```
s5 <- getdata(pop, sampling::strata(pop, "STR3", c(6, 7, 7), method="srswor",
description=TRUE))
Stratum 1

Population total and number of selected units: 33 6
Stratum 2

Population total and number of selected units: 33 7
Stratum 3

Population total and number of selected units: 34 7
Number of strata  3
Total number of selected units 20
s5$SamplingWeight <- 1/s5$Prob
```

We obtain a different sample from that in Table 3.17(a), but the numbers of sampled vessels by strata are the same, as well as, the resulting sampling weights.

```
 Obs  ID STR3    GT      CATCH              Prob SamplingWeight
   1  11    1   234   5458.752 0.181818181818182       5.500000
   2  17    1 286.2  2776.7124 0.181818181818182       5.500000
   3  33    1 263.9   3871.413 0.181818181818182       5.500000
   4  41    1   282     3651.9 0.181818181818182       5.500000
   5  42    1 291.4   8811.936 0.181818181818182       5.500000
   6  54    1 277.2  3009.8376 0.181818181818182       5.500000
   7   5    2   312   4687.176 0.212121212121212       4.714286
   8  16    2 310.5   4160.079 0.212121212121212       4.714286
   9  20    2 305.2    4158.35 0.212121212121212       4.714286
  10  27    2 345.1   6087.564 0.212121212121212       4.714286
  11  51    2 320.1   6638.874 0.212121212121212       4.714286
  12  69    2 316.8  8709.4656 0.212121212121212       4.714286
  13 100    2   336   3958.752 0.212121212121212       4.714286
  14  60    3   378   8551.872 0.205882352941176       4.857143
  15  67    3   399   7160.055 0.205882352941176       4.857143
  16  70    3 417.6   8218.368 0.205882352941176       4.857143
  17  73    3 377.4  8405.4528 0.205882352941176       4.857143
  18  77    3 377.4  6314.2794 0.205882352941176       4.857143
  19  88    3   418     8025.6 0.205882352941176       4.857143
  20  94    3 436.8 10615.9872 0.205882352941176       4.857143
 Sum                                                 100.000000
```

Function `strata` can also produce a stratified PPSWOR sample (as in Table 3.17(b)). However, for fixed stratum sample sizes, only the systematic sampling option is available. Other alternatives would be the balanced sampling method implemented as function `samplecube` in package `sampling` or separate PPS samples (Section 3.5) for each stratum.

```
s6 <- getdata(pop, sampling::strata(pop, "STR3", c(6, 7, 7), method="systemat
ic", pik=pop$GT_DAS))
s6$SamplingWeight <- 1/s6$Prob
```

```
 Obs ID STR3   GT_DAS     CATCH              Prob SamplingWeight
   1  6    1  59459.4 6481.0746 0.220962529577015       4.525654
   2 13    1    44280   3453.84 0.164552969079241       6.077071
   3 24    1    43152   4962.48 0.160361104826274       6.235926
   4 35    1    53508  6046.404 0.198845986212557       5.029018
   5 42    1    69936  8811.936 0.259895583683961       3.847699
```

```
   6 62    1    48720   5797.68 0.181052860287729       5.523249
   7  2    2    64310   6559.62 0.242821312113227       4.118255
   8 18    2    38640   3052.56 0.145896680143914       6.854166
   9 30    2  64365.5  7208.936 0.243030868680204       4.114704
  10 48    2  48470.4 6107.2704 0.183014245477421       5.464056
  11 56    2    78302  6185.858 0.295652221755402       3.382352
  12 69    2  75081.6 8709.4656 0.283492654759143       3.527428
  13 80    2    61420   6879.04  0.23190926745443       4.312031
  14 55    3    86904  6865.416 0.236646798794281       4.225707
  15 63    3    83496 10270.008 0.227366532174898       4.398185
  16 73    3  65667.6 8405.4528 0.178818320497369       5.592268
  17 82    3  76612.2  9959.586 0.208621373913597       4.793373
  18 88    3    83600    8025.6 0.227649732799433       4.392713
  19 93    3 103564.5  9942.192 0.282014721919938       3.545914
  20 99    3    62560   4879.68 0.170356067989623       5.870058
 Sum                                                   95.829826
```

Rather than using the samples drawn above, estimation from stratified samples is demonstrated using the samples listed in the main text (Table 3.17). While `strata` computed the inclusion probabilities for us, in this case we need to do it by hand.

```
n <- 20
SAMPLE5.Obs.ID <- data.frame(
  Obs=1:n,
  ID=c(1, 22, 23, 25, 42, 44, 12, 15, 30, 36, 46, 52, 75, 55, 57, 67, 82, 90,
91, 94)
)
SAMPLE5 <- merge(SAMPLE5.Obs.ID, pop)
w <- data.frame(STR3=1:3, Prob=c(6, 7, 7)/c(33, 33, 34))
w$SamplingWeight <- 1/w$Prob
SAMPLE5 <- merge(SAMPLE5, w)
SAMPLE5 <- SAMPLE5[order(SAMPLE5$Obs),]
```

```
 Obs ID STR3      CATCH              Prob SamplingWeight
   1  1    1    3541.44 0.181818181818182       5.500000
   2 22    1   3538.136 0.181818181818182       5.500000
   3 23    1    8402.75 0.181818181818182       5.500000
   4 25    1   4978.662 0.181818181818182       5.500000
   5 42    1   8811.936 0.181818181818182       5.500000
   6 44    1  4421.9175 0.181818181818182       5.500000
   7 12    2   8644.482 0.212121212121212       4.714286
   8 15    2  3786.0615 0.212121212121212       4.714286
   9 30    2   7208.936 0.212121212121212       4.714286
  10 36    2   5855.208 0.212121212121212       4.714286
  11 46    2   8100.048 0.212121212121212       4.714286
  12 52    2  4888.3428 0.212121212121212       4.714286
  13 75    2  9652.4416 0.212121212121212       4.714286
  14 55    3   6865.416 0.205882352941176       4.857143
  15 57    3  6364.0962 0.205882352941176       4.857143
  16 67    3   7160.055 0.205882352941176       4.857143
  17 82    3   9959.586 0.205882352941176       4.857143
  18 90    3    8803.08 0.205882352941176       4.857143
  19 91    3  7823.1153 0.205882352941176       4.857143
  20 94    3 10615.9872 0.205882352941176       4.857143
 Sum                                          100.000000
```

```
n <- 20
SAMPLE6.Obs.ID <- data.frame(
  Obs=1:n,
  ID=c(44, 41, 35, 11, 38, 37, 20, 48, 80, 12, 69, 56, 50, 58, 79, 43, 61, 57
, 91, 81)
)
SAMPLE6 <- merge(SAMPLE6.Obs.ID, pop)
# Compute inclusion probabilities proportional to GT_DAS by strata
# with stratum sample sizes 6, 7, 7
str_stats <- aggregate(GT_DAS~STR3, pop, sum)
str_stats <- within(str_stats,{
    n <- c(6, 7, 7);
    f <- n/GT_DAS # Here GT_DAS is the stratum sum, and inclusion probabiliti
es for population/sample elements thus obtained as f * GT_DAS
})
SAMPLE6 <- merge(SAMPLE6, subset(str_stats, select=c(STR3, f)))
SAMPLE6 <- within(SAMPLE6,{
    Prob <- f*GT_DAS;
    SamplingWeight <- 1/Prob
})
SAMPLE6 <- SAMPLE6[order(SAMPLE6$Obs),]
```

Obs	ID	STR3	CATCH	GT_DAS	Prob	SamplingWeight
1	44	1	4421.9175	46546.5	0.172975717598168	5.781158
2	41	1	3651.9	52170	0.193873721699729	5.157997
3	35	1	6046.404	53508	0.198845986212557	5.029018
4	11	1	5458.752	56862	0.21131009322005	4.732382
5	38	1	6288.66	64170	0.238468022263209	4.193434
6	37	1	6682.5	67500	0.250842940669575	3.986558
7	20	2	4158.35	38150	0.144046541084118	6.942201
8	48	2	6107.2704	48470.4	0.183014245477421	5.464056
9	80	2	6879.04	61420	0.23190926745443	4.312031
10	12	2	8644.482	68607	0.259045898929438	3.860320
11	69	2	8709.4656	75081.6	0.283492654759143	3.527428
12	56	2	6185.858	78302	0.295652221755402	3.382352
13	50	2	7179.0048	83476.8	0.315191200544448	3.172677
14	58	3	4519.008	57936	0.157764532529521	6.338560
15	79	3	5227.508	68783	0.187301813051954	5.338977
16	43	3	6359.2191	71451.9	0.194569449079088	5.139553
17	61	3	7173.6	85400	0.232551282070234	4.300127
18	57	3	6364.0962	87179.4	0.237396735832714	4.212358
19	91	3	7823.1153	101598.9	0.276662230116223	3.614516
20	81	3	13391.04	103008	0.280499326270382	3.565071
Sum						92.050773

B6.3 Estimation

Estimation using package `survey` proceeds in similar manner as in the earlier examples. We only need to provide the `strata` argument to `svydesign` and provide stratum-specific `fpc` (population size).

```
# first add to each sample element the population size of its stratum
# from table Ns created in the Prelimaries section
SAMPLE5 <- merge(SAMPLE5,
  data.frame(STR3=names(Ns), N=as.numeric(Ns)))
# then use those to determine appropriate fpc
```

```
des <- svydesign(
  ids = ~1,
  strata = ~STR3,
  fpc = ~N,
  data = SAMPLE5
)
res <- svytotal(~CATCH, des)
```

Table 3.19(a) STR_SRSWOR

```
  Total  s.e. lower 95% CL upper 95% CL      cv
 691976 41692     604014.1       779937 0.06025
```

```
SAMPLE6 <- merge(SAMPLE6,
  data.frame(STR3=names(Ns), N=as.numeric(Ns)))
des <- svydesign(
  ids = ~1,
  strata = ~STR3,
  fpc = ~N,
  weights=~SamplingWeight,
  data = SAMPLE6
)
res <- svytotal(~CATCH, des)
```

Table 3.19(b) STR_PPSWOR

```
  Total  s.e. lower 95% CL upper 95% CL       cv
 576254 22282       529243     623265.3 0.038667
```

B7 Section 4.2.3: Ratio and regression estimation examples

B7.1 Sample selection

We use SRSWOR samples SAMPLE1 and SAMPLE2 (Section 3.3.5), renamed SAMPLE7 and SAMPLE8 to emphasize the assumption that we now have access to auxiliary variables GT, DAS, and DOM01. We also assume that the population totals of the auxiliary variables are known.

```
SAMPLE7 <- subset(pop,ID %in% c(1,44,49,55,93))
n7 <- nrow(SAMPLE7)
SAMPLE7$SamplingWeight <- N/n7
GTtot <- sum(pop$GT)
DAStot <- sum(pop$DAS)
DOM01tot <- sum(pop$DOM01)
```

```
 Obs ID      CATCH    GT DAS DOM01 SamplingWeight
   1  1    3541.44   280 136     0             20
   2 44  4421.9175 282.1 165     1             20
   3 49 11355.9732 386.1 228     0             20
   4 55   6865.416   408 213     0             20
   5 93   9942.192 440.7 235     1             20
 Sum                                          100
```

```
SAMPLE8 <- subset(pop,
  ID %in% c( 1, 9, 29, 41, 47, 56, 63, 68, 69, 71,
             78, 94, 7, 20, 22, 24, 34, 37, 51, 79)
)
n8 <- nrow(SAMPLE8)
SAMPLE8$SamplingWeight <- N/n8
```

B7.2 Ratio estimation

C.f. Table 4.6; for more information, see the help file of `survey` function `calibrate`.

```
des7 <- svydesign(
  ids=~1,
  fpc=rep(N, n7),
  data=SAMPLE7)
( est.ratio <- svyratio(~CATCH, ~GT, des7) )
```

```
Ratio estimator: svyratio.survey.design2(~CATCH, ~GT, des7)
Ratios=
            GT
CATCH 20.10515
SEs=
           GT
CATCH 2.83109
```

```
predict(est.ratio, total=GTtot)
```

```
$total
          GT
CATCH 661387

$se
            GT
CATCH 93132.68
```

Using calibration:

```
des7.calib <- calibrate(des7, ~GT-1, pop=GTtot, variance=1)
svytotal(~CATCH, des7.calib)
```

```
       total    SE
CATCH 661387 93133
```

B7.3 Regression estimation

SAMPLE7 with one auxiliary variable GT. For more information, see the help file of `survey` function `svyglm`.

```
( reg.model <- svyglm(CATCH~GT, des7) )
```

```
Independent Sampling design
svydesign(ids = ~1, fpc = rep(N, n7), data = SAMPLE7)

Call:  svyglm(formula = CATCH ~ GT, design = des7)

Coefficients:
(Intercept)           GT
   -6238.68        37.46

Degrees of Freedom: 4 Total (i.e. Null);  3 Residual
Null Deviance:      4.6e+07
Residual Deviance: 15170000     AIC: 94.82
```

```
predict(reg.model, newdata=data.frame(GT=GTtot), total=N)
```

```
    link    SE
1 608586 68403
```

SAMPLE8 with auxiliary variables GT, DAS, and DOM01.

```
des8 <- svydesign(
  ids=~1,
  fpc=rep(N, n8),
  data=SAMPLE8)
( reg.model <- svyglm(CATCH~GT+DAS+DOM01, des8) )
```

```
Independent Sampling design
svydesign(ids = ~1, fpc = rep(N, n8), data = SAMPLE8)

Call:  svyglm(formula = CATCH ~ GT + DAS + DOM01, design = des8)

Coefficients:
(Intercept)           GT          DAS        DOM01
   -6329.23        20.55        33.35      -545.37

Degrees of Freedom: 19 Total (i.e. Null);  16 Residual
Null Deviance:      140800000
Residual Deviance: 29150000     AIC: 350.6
```

```
predict(reg.model,
  newdata=data.frame(GT=GTtot, DAS=DAStot, DOM01=DOM01tot),
  total=N)
```

```
    link    SE
1 637401 28485
```

The estimates are equal to those obtained with SAS SURVEYREG (Tables 4.8 and 4.10), but s.e.'s somewhat different.

B8 Section 4.3.3: Post-stratification example

B8.1 Sample selection

We use SRSWOR sample SAMPLE2 (Section 3.3.5), renamed SAMPLE9 to emphasize the assumption that we now have access to binary auxiliary variable DOM01. We also assume that its population frequencies (domain sizes) are known.

```
SAMPLE9 <- subset(pop,
  ID %in% c( 1, 9, 29, 41, 47, 56, 63, 68, 69, 71,
             78, 94, 7, 20, 22, 24, 34, 37, 51, 79)
)
n9 <- nrow(SAMPLE9)
SAMPLE9$SamplingWeight <- N/n9
Nps <- table(pop$DOM01)
Npop <- data.frame(DOM01=names(Nps), Freq=as.numeric(Nps))
```

B8.2 Estimation

C.f. Table 4.14.

```
des9 <- svydesign(
  ids=~1,
  fpc=rep(N, n9),
  data=SAMPLE9)
des.ps <- postStratify(des9, ~DOM01, Npop)
svytotal(~CATCH, des.ps)
```

```
       total    SE
CATCH 632759 55889
```

B9 Section 5.5. Example on treating nonresponse

B9.1 Sample selection

We use SRSWOR sample SAMPLE2 (Section 3.3.5), renamed SAMPLE10 to emphasize the assumption that we now have access to continuous auxiliary variable GT and categorical variable POST5 obtained by dividing the population to five equally-sized post-strata according to the values of GT. We also assume that the population total of GT and population sizes of the post-strata are known. Furthermore, measurements of the target variable CATCH are missing for two records.

```
pop$POST5 <- as.numeric(cut_number(pop$GT, 5))
( GTtot <- sum(pop$GT) )
```

```
[1] 32896.4
```

```
Nps <- table(pop$POST5)
( Npop <- data.frame(POST5=names(Nps), Freq=as.numeric(Nps)) )
```

```
  POST5 Freq
1     1   20
2     2   20
3     3   21
4     4   19
5     5   20
```

```
SAMPLE10 <- subset(pop,
  ID %in% c( 1, 9, 29, 41, 47, 56, 63, 68, 69, 71,
             78, 94, 7, 20, 22, 24, 34, 37, 51, 79)
)
n10 <- nrow(SAMPLE10)
SAMPLE10$SamplingWeight <- N/n10
SAMPLE10$CATCH[SAMPLE10$ID %in% c(37, 51)] <- NA
SAMPLE10$I <- as.numeric(!is.na(SAMPLE10$CATCH))
SAMPLE10 <- SAMPLE10[order(SAMPLE10$POST5, SAMPLE10$ID),]
```

```
 Obs ID  I      CATCH    GT POST5 SamplingWeight
   1  7  1    2642.64 218.4     1              5
   2  9  1  2752.9632 210.6     1              5
   3 22  1   3538.136 229.6     1              5
   4 24  1    4962.48   232     1              5
   5 29  1  7518.9576 266.8     1              5
   6 37  0       <NA>   270     1              5
   7  1  1    3541.44   280     2              5
   8 20  1    4158.35 305.2     2              5
   9 41  1     3651.9   282     2              5
  10 34  1   4363.008   312     3              5
  11 51  0       <NA> 320.1     3              5
  12 56  1   6185.858 319.6     3              5
  13 69  1  8709.4656 316.8     3              5
  14 78  1   6219.108 321.9     3              5
  15 47  1  8715.8907 359.7     4              5
  16 71  1  4031.7084 370.8     4              5
  17 63  1  10270.008   392     5              5
  18 68  1 11693.8944 399.6     5              5
```

```
  19 79  1   5227.508   407     5              5
  20 94  1 10615.9872 436.8     5              5
 Sum    18                                    100
```

B9.2 Estimation

No adjustment for non-response (Table 5.2 b)

```
des10 <- svydesign(
  ids=~1,
  fpc=rep(N, n10),
  data=SAMPLE10)
svytotal(~CATCH, des10, na.rm=TRUE)
```

```
       total    SE
CATCH 543997 65830
```

Post-stratification (Table 5.2 c)

```
des.ps <- postStratify(des10, ~POST5, Npop)
svytotal(~CATCH, des.ps, na.rm=TRUE)
```

```
       total    SE
CATCH 564206 47830
```

Regression estimation (Table 5.2 d)

```
reg.model <- svyglm(CATCH~GT, des10)
predict(reg.model, newdata=data.frame(GT=GTtot), total=N)
```

```
    link    SE
1 647368 47931
```

As in the regression estimation example (Section 4.2.3), the estimates are equal to those obtained with SAS SURVEYMEANS and SURVEYREG, but s.e.'s somewhat different.

B10 References

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Tillé, Y. and Matei, A. (2016). sampling: Survey Sampling. R package version 2.8. https://CRAN.R-project.org/package=sampling

Lumley T. (2019). survey: Analysis of complex survey samples. R package version 3.35-1. https://CRAN.R-project.org/package=survey