ISSN 1977-0375

eurostat
Methodologies and
Working papers

Handbook on precision requirements and variance estimation for ESS households surveys

2013 edition

eurostat
EUROPEAN COMMISSION

eurostat
Methodologies and
Working papers

Handbook on precision requirements and variance estimation for ESS household surveys

2013 edition

eurostat
EUROPEAN COMMISSION

Europe Direct is a service to help you find answers to your questions about the European Union.

Freephone number (*)**:**
00 800 6 7 8 9 10 11

(*) The information given is free, as are most calls (though some operators, phone boxes or hotels may charge you).

More information on the European Union is available on the Internet (http://europa.eu).

Cataloguing data can be found at the end of this publication.

Luxembourg: Publications Office of the European Union, 2013

ISBN 978-92-79-31197-0
ISSN 1977-0375
doi:10.2785/13579
Cat. No: KS-RA-13-029-EN-N

Theme: General and regional statistics
Collection: Methodologies & Working papers

Acknowledgments

The European Commission expresses its gratitude and appreciation to the following members of the ESS[1] Task Force, for their work on Precision Requirements and Variance Estimation for Household Surveys:

Experts from the European Statistical System (ESS):

Martin Axelson	Sweden — Statistics Sweden
Loredana Di Consiglio	Italy — ISTAT
Kari Djerf	Finland — Statistics Finland
Stefano Falorsi	Italy — ISTAT
Alexander Kowarik	Austria — Statistics Austria
Mārtiņš Liberts	Latvia — CSB
Ioannis Nikolaidis	Greece — EL.STAT

Experts from European universities:

Yves Berger	United Kingdom — University of Southampton
Ralf Münnich	Germany — University of Trier

Coordinators of the Task Force from Eurostat unit 'Quality, methodology and research' (formerly 'Methodology and research'):

Jean-Marc Museux	Chairman
Denisa Camelia Florescu	Coordinator

Methodologists from elsewhere:

Nicola Massarelli	Eurostat unit 'Labour market'
Albrecht Wirthmann	Eurostat unit 'Innovation and information society' (formerly 'Information society; tourism')
Onno Hoffmeister	FAO[2]

The handbook received significant contributions from **Guillaume Osier** (Luxembourg — STATEC), working for **SOGETI S.A.**, and specific contributions from **Michele D'Alò** (Italy — ISTAT) and from **AGILIS S.A.** experts.

The handbook received useful comments from the reviewers **Julia Aru** (Estonia — Statistics Estonia), **Harm Jan Boonstra** (Netherlands — CBS), **László Mihályffy** (Hungary — Hungarian Central Statistical Office), **Karim Moussallam** (France — INSEE) and **Paul Smith** (UK — Office for National Statistics), and from the **DIME**[3] members.

[1] European Statistical System.

[2] Initially attached to Eurostat unit 'Government and sector accounts; Financial indicators', member of the Eurostat network of methodologists.

[3] Directors of Methodology of the National Statistical Institutes (NSIs) of the European Statistical System (ESS).

Foreword

The quality of European statistics is essential for users. European legislation requires that a certain level of quality of statistics is guaranteed, and this quality has to be assessed.

One important dimension of quality is accuracy. The accuracy is in the general statistical sense the degree of closeness of estimates to the true values and its components are variance and bias. The scope of this initiative is the variance requirements and estimation.

Different statistical domains have been confronted with similar needs related to the variance of estimates. These needs range from setting up precision (variance) requirements and a process to assess compliance to the requirements (in EU Labour Force Survey – LFS), to developing procedures to streamline the production of standard errors of European statistics on the basis of the information provided by National Statistical Institutes (NSIs) of the European Statistical System - ESS (in the Community Survey on ICT[4] Usage in Households and by Individuals).

The initiatives launched by different statistical domains on similar issues called for a harmonisation of the methods in the ESS. In agreement with the ESS Directors of Methodology (DIME), the Eurostat Directorate 'Methodology, corporate statistical and IT services' (in charge of the methodological coordination and support both at Eurostat and ESS level) set up a Task Force (TF) with a generic mandate to issue general recommendations on variance requirements and estimation for ESS household surveys. The implementation of the general recommendations and the specific agreements at stake for LFS and ICT are decided by the domain specialists. Actually, a LFS domain specialist in Eurostat set up a domain specific TF which run in parallel with the DIME TF, discussed the general recommendations and provided valuable feedback to the DIME TF. The coordination between the two TFs was ensured by the Eurostat methodologists. With respect to the ICT, domain specialists in Eurostat are currently assessing the use of methods to estimate standard errors centrally in Eurostat.

For efficiency reasons, the DIME TF was a think tank composed of a limited number of high profile experts in Member States, two high profile academic experts and the Eurostat methodologists involved in their respective projects. The handbook prepared under the auspices of the DIME TF was additionally submitted for review by experts designated from other Member States than those who participated to the DIME TF and was subsequently developed on specific issues.

We expect that the general recommendations issued within this coordinated approach will provide a basis for a more harmonised/standardised approach of similar problems in other surveys.

Antonio Baigorri
Head of Unit

Daniel Defays
Director

[4] Information and Communication Technology.

Table of contents

1. Introduction

The objective of this handbook is to present the results of the work on Precision Requirements and Variance Estimation for ESS Household Surveys, and more specifically the general recommendations issued by the Task Force, set up under the auspices of the DIME. The handbook covers only the variance component of accuracy (and not the bias).

The recommendations are in line with the 'ESS handbook for quality reports' (Eurostat, 2009a) and with the mandate of the Task Force. They comprise:

- a recommendation for a standard formulation of precision requirements in EU regulations, by taking into account survey specificities such as indicators and regional disaggregation. The point of formulating these requirements is to achieve uniform and unambiguous understanding between the National Statistical Institutes (NSIs) and Eurostat;
- a review of variance estimation methods, with a view to establishing a more harmonised approach when computing standard errors and confidence intervals for statistics at national and EU levels. The handbook recommends good practices and identifies bad ones which should be avoided;
- a recommendation for an integrated approach to the increased availability of standard errors in the ESS, with a view to achieving a fully harmonised approach. The recommendation assesses a range of possible approaches — from fully centralised to decentralised;
- a recommendation on how to assess NSIs' compliance with the precision requirements.

2. Precision requirements

One objective of this chapter is to identify and assess the existing approaches for setting up precision requirements and to propose one of them as best practice. Other objectives are to provide the appropriate precision measures for each type of indicator and to introduce the concept of domains and how to handle them when defining precision requirements. This chapter sets out known examples of precision thresholds/sizes used in different contexts and by different institutions. These are not meant to be prescriptive but to give some feasible benchmark when defining precision thresholds. Finally, the chapter proposes a set of standard formulations of precision requirements for regulations, for level (annual, quarterly, monthly, etc.) estimates, and for estimates of net change (for overall national populations and for national breakdowns).

2.1 The two approaches for specifying precision requirements

Michele D'Alò and Stefano Falorsi (ISTAT)

There are two main strategies for setting up precision requirements: specifying ***minimum effective sample sizes*** with which the NSIs have to comply, or ***precision thresholds*** that have to be met by the main target indicators of the survey. Both strategies can be defined at either the planning or estimation stage, after the survey has been carried out.

The approach to setting up ***precision thresholds*** for survey estimates has already been used in the EU Labour Force Survey (EU-LFS) and in the Community Survey on Information and Communication Technology (ICT).

The EU-LFS Framework Regulation (Council Regulation No 577/98 of 9 March 1998) introduced precision requirements in the form of precision thresholds which have to be met over certain sub-populations by estimates of annual averages and estimates of changes over two consecutive quarters. This ensures that the EU-LFS national samples can achieve a significant degree of 'representativeness'.

Figure 2.1.1: Precision thresholds — EU-LFS

Article 3

Representativeness of the sample

1. For a group of unemployed people representing 5 % of the working age population the relative standard error for the estimation of annual averages (or for the spring estimates in the case of an annual survey in the spring) at NUTS II level shall not exceed 8 % of the sub-population in question.

Regions with less than 300 000 inhabitants shall be exempt from this requirement.

2. In the case of a continuous survey, for sub-populations which constitute 5 % of the working age population the relative standard error at national level for the estimate of changes between two successive quarters, shall not exceed 2 % of the sub-population in question.

For Member States with a population of between one million and twenty million inhabitants, this requirement is relaxed so that the relative standard error for the estimate of quarterly changes shall not exceed 3 % of the sub-population in question.

Member States whose population is below one million inhabitants are exempt from these precision requirements concerning changes.

It is important to note that the current LFS precision requirements refer to a theoretical situation in which unemployed people account for 5 % of the working-age population. These requirements are thus a reference for designing the survey, but cannot tell us anything about the quality of the actual survey results. This approach, within the context of the current arrangements for formulating precision requirements, prevents the relative standard errors used as precision thresholds from becoming meaningless (when the proportion approaches zero, the relative standard error approaches infinity). However, this is a critical point in the current formulation of precision requirements. In particular, it is difficult to say whether or not such requirements — for theoretical situations — are met when the relative standard errors can only be reliably computed for the actual estimates (Eurostat, 2010c).

A more straightforward formulation of precision requirements in the form of ***precision thresholds*** (referring to the quality of the actual estimates) is provided in the methodological manual for the ICT survey (Eurostat, 2010b). For the household survey:

> 'The estimated standard error (…) shall not exceed 2 percentage points of the overall proportions and shall not exceed 5 percentage points for the proportions related to the different subgroups of the population, where these subgroups constitute at least 10 % of the total population in the scope of the survey'.

Another approach is to formulate precision requirements in terms of ***minimum effective sample sizes*** to be achieved by the countries.

This is the case with the EU-SILC[5] Framework Regulation (European Parliament and Council Regulation No 1177/2003 of 16 June 2003).

Figure 2.1.2: Minimum effective sample sizes — EU-SILC

ANNEX II

Minimum effective sample sizes

	Households		Persons aged 16 or over to be interviewed	
	Cross-sectional	Longitudinal	Cross-sectional	Longitudinal
	1	2	3	4
EU Member States				
Belgium	4 750	3 500	8 750	6 500
Denmark	4 250	3 250	7 250	5 500
Germany	8 250	6 000	14 500	10 500
Greece	4 750	3 500	10 000	7 250
Spain	6 500	5 000	16 000	12 250
France	7 250	5 500	13 500	10 250
Ireland	3 750	2 750	8 000	6 000
Italy	7 250	5 500	15 500	11 750
Luxembourg	3 250	2 500	6 500	5 000
Netherlands	5 000	3 750	8 750	6 500
Austria	4 500	3 250	8 750	6 250
Portugal	4 500	3 250	10 500	7 500
Finland	4 000	3 000	6 750	5 000
Sweden	4 500	3 500	7 500	5 750
United Kingdom	7 500	5 750	13 750	10 500
Total of EU Member States	**80 000**	**60 000**	**156 000**	**116 500**
Iceland	2 250	1 700	3 750	2 800
Norway	3 750	2 750	6 250	4 650
Total including Iceland and Norway	**86 000**	**64 450**	**166 000**	**123 950**

The concept of effective sample size basically refers to the minimum sample size that would be required, under simple random sampling without replacement, to obtain the same level of precision as with the actual sampling design. In practice, however, many samples are selected with 'complex' designs (multi-stage selections, weight adjustment for non-response, calibration, etc.). It follows that the minimum effective sample sizes under simple random sampling have to be adjusted for design effects *Deff* , viz. the variation in design efficiency caused by sampling design components such as stratification or clustering. This leads us to the concept of achieved sample size. Design effect is found to be subject to interpretation and is not easy to forecast because it also depends on indicators, domains and the estimation methods used. See *Appendix 7.2* for more information.

If n denotes the achieved sample size, then the effective sample size n_{eff} is given by (Kalton *et al*, 2005):

$$n_{eff} = n / Deff \ . \quad (2.1.1)$$

[5] EU Statistics on Income and Living Conditions.

The achieved sample size refers to the number of (ultimate) respondents. *Therefore, the real sample size at the planning stage should be adjusted to the anticipated non-response.*

In practice the value of $Deff$ is unknown and has to be estimated. The design effect of an estimator $\hat{\theta}$ of the parameter θ is defined as the ratio between the variance $V(\hat{\theta})$ of the estimator under the actual sampling design and that which would be obtained from a hypothetical simple random sample without replacement of the same size:

$$Deff = Deff\left(\hat{\theta}\right) = \frac{V(\hat{\theta})}{V_{SRS}(\hat{\theta}^*)}. \quad (2.1.2)$$

$\hat{\theta}^*$ is an 'equivalent' estimator of θ under simple random sampling without replacement. See *Appendix 7.2* for more information.

When designing a survey, defining the minimum level of precision is a very important step: very high precision attends to waste resources, while very low precision makes the results less usable. From an EU perspective, it is desirable to have accurate statistics at national level so that we can compare not only the performance of countries against specified targets but also their performance between each other.

Specifying a minimum sample size makes it possible to calculate a confidence interval that includes the true value of the parameter with probability $(1-\alpha)$ close to 1. A common practice when determining confidence intervals consists of assuming that the estimator follows a normal distribution. A confidence interval with a given confidence level is then derived using percentile values of the normal distribution of mean 0 and variance 1. The half-length of the confidence interval represents the (absolute) margin of error of the estimator, while the relative margin of error is obtained by dividing the absolute margin of error by the estimated value of the parameter (see *Appendix 7.1*).

As a general formula, let us consider a simple random sampling without replacement, and assume that we are seeking a relative margin of error of $100 \cdot k\,\%$ for the total Y of a study variable y. Thus, the minimum sample size is given by:

$$n_{\min} = \frac{z_{1-\alpha/2}^2 N^2 S_y^2}{k^2 Y^2 + z_{1-\alpha/2}^2 N\ S_y^2}, \quad (2.1.3)$$

where S_y^2 is the variance of y over the whole target population (see *Appendix 7.1*) and $z_{1-\alpha/2}$ is the percentile value at $100(1-\alpha/2)\%$ of the normal distribution of mean 0 and variance 1. In many practical applications $\alpha = 0.05$. The population quantities Y and S_y^2 are actually unknown and have to be estimated using data from auxiliary sources (previous surveys, administrative sources, expert judgment, etc.). In the above formula k is the relative margin of error expressed as a proportion, while $100 \cdot k\,\%$ is the relative margin error expressed as a percentage. Equation (2.1.3) accounts for the finite population correction.

Calculation of minimum sample sizes had so far relied on a single estimator for which a specified level of precision was desired. This made it difficult for Eurostat to assess compliance since it required knowledge and monitoring of the actual sampling design.

Nonetheless, there are also practical situations where minimum sample sizes are not determined on the basis of a precision criterion. For instance, in many ESS surveys, budgetary constraints may be such that they put a strict limit on the total number of interviews which can

be conducted at EU level. Minimum sample sizes at country level are then determined by allocating the total number of EU-level interviews among the countries, basing the allocation method on a general compromise between EU and country accuracies. This is done by allocating a minimum number of units to the small countries, thereby ensuring a minimum level of precision in each of them.

There are also intermediate situations where budgetary constraints at EU level are weaker, so a more ambitious EU precision target can be set. *In such cases, sample sizes at country level should be adjusted for design effects and anticipated non-response.* The choice of the actual sampling design is dictated by a trade-off between reducing cost and reducing variability. *This adjustment of national sample sizes triggers additional costs which should be considered in the total budget at EU level* (given that budgetary constraints are weaker), *on condition that the non-response and design effects are kept under control*.

Determining minimum sample size works only if precision thresholds have been set first. Conversely, the only practical way for countries to take on board precision thresholds is to ensure that a minimum number of units have been sampled. The two main strategies are therefore equivalent in theory, but may differ in practice, especially for multi-purpose surveys.

Large-scale surveys are usually designed to estimate a great number of parameters with reference to many different domains of interest. In this context, precision requirements expressed as precision thresholds seem to be more flexible, even though they also refer to a reduced set of target indicators. A given effective sample may achieve satisfactory precision for one indicator but may be less satisfactory for others. Besides, sampling designs that meet design requirements may end up producing low-quality output (e.g. a minimum sample size does not continuously achieve satisfactory precision in case of dynamic phenomena, a minimum sample size does not naturally cover for all sources of variability like calibration). What really matters to data users is output quality. *Therefore, for EU regulations, precision requirements expressed as precision thresholds are recommended.* They are an important instrument in terms of quality assurance. Preference is given to output quality under the assumption that it includes all of the effects (sampling design, non-response, calibration, imputation, etc.).

With regard to allocation in large-scale surveys, bear in mind that surveys have, in most cases, multiple objectives. This means that it is unrealistic to hope for sample dimensions that guarantee predetermined levels of precision for all estimates of interest. A further problem arises from the need to produce parameter estimations for quite a large number of domains. *It is recommended to limit precision requirements to the main target indicators for key reporting domains in order to avoid cumulative conflicting constraints on a single data collection instrument*.

Though we may be seeking the optimum autonomous solution in terms of precision of estimates for each domain, the result, in practice, is a compromise of different aims, each of them demanding a specific type of response and where the solution may be at odds with the other solutions available. So we need to establish an optimum sample size and allocation in a multivariate framework, assuming that an optimum for each variable considered individually may not be optimum for the overall set of variables of interest. The extensive range of objectives and ties calls for multivariate allocation methodologies, in order to get an overall picture of how to achieve optimum determination of sample size. More precisely, determining the sample size and the sampling allocation has to be able to guarantee precision thresholds for each variable of interest and for a variety of domains. The method proposed by Bethel (1989) aimed at determining optimum size from a multivariate viewpoint in the case of a

design with one stage of stratification and with reference to a single domain. The method has been generalised for multi-purpose surveys in the context of multi-stage sampling designs when multiple domains are under study. The method is based on inflating the variance of the estimator $\hat{Y}_h$ of the total Y_h in stratum h under simple random sampling by means of an estimator of the design effect $Deff$ for each domain of interest (Falorsi and Russo, 2001).

Summary

The two main strategies for setting up precision requirements refer to:

- precision thresholds, to be met by a few main target indicators of the survey, and
- minimum effective sample sizes, to be ensured by the NSIs.

For regulations, it is recommended to express requirements by defining minimum precision thresholds to be met by a few main target indicators. Precision and accuracy are concepts that are well defined and documented in the ESS quality framework and are easily understood by users of statistics.

It is recommended to limit precision requirements to the main target indicators for key reporting domains in order to avoid cumulative conflicting constraints on a single data collection instrument.

Minimum sample size is a meaningful concept for data producers who need to design instruments to collect data and estimate costs. However, precision requirements expressed as precision thresholds seem to be more flexible, even though they also refer to a reduced set of target indicators. A given effective sample may achieve satisfactory precision for one indicator but may be less satisfactory for others. Besides, sampling designs that meet design requirements may end up producing low-quality output (e.g. a minimum sample size does not continuously achieve satisfactory precision in case of dynamic phenomena, a minimum sample size does not naturally cover for all sources of variability like calibration). Precision requirements expressed as precision thresholds — which are assumed to cover these sources — are an important instrument for quality assurance. What really matters to data users is output quality. *Therefore, for EU regulations, precision requirements expressed as precision thresholds are recommended.*

The two frameworks are nevertheless equivalent in theory: minimum sample sizes are a translation of precision thresholds in an ideal survey sampling context. The technical difficulty associated with the effective sample size framework has to do with determining the design effect that measures the distance between the actual design and the ideal situation. Design effect is found to be subject to interpretation and is not easy to forecast because it depends also on indicators, domains and the estimation methods used.

2.2 Precision measures in relation to type of indicators

Denisa Camelia Florescu and Jean-Marc Museux (Eurostat)

When setting up precision requirements for a survey, the precision measures should be geared to the type of indicators. Most of the commonly used indicators in ESS surveys belong to one of the following categories:

- the total or the mean of a continuous variable (e.g. the total or the mean household income);
- in the case of a qualitative variable, the interest generally lies in the total or the proportion of population elements in a given category (e.g. total number or proportion of unemployed people in the population);
- a non-linear function of several totals, means or proportions (ratios, regression coefficients, etc.).

First of all, we need to clarify the difference between percentages and proportions, and between proportions and ratios.

- ***Percentages*** and ***proportions*** are conceptually equivalent but are expressed in different ways. An indicator may refer to the percentage of individuals having access to Internet — which can, for instance, take the value of 50 % — or to the proportion of individuals having access to Internet — which is, in the same case, 0.5.
- Both ***ratios*** and ***proportions*** are parameters of a population. A ratio is a ratio of two totals or means. A proportion is a special case of a ratio. The numerator and the denominator are counts of elements, in the case of a proportion. The numerator is the count of elements in a domain A, and the denominator is the count of elements in a domain B. Domain A has to be a subset of domain B.

  Ratios and proportions are usually estimated by individually estimating the numerator and the denominator. It is not important whether the population parameter is a ratio or a proportion when it comes to variance estimation. The important point in variance calculations is whether or not the estimator of the denominator has sampling variability. In many practical cases, the variance of the estimator in the denominator is zero. This happens, for instance, with proportions when the population total in the denominator is known from external sources. In this case, we have to estimate the variance of the estimator in the numerator and divide the result by the squared value of the denominator. An example is the proportion of individuals having broadband connection, provided the whole number of individuals is known from an external source. On the other hand, the variance of the estimator in the denominator may be strictly positive; this means that the variability of the denominator has to be taken into account in calculations. This occurs, for example, with domain estimators, viz. estimators for sub-populations. For instance, a statistic may be the unemployment rate, where the total labour force (the population of employed and unemployed persons) is estimated from a sample of observations. In order to estimate the variance of such non-linear statistics, we often resort to the Taylor linearisation method (Tepping, 1968; Woodruff, 1971; Wolter, 2007; Osier, 2009).

  According to the above definitions, both proportions and ratios can have a constant denominator (variance of the denominator is zero) or a variable denominator (variance of the denominator is not zero).

However, for simplification purposes and in the variance estimation context, the concept of ratio is used to designate a ratio of two estimators where the denominator has a non-zero variance (a non-linear statistic), while the concept of proportion is used to designate a linear statistic (with constant denominator).

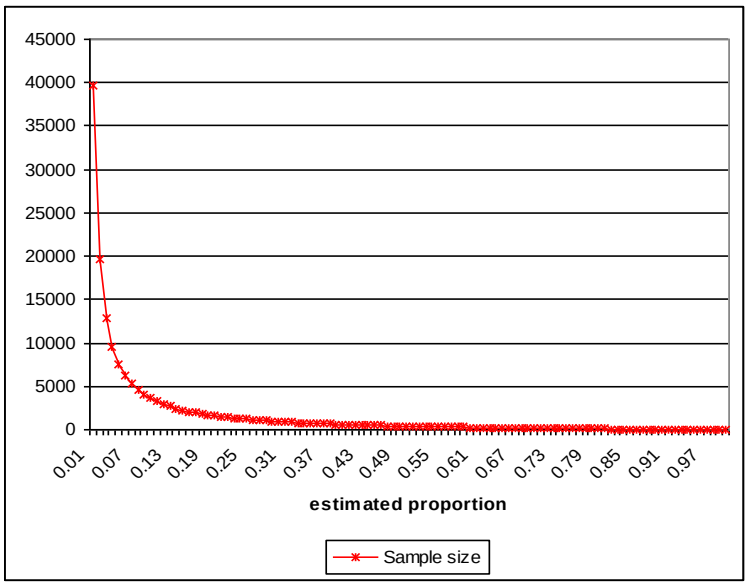
Coefficients of variation (relative standard errors) are generally not recommended for estimating the precision of percentages/proportions. This is because the value of the percentage/proportion has a strong impact on the value of the coefficient of variation, especially when the percentage/proportion is low, and because the coefficients of variation for the percentages/proportions of any characteristic are not symmetrical.

Consider a simple random sample without replacement of size n. Let us assume that we want to estimate a proportion P $(0 < P \leq 1)$ over the entire population, and that the exact size of the population (denominator of the proportion) is known from external sources. Thus, the coefficient of variation CV of the estimated proportion is given by:

$$CV = \sqrt{\frac{1-P}{nP}} \quad . \qquad (2.2.1)$$

Therefore, the lower the value of proportion P, the higher will be the coefficient of variation CV. Furthermore, let us examine the impact of the proportion on the minimum sample size needed to achieve a coefficient of variation of 5 %.

Figure 2.2.1: The impact of the proportion on the minimum sample size needed to achieve a coefficient of variation of 5 %, under simple random sampling



When a precision threshold is expressed as a coefficient of variation, the proportion has a strong impact on the minimum sample size needed to attain this threshold: the sample size tends towards infinity, as the proportion approaches zero. Therefore, the use of coefficients of variation in precision requirements would lead to more stringent conditions for countries/regions with small values of proportion, and would thus require a huge increase in sample size. This could result in values of sample sizes which can never be attained under standard budgets.

In particular, it can pose serious accuracy problems when estimating a proportion of individuals in 'rare' sub-populations, for instance, individuals having a particular profession or activity status:

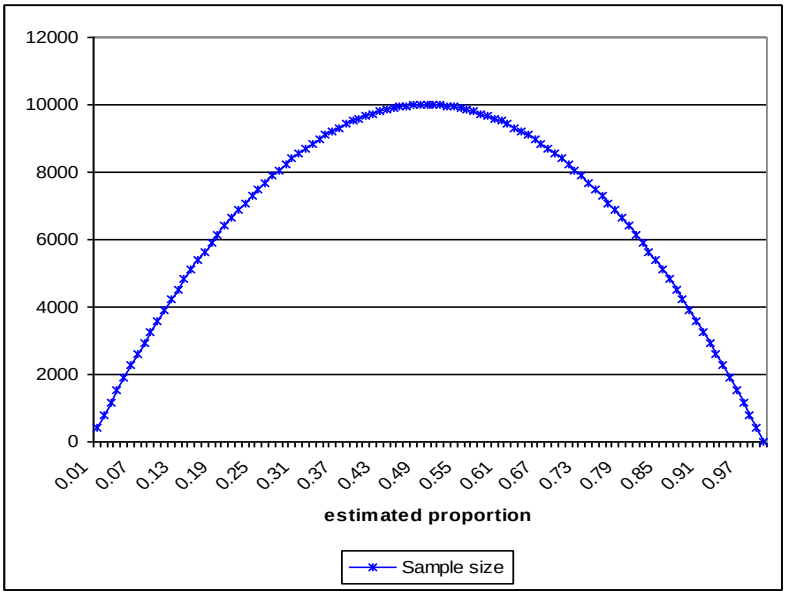
Table 2.2.1: Minimum sample size to ensure a coefficient of variation of 5 %

	Clergy	Administrators, public sector	Scientists	Retired farmers	Retired factory workers
Proportion in the total population (P)*	0.001	0.005	0.01	0.02	0.05
Minimum sample size needed (n)	400 000	80 000	40 000	20 000	8 000

**Source: French Census, 1990*

On the other hand, whenever the precision threshold is expressed using an absolute measure of accuracy like standard error, then the minimum sample size needed increases as the proportion approaches 0.5 (from both directions), albeit not to infinity. The use of standard errors in precision requirements would therefore impose less stringent conditions. As a result, survey targets expressed in terms of standard errors are more tractable.

Figure 2.2.2: The impact of the proportion on the minimum sample size needed to obtain a standard error of 0.5 percentage points, under simple random sampling



The above therefore discourages the use of coefficients of variation for percentages/proportions, but encourages and recommends the use of standard errors.

However, for specific surveys, experts should decide to use that precision measure which is the most demanding in the case of that proportion value which makes the study variable the most relevant.

This means:

- *the use of the standard error may be preferred if the study variable becomes more relevant as the estimated proportions get closer to 0.5, since it imposes more demanding requirements (bigger sample size) for proportions nearer to 0.5;*
- *the use of the coefficient of variation may be preferred if the study variable becomes more relevant as the estimated proportions tend to 0, since it imposes more demanding requirements (bigger sample size) for proportions nearer to 0. However, we should note the huge increase in the sample size whenever the proportion approaches 0 and should set a low threshold of the proportion under which the requirement does not apply;*
- *the use of either standard error or coefficient of variation is equally preferable if the study variable becomes more relevant as the estimated proportions approach 1, since they both relax the burden (lower the sample size needed).*

Summary

It is recommended to use precision measures which are geared to the type of indicators they refer to.

The general definitions of ratio and proportion are: a ratio is a ratio of two totals or means, while a proportion is a special case of a ratio where the numerator and the denominator are counts of elements in domain A and domain B respectively, where domain A is a subset of domain B. However, for simplification purposes and in the variance estimation context, the concept of *ratio* is used to designate a ratio of two estimators where the denominator has a non-zero variance (a non-linear statistic), while the concept of *proportion* is used to designate a linear statistic (with constant denominator).

Recommended precision measures are:

- *coefficients of variation and other precision measures expressed in relative terms for totals and means of continuous variables;*
- *standard errors and other precision measures expressed in absolute terms for proportions, but also for ratios and changes which are close to 0.*

The second recommendation aims to avoid situations where precision requirements lead to a huge increase in the sample size when the indicator approaches 0. Moreover, absolute precision measures for the percentages/proportions of any characteristic are symmetrical.

However, for specific surveys, experts should decide to use that precision measure which is the most demanding in the case of that proportion value which makes the study variable the most relevant.

This means:

- *the use of the standard error may be preferred if the study variable becomes more relevant as the estimated proportions get closer to 0.5;*
- *the use of the coefficient of variation may be preferred if the study variable becomes more relevant as the estimated proportions tend to 0. However, we should set a low threshold of the proportion under which the requirement does not apply;*
- *the use of either standard error or coefficient of variation is equally preferable if the study variable becomes more relevant as the estimated proportions approach 1.*

2.3 Precision requirements and reporting domains

Denisa Camelia Florescu and Jean-Marc Museux (Eurostat)

Precision thresholds and/or minimum effective sample sizes can be defined at EU level, at country level or at domain level. There are usually no precision requirements for EU estimates as their reliability is a direct consequence of the reliability of national estimates. Thus, in practice, precision requirements are mostly laid down at national and domain levels.

It is recommended that precision requirements be formulated for a certain domain level for a specific survey, whenever a regulation stipulates that reliable estimates are required at that domain level.

A ***domain*** is a subgroup of the whole target population of the survey for which specific estimates are needed. A domain may consist of a geographical area, such as a NUTS2 region, or a major population centre, e.g. capital cities. It may also comprise a specified population category, such as a major national or ethnic group (OECD). For instance, the focus may be on not only the unemployment rate of the entire population, but also of breakdowns by age, gender and education level.

Units in a domain may sometimes be identified prior to sampling. In such cases, the domain can be treated as a separate stratum from which a specific sample may be taken. Stratification ensures a satisfactory level of representativeness of the domain in the final sample: these are called ***planned domains***.

Precision thresholds and/or minimum effective sample sizes are mostly set up for planned domains. On the basis of the result (2.1.3) in Section 2.1, the minimum sample size required to achieve a relative margin of error of $100 \cdot k\,\%$, for the total Y_d of a study variable y, over a domain U_d of size N_d, is given by:

$$n_{d\_\min} = \frac{z_{1-\alpha/2}^2 N_d^2 S_{yd}^2}{k^2 Y_d^2 + z_{1-\alpha/2}^2 N_d S_{yd}^2}, \qquad (2.3.1)$$

where S_{yd}^2 is the variance of y over the domain and $z_{1-\alpha/2}$ is the percentile value at $100(1-\alpha/2)\%$ of the normal distribution of mean 0 and variance 1. In the above formula k is the relative margin of error expressed as a proportion, while $100 \cdot k\,\%$ is the relative margin of error expressed as a percentage. The population values Y_d and S_{yd}^2 are unknown and have to be estimated using data from auxiliary sources (previous surveys, administrative sources, expert judgment, etc.). Equation (2.3.1) accounts for the finite population correction. The formula (2.3.1) is applicable when the whole sample is selected by simple random sampling, in which case it can be assumed that the sample s_d of size n_d (which is supposed to be constant) from the domain U_d has been selected by simple random sampling without replacement.[6]

When several precision thresholds are defined (at both national and domain level), the minimum effective sample size should attain each precision target as specified in (2.1.3) and (2.3.1). Whenever constraints are tight, such that they lead to a sample size which cannot be

[6] The formula (2.3.1) cannot be applied if the selection probabilities are different for the units included in the same domain.

attained given the available resources, there will need to be a compromise solution, which will involve removing and/or relaxing certain objectives.

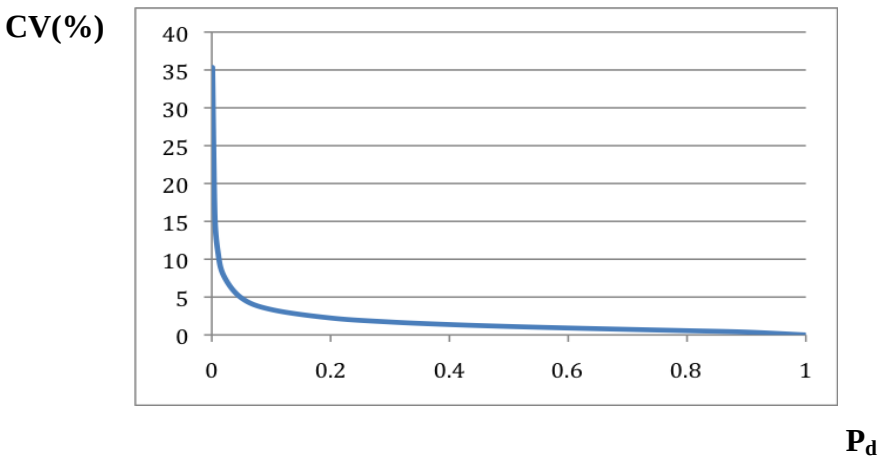
On the other hand, there are many ***unplanned domains*** for which units cannot be identified prior to sampling. The need for estimates of certain domains is often evident only after the sampling design has been decided or after the sampling and the fieldwork have been completed. However, *it is recommended that survey managers avoid setting requirements for unplanned (reporting) domains, especially for domains which represent a small share of the total population*. Sample sizes for sub-populations are random variables, since formation of these sub-populations is unrelated to sampling design. The survey manager cannot control the size of an unplanned domain sample needed to ensure compliance with established requirements. Besides, the random size of the sample builds an additional component of variability into the domain estimates (Eurostat, 2002). These two issues occur in particular in relation to rare sub-populations[7] (say, where the domain accounts for less than 5 % of the total population).

To illustrate the variability of the sample size from unplanned domains, consider a simple random sample s without replacement of size n selected from a target population U, of size N. Let s_d be the part, of size n_d, of the whole sample s which falls into a domain U_d $(U_d \subseteq U)$. n_d is a random variable which satisfies the following properties:

$$\begin{aligned} E(n_d) &= nP_d \\ V(n_d) &\cong nP_d(1-P_d) \end{aligned}, \qquad (2.3.2)$$

where $P_d = \frac{N_d}{N}$ denotes the relative size of the domain U_d in the population U. For example, with $n = 8000$ and P_d =0.05, we get a coefficient of variation for the sample n_d of nearly 5 % and a relative margin of error of nearly 10 %, which is not negligible. For a fixed sample size $n = 8000$, the lower the relative size of the domain (P_d), the higher will be the coefficient of variation for the sample size n_d from the domain.

Figure 2.3.1: The coefficient of variation (CV) (%) for the sample size n_d from the domain, plotted against the relative size P_d of the domain (fixed whole sample size n= 8000)



[7] Small domain estimation is excluded when we are talking about precision requirements for rare populations.

When unplanned domains form a part of the whole population, e.g. age groups, gender, education levels, the survey manager can estimate what the domain sample sizes would be if the national sample structure mirrored the population structure by domain. Precision measures can be calculated for unplanned domain estimates to interpret the expected domain sample sizes in terms of statistical accuracy.

The precision of estimates for unplanned domains can be improved by post-stratification. However, bias can be introduced at the same time. *The effect on the mean square error should be considered before post-stratifying.*

Summary

It is recommended that precision requirements be formulated for a certain domain level for a specific survey, whenever a regulation stipulates that reliable estimates are required at that domain level.

When several precision thresholds are defined (at both national and domain level), the minimum sample size should attain each precision target. Whenever constraints are tight, such that they lead to a sample size which cannot be attained given the available resources, a compromise solution should be sought by removing and/or relaxing certain objectives.

In sample surveys, some of the domains are unplanned, i.e. the domain units cannot be identified prior to sampling. *It is recommended that survey managers avoid setting requirements for unplanned (reporting) domains, especially for domains which represent a small share of the total population*. The survey manager cannot control the size of an unplanned domain sample to ensure compliance with requirements. In addition, the precision of estimators over unplanned domains is known to have a variance component related to the uncertainty of the sample size from such domains. These occur in particular with rare sub-populations (say, where the domain accounts for less than 5 % of the total population).

2.4 Examples of precision thresholds/sizes

Denisa Camelia Florescu and Jean-Marc Museux (Eurostat)

Specifying what degree of precision is desired is an important step when planning a sample survey. A very high level might mean wasting of resources, while a very low one might make the results less usable. In practice, questions arise as to which precision thresholds are linked to acceptable quality, but there is no common standard.

Precision thresholds are generally survey-specific and depend on users' needs and on the required reliability. Furthermore, and over and above statistical concerns, determining precision thresholds is also a political and resource-related decision.

Some examples of precision sizes/thresholds used in different contexts by different institutions are given below. They are not meant to be prescriptive but rather to give some feasible benchmarks to be used when defining precision thresholds:

- A ***coefficient of variation*** of 5 % or less means a satisfactory level of reliability for estimates, while a coefficient of variation of more than 5 % means lower reliability (Ardilly, 2006).

- In the ICT household survey, the estimated ***standard error*** may not exceed 2 percentage points for the overall proportions and 5 percentage points for the proportions relating to the different subgroups of the population, where these subgroups comprise at least 10 % of the total population within the scope of the survey (Eurostat, 2010b).
- In the EU-SILC, a methodological document (Eurostat, 2001) sets out how to use the 'compromise power allocation' method to allocate the EU sample size (which should not exceed 80 000 to 100 000 sample households) to countries. The main household income measure is the poverty rate, and varies roughly in the 5-25 % range. At national level, taking a proportion (percentage) of 15 % as the basis for computations, a simple random sample of 5 000 households is required (except perhaps for the smallest countries) to estimate this with 1 percentage point error (***the absolute margin of error***) (95 % confidence interval). This corresponds to an ***absolute standard error*** of around 0.5 percentage points.
- The European Health Interview Survey (EHIS) methodological sampling guidelines (Eurostat, 2009b) show how to allocate the EU sample size (270 000 individuals) to countries by using the 'compromise power allocation' method. This sample size derives from the consideration that an average of 10 000 or 7 500 individuals per country would make for good precision (for a sample size of 7 500 and a percentage of 10 %, ***the absolute standard error*** is 0.3 percentage points). National effective sample sizes have been computed by taking the percentage of people with severely hampered activity as the most critical indicator. This indicator was selected because of low prevalence in some Member States, which could lead to precision problems for some sub-groups. The corresponding errors in absolute percentage points (standard error in absolute terms) were then computed for the national effective samples. At national level, ***the absolute standard error*** varies from 0.1 to 0.4 percentage points. This corresponds to an ***absolute margin of error*** of between 0.2 and 0.8 percentage points (95 % confidence interval).
- At ISTAT, ***coefficients of variation*** should not exceed 15 % for domains and 18 % for small domains; when they do, this serves as an indication to use small area estimators. Note that this is just a rule of thumb and that not all domains are equivalent because they are associated with the percentage of the population they represent, and this population can vary.
- Statistics Canada applies the following guidelines on LFS data reliability (Statistics Canada, 2010):
  - if the ***coefficient of variation*** (CV) $\leq$ 16.5 % , then there are no release restrictions;
  - if 16.5 % < $CV \leq$ 33.3 %, then the data should be accompanied by a warning (release with caveats);
  - if CV > 33.3 %, then the data are not recommended for release.

Summary

There are no general precision thresholds/sizes that would hold good for all ESS surveys. They tend to be survey-specific and purpose-specific, depend on users' needs in terms of reliability, and are related to available resources.

The handbook nevertheless presents some (non-prescriptive) examples of precision thresholds/sizes used by different institutions for specific cases.

2.5 Recommendations for a standard formulation of precision requirements

Denisa Camelia Florescu and Jean-Marc Museux (Eurostat)

The DIME Task Force issued a proposal for a standard formulation of national precision requirements for percentages/proportions[8] (as this is the type of indicator most often encountered in household surveys) in EU Regulations. This is linked to the strategy of setting requirements in terms of precision thresholds (see Section 2.1). Precision thresholds refer to the actual value of the estimated indicator. Unlike precision thresholds, a compliance criterion fixed at the design stage would be difficult for Eurostat to monitor and may be fruitless since the main aim is the quality of the output.

The proposed standard formulation of precision requirements is issued for indicators of the proportion type, for national estimates of level, and for net changes in the national estimates of level, as follows:

- Precision requirements for ***estimates of level*** (e.g. annual, quarterly, etc. estimated percentages):
  - ***For overall national estimates:***

    'The survey should be designed such that *the estimate of the standard error* does not exceed ... *percentage points* for the estimated percentage ...[9] for the total reference population'.
  - ***For estimates of national breakdowns (domains):***

    'The survey should be designed such that *the estimate of the standard error* does not exceed ... *percentage points* for the estimated percentage ...[10] for the …[11] population breakdowns, where such population breakdowns comprise at least ...%[12] of the total reference population'.
- Precision requirements for ***net changes in the estimates of level*** (absolute changes in the estimated percentage between successive years, quarters, etc.)
  - ***For overall national estimates:***

    'The survey should be designed such that *the estimate of the standard error* does not exceed ... *percentage points* for the change between ...[13] of the estimated percentage ...[14] for the total reference population'.
  - ***For estimates of national breakdowns (domains):***

[8] Percentages and proportions are conceptually equivalent. See Section 2.2 for more information.

[9] These ellipses will be replaced by the name of the main target indicator.

[10] These ellipses will be replaced by the name of the main target indicator.

[11] According to Section 2.3, survey managers should avoid setting requirements for unplanned domains, especially for domains which represent a small share of the total population. Precision requirements can be set for planned domains, e.g. NUTS2, where the sampling design provides for stratification by NUTS2.

[12] Breakdowns can also be defined according to their contributions to the target indicators.

[13] The period of time concerned by the change will be mentioned here. For example, the ellipses can be replaced by 'two successive quarters'.

[14] These ellipses will be replaced by the name of the main target indicator.

'The survey should be designed such that *the estimate of the standard error* does not exceed ... *percentage points* for the change between ...[15] of the estimated percentage ...[16] for the …[17] population breakdowns, where such population breakdowns make up at least ...*%*[18] of the total reference population'.

If the confidence interval of the net change includes the value 0, then the change in the estimates is not significantly different from 0 at the corresponding significance level.

- *The requirements for the estimates of level and of net change should be accompanied by additional* ***provisions for the relaxation and/or exemption of requirements for small and very small geographical domains (breakdowns) (e.g. countries, NUTS2 or NUTS3 regions).*** These provisions are particularly relevant to estimates of national breakdowns, where there are only few population units in small countries' breakdowns, thus requiring a higher sampling fraction than for bigger countries. The provisions address a political concern concerning the burden on small countries/regions. The following provisions can be used:

  'The same requirement is relaxed to a threshold of ... *percentage points* for geographical domains with a population of between ... and ... inhabitants'.

  'Geographical domains whose population is below ... inhabitants are exempted from these precision requirements concerning changes'.

This proposal on formulating a common standard of national precision requirements is accompanied by the following explanations and clarifications:

- The type of estimate to which the standard formulation refers to is the estimated percentage (conceptually equivalent to the estimated proportion).
- The concept of standard error is closely related to survey design since it reflects the expected variability of the parameter estimator (the parameter in this case is the population percentage). Typically, the standard error remains an unknown value which itself has to be estimated, by using an appropriate estimator (called the 'estimator of the standard error'). *Consequently, 'standard error' should be used in conjunction with 'estimator'.* For determining a particular sample and a particular estimated percentage, we can calculate an estimate of the standard error by using an appropriate estimator. Hence, as the requirements concern the survey output, the formulation refers to the 'estimate of standard error' rather than just to 'standard error'. *'Estimate of the standard error' should be used in conjunction with 'estimated percentage'.*
- The measurement unit of a percentage is percentage points. Both standard error and absolute margin of error conserve the measurement unit of the target indicator. An estimate of standard error is therefore expressed in percentage points in the formulation of requirements.

[15] The period of time concerned by the change will be mentioned here. For example, the ellipses can be replaced by 'two successive quarters'.

[16] These ellipses will be replaced by the name of the main target indicator.

[17] According to Section 2.3, survey managers should avoid setting requirements for unplanned domains, especially for domains which represent a small share of the total population. Precision requirements can be set for planned domains, e.g. NUTS2, where the sampling design provides for stratification by NUTS2.

[18] Breakdowns can be also defined according to their contributions to the target indicators.

Let us consider a net sample size of 8 000 units (individuals). Assuming simple random sampling, if the estimated percentage of individuals with access to the Internet is 50 % (50 percentage points), then the estimate of the standard error is around 0.56 percentage points. The half-length of the confidence interval (the estimate of the absolute margin of error) is around 1.1 percentage points, for a confidence level of 95 %. The upper and lower limits of the confidence interval are determined by adding and subtracting 1.1 percentage points to and from 50 percentage points. Thus, the lower limit of the confidence interval is 48.9 % (48.9 percentage points) and the upper limit is 51.1 % (51.1 percentage points). Confusion may arise if the percentage sign '%' is used instead of 'percentage points' to express the estimate of standard error or of the absolute margin of error. The risk is that the upper and lower limits of the confidence interval are determined after calculating the percentage of 1.1 % out of 50 percentage points and then adding and subtracting the result to and from 50 percentage points. For this reason, the threshold for the estimate of standard errors is expressed in 'percentage points' (and not in '%') in the formulation of requirements.

- The precision requirements concern the survey output (the actual estimates), while measures have to be taken at the design stage to ensure compliance with the requirements. This is the rationale for the expression 'the survey should be designed such that...' in the formulation. *At the design stage, survey designers should take into consideration the expected non-response, the variability of the variable of interest in the population, the design effect, etc. in order to estimate the sample size needed.* Meeting survey output requirements by adopting measures at the survey design stage is not an easy task. This is because of the variability of the variance estimates across all possible sample realisations and the fact that the variance estimate of the point estimate obtained with one sample is subject to this variability. It is one of the reasons why compliance with requirements is accompanied by provisions on tolerance (see chapter 5 for more details).
- Unlike the margin of error, the use of the standard error in precision requirements does not assume a normal distribution of the sample means across all possible sample realisations.[19]
- The requirements cover only precision and not the bias, so the whole concept of accuracy is not fully covered. *The precision should incorporate the effects, e.g. of non-response, calibration, etc.* However, the elimination of bias cannot be guaranteed. It is common practice to set up a control mechanism for the level of non-response:
    - In EU-SILC, under Commission Regulation No 1982/2003, the precision requirements for publication of the data must be expressed in terms of the number of sample observations on which the statistic is based and on the level of item non-response (besides the total non-response at unit level).
        - The Commission shall not publish an estimate if it is based on fewer than 20 sample observations, or if non-response for the item concerned exceeds 50 %.
        - The Commission shall publish the data with a flag if the estimate is based on 20 to 49 sample observations, or if non-response for the item concerned exceeds 20 % and is 50 % or below.

[19] However, bootstrap confidence intervals, for instance, are not based on the normality assumption.

- The Commission shall publish the data in the normal way when they are based on 50 or more sample observations and the item non-response does not exceed 20 %.

All data publications must include technical information for each Member State on the effective sample size and a general indication of standard errors for at least the main estimates.

- The OECD Programme for International Assessment of Adult Competencies (PIAAC) outlines a minimum overall response rate of 70 % as the goal, and goes on to state that (OECD, 2010):
  - data from all countries with a minimum response rate of 70 % will generally be included in international indicators and reports unless sample monitoring activities and/or non-response bias analyses indicate serious levels of bias in the country data;
  - results from countries with response rates of between 50 % and 70 % will typically be included in international indicators and reports unless problems resulting from such response rates are compounded by other factors, such as under-coverage bias;
  - results from countries with response rates below 50 % will not be published unless the country can provide evidence that the potential bias introduced by the low response rates is unlikely to be greater than the bias associated with response rates of between 50 % and 70 %.

- The proposed standard formulations have some limits caused by the dependence of the estimated standard error on the actual value of the percentage (estimated percentage). For dynamic phenomena in particular, the change in the value of the indicator may trigger a need to readjust the sample size to ensure continued compliance with requirements. However, continuously adapting the sample size is neither feasible nor desirable. *The following possibilities should therefore be envisaged:*
  - *The survey designers may consider the most demanding value possible of the estimated percentage when they estimate the sample size needed.* If the requirements set thresholds for the estimate of the standard error, then theoretically this value is 50 %; in practice, however, it can be the nearest percentage value to 50 % out of the actual range of relevant values for the specific survey.[20] *The feasibility of such reflections should be assessed by the domain specialists for each survey.*
  - *Both requirements for the estimates of level and net change may use multiple thresholds for the estimates of standard error, which should be set up in function of the values of the estimated percentages.* The rationale behind this is to alleviate the different treatment (burden) imposed on countries with different values of estimated percentages. The thresholds may be determined as:
    - the standard errors that correspond to the upper boundaries of each band defined by the values of estimated percentages;
    - the standard errors that correspond to the mid-points of each band defined by the values of estimated percentages.

[20] If the requirements set thresholds for the estimate of the relative standard error (coefficient of variation), then the most demanding percentage in terms of sample size is 0 %, or percentages that tend to 0 %.

- *The threshold for the estimate of the standard error may be expressed as a model function of the estimated percentage for the requirements of the estimates of both level and net change.*

  This is in fact the approach for revising the current precision requirements (Eurostat, 2010d) preferred by the Group of Experts on Precision Requirements for the Labour Force Survey (LFS). The main principle that guided the choice of this approach is that the new precision requirements for the LFS should be neither more restrictive nor more relaxed than the current ones. And in practice, it imposes a neutrality constraint in the revised text.

  The advantages of this approach are:

  - the required precision of design would be fixed, while allowing the threshold to vary with the actual value of the estimate;
  - compliance with requirements will not be influenced by the actual value of the estimate. On the one hand, this approach avoids having to tighten up the requirements just because of a change in the value of the estimate; it also avoids any pressure to increase the sample size, even with an efficient sampling design. On the other hand, it avoids relaxing the requirements and having to deal with a possible budgetary request to cut the sample size, thereby reducing the quality of the survey estimates as a whole.

- The requirement for the precision of the net change of estimates may be adapted by establishing the required difference at which an estimate of change has to be significant.[21]

  - ***For overall national estimates:***

  'The survey should be designed such that a difference of ...[22] or more *percentage points* in the estimated percentage ...[23] between ...[24] is significant at the 0.05 level, for the total reference population'.

  - ***For estimates of national breakdowns (domains)***, the formulation can be adapted analogously.

  Determination of the (maximum) estimated variance (of the estimator of net change) which allows us to conclude that the actual change is significant can be done by using the following statistical test:

  $$H_0: P_2 - P_1 = 0$$

  $$H_1: P_2 - P_1 \neq 0 .$$

  We reject the null hypothesis if:

  $$\left| \frac{\hat{P}_2 - \hat{P}_1}{\sqrt{\hat{V}(\hat{P}_2 - \hat{P}_1)}} \right| > z_{1-\frac{\alpha}{2}} , \quad (2.5.1)$$

  where $\hat{P}_1$ = estimated percentage of time t_1,

[21] This is, for instance, the approach of the U.S. Current Population Survey. See U.S. Census Bureau (2006), p.3-1.

[22] The specific value for the absolute change required to be significant will replace the ellipses.

[23] These ellipses will be replaced by the name of the main target indicator.

[24] The period of time concerned by the change will be mentioned here. For example, the ellipses can be replaced by 'two successive quarters'.

$\hat{P}_2$ = estimated percentage of time t_2,

$\hat{V}(\hat{P}_2 - \hat{P}_1)$ = estimate of the variance of the estimator of the net change,

$z_{1-\frac{\alpha}{2}}$ = the $1-\frac{\alpha}{2}$ quantile of the standard normal distribution.

From the above formula, the null hypothesis will be rejected when the net change of estimates is higher than its estimated absolute margin of error. In other words, rejection of the null hypothesis occurs when the confidence interval of the change does not include the value 0.

Let us take a numerical example. Suppose the net change of estimates between t_1 and t_2 is equal to 10 percentage points. If the estimated absolute margin of error is 5 percentage points (for α = 5%), then the confidence interval ranges from 5 to 15 percentage points. As the confidence interval does not include the value 0, the change is statistically significant. For the same example, the absolute net change is higher than its estimated absolute margin of error, in which case the null hypothesis is rejected.

This approach explicitly requires significance of net change. However, significance can also be a consequence of applying the requirements and not just the basis of the requirement itself.

Precision requirements for gross changes[25] can also be established. *However, all requirements for a survey (for estimates of level, of net changes and of gross changes) should be parsimonious and should be assessed from the point of view of redundancy and consistency.*

The specific choice for formulating requirements should be analysed and decided by the specialists for each survey. Precision thresholds should be agreed by the specialists of the statistical domain, based on technical feasibility studies.

Summary

It is recommended to follow a standard formulation of precision requirements for EU regulations which aims at uniform and unambiguous understanding within the ESS. This formulation (which is presented in this section) is issued for indicators of the proportion type, for both:

- estimates of level (e.g. annual, quarterly, etc. estimated proportions), for the overall national estimates and estimates of national breakdowns (domains);
- net changes of estimates of level (absolute changes of the estimated proportions between successive years, quarters, etc.), for the overall national estimates and estimates of national breakdowns.

Both requirements should be accompanied by additional provisions for relaxing and/or exempting requirements for small and very small geographical breakdowns.

The proposed standard formulations have some limits caused by the dependence of the estimated standard error on the actual value of the estimated proportion. *The following possibilities should therefore be envisaged:*

[25] See Section 3.7.3 for the definition of gross changes.

- *the survey designers may consider the most demanding value possible of the estimated percentage when they estimate the sample size needed;*
- *the requirements may use multiple thresholds for the estimate of standard error, to be set as a function of the values of the estimated percentages;*
- *the threshold for the estimate of the standard error may be expressed as a model function of the estimated percentage.*

Precision requirements for gross changes can also be established. *All requirements for a survey (for estimates of level, of net changes and of gross changes) should be parsimonious and should be assessed from the point of view of redundancy and consistency.*

The specific choice for the formulation of requirements should be analysed and decided by the specialists for each survey. Precision thresholds should be agreed by the specialists of the statistical domain, based on technical feasibility studies.

3. Best practices on variance estimation

This chapter starts with information on the main sampling designs used in the ESS household surveys. This is followed by reviews of all variability sources of estimates which should be taken into account, as far as possible, when estimating variance.

The chapter then reviews the variance estimation methods, their characteristics and applicability, and evaluates the methods against defined criteria. Later on, it introduces a matrix (*Appendix 7.4*) which offers guidance on the choice of suitable variance estimation methods in relation to sampling design and type of indicators, also listing unsuitable methods. There is also guidance on available methods for incorporating the effect on variance of indirect sampling, implicit stratification, unequal probability sampling, calibration, unit non-response, imputation, coverage errors and measurement/ processing/ substitution errors. Then the chapter presents tools for variance estimation. Some examples of current methods and tools used in NSIs and Eurostat are briefly described.

The last part of the chapter discusses surveys using sampling over time and guides the reader towards estimation of variance for an annual average of estimates and for estimators of net change that present a covariance structure induced by a rotation pattern (e.g. as in the LFS). The final part introduces the concept of gross changes and sets out basic ideas for variance estimation of gross changes.

3.1 Overview of sampling designs in household surveys

Denisa Camelia Florescu (Eurostat) and Onno Hoffmeister (FAO)

A draft inventory of sampling designs used in the EU Labour Force Survey (EU-LFS), the Information and Communication Technology (ICT) household survey, and EU Statistics on Income and Living Conditions (EU-SILC) highlights their diversity and complexity.

The sampling designs used are:

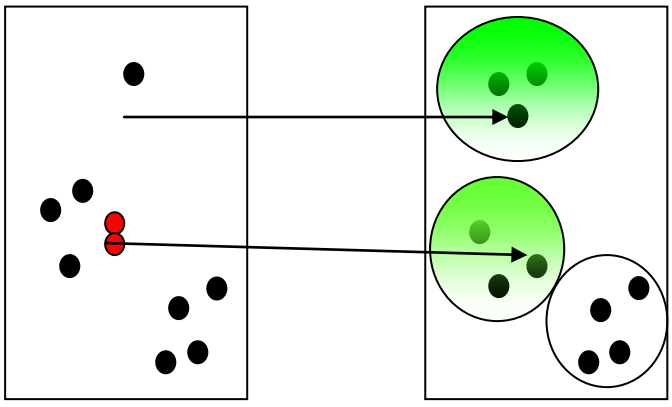
- **simple random sampling** of individuals;
- **stratified random sampling** of individuals;
- **(stratified) cluster sampling** of households, addresses, etc. Clusters are selected in the first stage and *all* eligible individuals in the clusters are interviewed;
- **(stratified) indirect cluster sampling** of households, addresses, etc. *Some* individuals are selected in the first stage and then all eligible individuals living in the household or at the address of the selected individuals are interviewed;
- **(stratified) multi-stage sampling** of individuals. Clusters are selected in the first stage(s) and *some* individuals are selected from clusters in the last stage;
- **(stratified) multi-phase sampling** of individuals. A master sample, a microcensus or a sample drawn from another survey sample is used as sampling frame. *Some* individuals are selected from that frame or from clusters formed in later phases;

- **(stratified) multi-stage cluster sampling** of households, addresses, etc. Clusters are selected in a higher stage than the first and *all* eligible individuals in the clusters formed in the last stage are interviewed;
- **(stratified) indirect multi-stage cluster sampling** of households, addresses, etc. *Some* individuals are selected in a stage higher than the first one and all eligible individuals living in the household or at the address of these selected individuals are included;
- **(stratified) multi-phase cluster sampling** of households, addresses, etc. A master sample, a microcensus or a sample drawn from another survey sample is used as a sampling frame. Clusters are selected from that frame or from clusters formed in later phases and *all* eligible individuals in the clusters selected in the last sampling phase are interviewed;
- **(stratified) indirect multi-phase cluster sampling** of households, addresses, etc. A master sample, a microcensus or a sample drawn from another survey sample is used as sampling frame. *Some* individuals are selected from that frame or from clusters formed in later phases and all eligible individuals living in the household or at the address of these selected individuals are included.

Systematic sampling (with equal or unequal selection probabilities) is often used as a sampling scheme in the different sampling stages.

The above classification makes a clear distinction between **direct** and **indirect sampling**. For example, in direct cluster sampling, the ultimate sampling units are clusters (e.g. households), while in indirect cluster sampling a sample of clusters is obtained from a sample of other units (e.g. individuals). A sample of individuals may be selected from a population register and then a sample of households is obtained by taking all households that have at least one of their current members in the original sample of individuals.

Figure 3.1.1: Indirect sampling of households through individuals



Selection of individuals

Selection of households

In practice, alternative names are sometimes used for indirect sampling, such as network sampling.

The distinction between direct and indirect cluster sampling is deemed relevant, since different weights should be applied to these designs. When a simple random sample of

households is selected, every household has an equal probability of selection. On the other hand, an indirect selection of households through individuals leads to the selection of households with probabilities proportional to their size (according to the number of household members). Weighting for the selected households is by the generalised weight share method. See Section 3.4 for more information on adjustment of weights and for variance estimation for indirect sampling.

An additional distinction should be made between **multi-stage sampling** *and* **multi-phase sampling.**

Both sampling procedures involve sampling at different stages or phases.

However, multi-stage sampling refers to sampling designs in which the population units are arranged hierarchically and the sample is selected in stages corresponding to the levels of the hierarchy. The sampling units are different for the different stages. On the other hand, in multi-phase sampling the same type of sampling unit (e.g. individuals) is sampled multiple times. In the first phase, a sample of units is selected and every unit is measured on some variable. Then, in a subsequent phase, a subsample of units of the same type is selected only from those units selected in the first phase and not from the entire population.

In multi-stage sampling, sampling units are selected in various stages but only the last sample of units is studied. In multi-phase sampling, the sample of units selected in each phase is studied properly before another sample is drawn from it.

Unlike multi-stage sampling, in multi-phase sampling information may be collected at the subsequent phase at a later time; in this event, information obtained on all sampled units of the previous phase may be used if this appears advantageous.

Multi-phase sampling can be used when we do not have a sampling frame with sufficient auxiliary information to allow for stratification, or when we cannot identify in the sampling frame the population subgroup of interest. The first phase is used to measure the stratification variable on an initial sample of units or to screen out the initial sample of units on the basis of some variable. Then, using only the strata or the part of the sample for which we want additional information, a probability sample of those elements is selected for additional data collection on a second variable. For example, a first phase can screen out a sample of individuals to identify only those who have been a victim of a robbery, while the second phase can ask more detailed information (e.g. whether the individuals reported the robbery to the police) to a sub-sample of the identified victims of a robbery. Multi-phase sampling reduces costs, time and the response burden. Moreover, the information from both phases can then be used to compute a regression or a ratio estimate. For instance, a ratio can be the share of individuals who reported a robbery to the police in the total number of individuals who have been a victim of a robbery.

Multi-stage sampling is a particular case of multi-phase sampling arising by imposing the requirements for invariance and independence of the second phase designs. Invariance means that every time the i^{th} PSU (primary sampling unit) is included in the first stage sampling, the same subsampling design must be used. Independence means that subsampling in a given PSU is independent of subsampling in any other PSU. See Särndal *et al* (1992), Section 4.3.1.

Two-phase sampling is sometimes called ‘double sampling’.

Multi-phase sampling can be identified when a survey sample is drawn from a master sample, a microcensus or from another survey sample. *When calculating weights, it is recommended that selection probabilities for those first-phase units be taken into account and that the*

additional variability resulting from multi-phase sampling be accurately incorporated in the calculation of variance estimates.

An additional distinction should be made between **interviewing all, some or one of the eligible members of the selected households.** *Variance estimation should take this into account.*

Most surveys employ multi-stage designs, whereby a sample of households is drawn using any conventional sampling design (simple random sampling without replacement, systematic sampling, stratified sampling, multi-stage sampling and so on) and then individuals are selected for interview from every sampled household. There are two main options at this stage. The first option consists of selecting and interviewing one person per sampled household. The respondent is generally selected by the next/last birthday method or the Kish grid method. An alternative is to survey some or all of the household members above a certain age limit.

Osier (2011) discusses how many people should be interviewed per household in the EU Safety Survey (EU-SASU), particularly whether one or all members should be interviewed in every sampled household. The paper starts with a review of some technical aspects in relation to selecting and interviewing all members of a household, rather than one. The choice has implications for data quality, mainly sampling variance, non-response rate and measurement errors, and for the overall cost of the survey.

The advantages of interviewing all household members are:

- The survey costs are reduced: in order to achieve a target sample size, this option means contacting far fewer households than if one person was interviewed per household. For face-to-face surveys, the number of trips to a segment area can be minimised, which helps save money by reducing travel costs. Nevertheless, with telephone or web surveys the cost of contacting a household is generally small.
- Household respondents may help interviewers by providing contact information for the other household members and the times when they are likely to be available. Further, if their experience was positive, household respondents help to locate and motivate other household members to respond, a burden which would otherwise fall on interviewers. Thus, because the fieldwork can be supervised more easily, non-response is likely to be reduced.
- Having all the members of a household interviewed may also produce more accurate results, especially to household-level questions.

The disadvantages of interviewing all household members are:

- It often leads to less accurate results in terms of sampling variance, mainly because the members of a household tend to be more homogeneous than the general population with regard to the variable of interest.
- Data for multi-respondent households may be subject to certain biases on sensitive topics (e.g. domestic violence, personal attitudes). This measurement bias could be reduced if only one person per household were interviewed.

To compare the effect of selecting one, some or all persons from each sampled household, Osier (2011) uses variance estimation formulae (which assume a simple random sampling of households, a constant number of household members and a constant overall sample of individuals) under different scenarios related to different values of victimisation rates and

intra-cluster correlation coefficients. The design effect is also considered. See Osier (2011) for more details, including consideration of cost.

Finally, some NSIs draw household samples using **balanced designs**. A sampling design is said to be balanced if it ensures that the Horvitz-Thompson estimators of some 'balancing' variables are equal to the known totals. The cube method proposed by Deville and Tillé (2004) enables balanced samples to be selected. As there is often no such thing as an exact balanced sampling design, the cube method generally proceeds in two steps: a 'flight phase' in which exact balance is maintained, and a 'landing phase' in which the final sample is selected while complying as closely as possible with the balance conditions. Deville and Tillé (2005) derive a variance approximation for balanced sampling. It stems from considering balanced sampling as a calibration exercise at the design stage and, like calibration (see Section 3.4), it relies on the residuals of regression of the study variable on the balancing variables.

Summary

Sampling designs used in household surveys are highly diverse and complex.

Variance estimation should take into account sampling design. It should distinguish between direct and indirect sampling, between multi-stage and multi-phase sampling and, in the case of household surveys, between enumerating only one or more members of the same household.

3.2 Sources of variability of an estimator

Mārtiņš Liberts (CSB)

This section reviews the main sources of variability of an estimator and gives general guidelines on how to incorporate such variability components into variance estimation.

We have a parameter of interest denoted by θ which we would like to estimate from a sample $s \in S_0$ (S_0 denotes the set of all possible samples that can be drawn from the target population). We have an estimator denoted by $\hat{\theta}$ which we are using to estimate θ. By definition, $\hat{\theta}$ is a stochastic variable. It is a function of the set-valued random variable $\tilde{S}$ whose realisations are the possible samples (see *Appendix 7.1*).

$$\hat{\theta} = \hat{\theta}(\tilde{S}) : s \in S_0 \rightarrow \hat{\theta}(s) = \hat{\theta}_s \ . \quad (3.2.1)$$

In practice, the expected value of $\hat{\theta}$ denoted by $E(\hat{\theta})$ is often not equal to θ (because of non-response errors, measurement errors, coverage errors and other errors). The difference between $E(\hat{\theta})$ and θ is the **bias** of $\hat{\theta}$:

$$B(\hat{\theta}) = E(\hat{\theta}) - \theta = \sum_{s \in S_0} p(s) \cdot \hat{\theta}_s - \theta \ . \quad (3.2.2)$$

Bias is a component of the mean square error. It is one of the accuracy measures of a population parameter estimator. The bias of an estimator is the average error of the estimator

over all possible samples. An estimator is biased if, on average, its value differs from the true value. In this handbook, however, we are not concerned with bias.

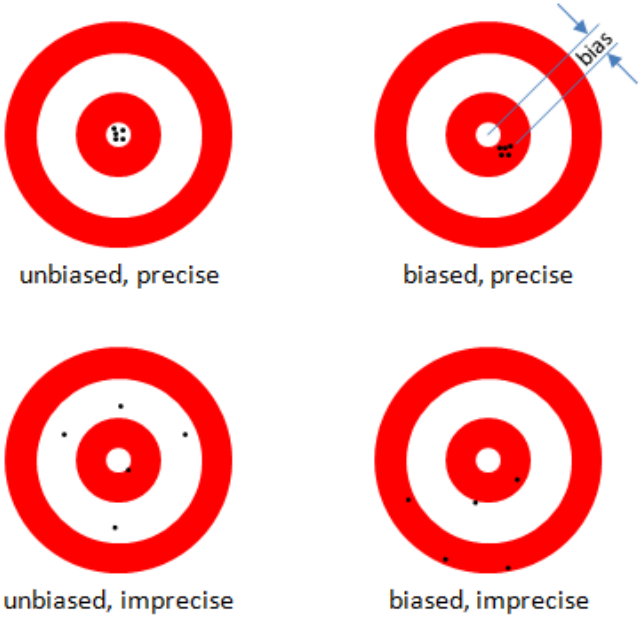
The sample-to-sample **variability** of $\hat{\theta}$ around $E(\hat{\theta})$ is the other component of the mean square error of an estimator. We will express this variability by variance $V(\hat{\theta})$:

$$V(\hat{\theta}) = E\left[\hat{\theta} - E(\hat{\theta})\right]^2 = \sum_{s \in S_0} p(s) \cdot \left[\hat{\theta}_s - E(\hat{\theta})\right]^2 = E(\hat{\theta}^2) - \left[E(\hat{\theta})\right]^2. \qquad (3.2.3)$$

The sampling variability is the variability of the statistic computed for all possible samples taken from a population. **Precision** refers to how close estimates from different samples are to each other.

The concepts of bias and variability are illustrated in the figure below.

Figure 3.2.1: Bias and variability (precision) of an estimator[26]



We need to know the variance $V(\hat{\theta})$ in order to be able to measure the variability of an estimator. It is not possible to measure the true value of $V(\hat{\theta})$ from a sample of the population. But it is possible to build a variance estimator $\hat{V}(\hat{\theta})$ for $V(\hat{\theta})$.

Before building $\hat{V}(\hat{\theta})$ *it is recommended to explore the different sources of variability for an estimator*. There are several sources of variability for an estimator $\hat{\theta}$, namely:

Sampling design and estimator

The first source of variability of an estimator comes from the procedure used in selecting the sample (commonly known as the sampling design). Consider a finite population U of size N. A random sample s of size n is selected from the population according to a sampling

[26] Charles Annis, P. E. Statistical Engineering. http://www.statisticalengineering.com/Weibull/precision-bias.html.

design $p = \{p(s), s \in S_0\}$.[27] The variability of $\hat{\theta}$ depends on sampling design p in the sense that the estimator would take different values from one realisation of the sample to another. The variability caused by observing a sample instead of the whole population is called sampling error.

The form of an estimator also has an impact on variance. For example, calibration estimators are known to be generally more accurate than 'non-calibrated' estimators. Not taking this feature into account in calculations would lead to misleading results, particularly when calibration information is strongly correlated with what the survey intends to measure (Deville and Särndal, 1992). See Section 3.4 for how to account for variability caused by calibration.

There is a great deal of statistical literature (e.g. Särndal *et al*, 1992) that deals with variance estimation under 'ideal' survey conditions. By 'ideal' conditions we mean a full response, a sampling frame that perfectly represents the target population (no frame errors) and absence of measurement, processing and any other non-sampling errors.

Unit non-response

All surveys have to deal with unit non-response, that is, the failure to collect information on a sample unit (due to non-contact, refusal or other reasons). Unit non-response is a source of both bias and variability for $\hat{\theta}$. Variability comes from the fact that we have a subset r of respondents selected as a subset from sample s with the conditional probability $q(r|s)$. The variance of the estimator increases because the size of the subset r of respondents is smaller than the size of sample s.

See Section 3.4 for how to account for the variability caused by unit non-response.

Item non-response

Another practical problem is item non-response, that is, failure to collect information on certain items only. Item non-response is usually handled by imputation. A common source of error in variance estimation is to treat imputed values as exact values. Ignoring imputation and treating imputed values as if they were observed values may lead to valid point estimates (under missing at random scenario) but it could lead to underestimation of variance (if standard variance estimation methods are naively applied).

The impact of imputation on variance can be large if there is considerable item non-response. A non-response rate of 30 % may lead to 10-50 % underestimation of standard error (Kovar and Whitridge, 1995, as cited by Eurostat, 2002).

See Section 3.4 for how to account for the variability caused by item non-response.

Coverage (frame) errors

Frame imperfections such as over-coverage and multiple listings are other potential sources of variability in estimates. Under-coverage is usually a source of bias. Under-coverage occurs when target population units are not accessible via the sampling frame. Sometimes the sampling frame is incomplete, some units are omitted and potential respondents cannot be sampled with a view to participating in the survey.

- Over-coverage generally increases variance because it results in the number of sampled eligible units being lower than the sample size (non-eligible units which do not belong

[27] See Appendix 7.1.

to the target population do not provide information on the study variables of a survey). The random number of sampled eligible units introduces a further component of variance.

- Multiple listings can increase variance. One of the main reasons is that multiple listings, like over-coverage, reduce the size of the final effective sample (population elements that appear in the sample more than once are excluded from it).

See Section 3.4 for how to account for variability caused by coverage errors.

Measurement errors

Measurement errors introduce another component of variance — response variance (Wolter, 2007). Measurement errors arise from the fact that observed values which are collected through a fixed survey procedure may differ from true values. There are many explanations for this e.g. respondents might not understand a question, or be unwilling to provide true answers to certain questions. Interviewers might also influence respondents to give erroneous answers. For further details regarding variance estimation under measurement errors, see Section 3.4.

Processing errors

Processing errors are of the same nature as measurement errors. Possible sources of processing errors are data entry, data editing (checks and corrections) or coding. Just as with measurement errors, we can assume processing errors to be a random variable with some unknown characteristics. Processing errors can introduce an extra component of variability in $\hat{\theta}$. For further details, see Section 3.4.

Substitution errors

Substitution errors are likewise of the same nature as measurement errors. Substitution errors are caused by substituting a unit in a sample by another unit considered as a good proxy for the original one. The value collected on a substituted unit generally differs from what would have been collected on the original unit, which makes substitution errors equivalent to measurement errors. For further details, see Section 3.4.

Substitution is, in practically all cases, bad practice and should be avoided because there is a high risk that the process of identifying the units to be substituted is informative.

The main sources of variability of an estimator are summarised in Table 3.2.1:

Table 3.2.1: The main sources of variability of an estimator

Source	Component of variance	Estimation methods (see Section 3.4 for a description of methods)
Sampling design and estimator	**Sampling variance**	Estimator based on sampling design, type of parameter of interest θ and estimator $\hat{\theta}$. Methods which account for the effect of implicit stratification, rotating samples, indirect sampling and unequal probability sampling on variance are presented in this handbook. Among methods which can account for the effect of calibration on variance, the main one presented is the Deville and Särndal method (1992).
Unit non-response	**Non-response variance**	Adjusting the variance estimator $\hat{V}(\hat{\theta})$ to take unit non-response into account can be done by using methods which assume that respondents are missing at random or completely at random within e.g. strata or constructed response homogeneity groups or by using the two-phase approach or Fay's approach (Fay, 1991; Shao and Steel, 1999).
Item non-response	**Imputation variance**	Multiple imputation can be used to account for the imputation variance. Replication methods and analytical methods can also be used to incorporate imputation into variance estimation. Deville and Särndal (1994) proposed a method for the regression imputed Horvitz-Thompson estimator.
Over-coverage	**Over-coverage variance**	Methodology of domain estimation can be used. Target population has to be defined as a domain of the frame population. The related loss of precision can be quantified.
Multiple listings	**Multiple listings variance**	Same as over-coverage. Possible to estimate if correct sampling probabilities can be computed.
Measurement errors	**Measurement (response) variance**	Require several sets of repeated measurements. Each new set must be conducted under the same conditions as the others and, above all, must be uncorrelated with the other sets (very hard to achieve in practice).
Processing errors	**Processing variance**	Same as measurement errors.
Substitution errors	**Substitution variance**	Same as measurement errors.

In surveys based on samples, the total variance comprises sampling variance and non-sampling variance, while in censuses and take-all strata, the total variance consists only of non-sampling variance.

The total variance of an estimator $\hat{\theta}$ can be split into many components, as summarised in the previous table. In theory, we should estimate each component individually. However, certain components of variance may be difficult to estimate because the necessary information for variance estimation is missing. This is what happens, for instance, with the response variance caused by measurement errors, as we never have repeated measurements of the same variable at our disposal. The fact is that it is hard to estimate variance components separately, even though the users of statistics are usually interested in the overall variance of $\hat{\theta}$ by taking all the different sources of variability into account. It therefore makes sense to devote most of our efforts to determining the main components of variance and ignore the others. This approach almost certainly leads to the overall variance being underestimated, but it is still probably the most reasonable solution one can hope for under such circumstances.

Variance estimators can be tested by simulation experiments. It might be possible to simulate artificial extra non-response (unit or item), measurement and processing errors. Experiments could be done to examine how the variance of population parameter estimates is affected by increasing the level of artificial errors. These could lead us to draw some conclusions about the relationship between the variance of population parameter estimates and the amount of errors.

The overall recommendations for constructing suitable variance estimators are to:

- *consider all possible sources of variability (see Table 3.2.1), or at least those sources which account for most of the total variance;*
- *consider those sources of variability which can be estimated;*
- *consider those sources of variability which can be described with some other indicative information (for example, level of processing errors).*

Summary

The total variance of population parameter estimates is made up of several components: sampling variance, non-response variance, imputation variance, over-coverage variance, multiple listings variance, measurement (response) variance, processing variance and substitution variance.

The recommendation is to make an a priori impact assessment of the different sources of variability and to choose methods that allow the most important sources of variability to be accounted for as much as possible.

3.3 Variance estimation methods: description, evaluation criteria and recommendations

Yves Berger (University of Southampton), Alexander Kowarik (Statistics Austria), Mārtiņš Liberts (CSB) and Ralf Münnich (University of Trier)

This section describes the variance estimation methods. It also makes a general comparative assessment of the methods on criteria related to applicability (the sampling design used and the type of statistics), accuracy (confidence interval coverage probabilities, unbiasedness and stability) and administrative considerations (cost, timeliness and simplicity). It introduces *Appendix 7.4*, which offers guidance on which variance estimation methods to choose in relation to sampling design and type of indicators, and which sets out unsuitable methods (bad practices).

There are various variance estimation methods. There are basically three groups that we consider here: analytical methods, replication methods and methods based on generalised variance functions. In addition, we consider linearisation methods. These are used to find a linear approximation of a non-linear estimator, after which a variance estimation method is applied. Valliant, Dorfman and Royall (2000) synthesise much of the model-based literature on variance estimation. However, model-based estimation of variance is not within the scope of this handbook.

Description of variance estimation methods

- **Analytical methods**

Analytical methods provide direct variance estimators which seek to reflect the main features of the sampling design (stratification, clustering, unequal probabilities of selection, etc.). Analytical methods can be:

  - **Exact analytical methods**

    By nature, exact analytical methods lead to tailor-made variance formulae which reflect the main sampling design components As a result, the variance estimators are generally design-unbiased, or nearly design-unbiased, which makes them attractive. The first rule is that, *whenever possible, we should strive to establish variance formulae which adhere as strongly as possible to the sampling design.*

  - **Approximate analytical methods**

    When a sampling design is too complex, we will not be able to fix any exact variance estimator unless we make additional assumptions and/or use approximate methods. Assumptions always provide an approximate picture of reality, and so will any variance estimators thus obtained.

    Approximate analytical methods rely therefore on assumptions; they are born of the barriers that prevent the large-scale implementation of exact analytical methods:

    - Firstly, sampling designs might happen to be so ‘complex’ that we cannot get a variance estimator unless we make further assumptions. These assumptions consist of approximating the sampling design. Hence, this leads to simplified biased variance estimators.

- A second barrier to the extensive use of exact analytical methods is 'mathematical' difficulties. For example, the exact calculation of probabilities of selection of the order two — that is, the probability that two distinct units *i* and *j* be selected in the sample — is not feasible for certain sampling designs, especially where units are selected with unequal probabilities. To overcome this, the approximate variance formulae (3.4.21), (3.4.22) and (3.4.23) are proposed, wherein double inclusion probabilities are not used.
- Unit non-response requires further assumptions to be made regarding the nature of the response mechanism: a Poisson selection, a post-stratified selection, a simple random selection, etc. The accuracy of any analytical variance estimator that would take into account both sampling design and non-response mechanism is closely linked to the validity of the non-response model.

  The same is true of measurement errors, since the only way to handle them analytically is to make model-based assumptions. If the model turns out to be true, then the variance expression should be unbiased; otherwise, it might result in a high level of bias.
- *In the case of multi-stage sampling, variance estimation at the first stage only is insufficient and not recommended when variances at subsequent stages are comparatively large.* It can be extremely cumbersome to estimate the variance at each stage, especially if samples are drawn with unequal probabilities at several stages within the sampling design. The problem of calculating selection probabilities of the order two is given at each stage (Särndal *et al* (1992)). Münnich *et al* (2011a) also show some variance estimation methods for complex sampling designs, like multi-stage sampling or unequal probability sampling.

  For a two-stage sampling design, according to Raj (1968), an unbiased estimator of the variance of the Horvitz-Thompson estimator of a total *Y* is given by

$$\hat{V}\left(\hat{Y}\right) = f\left(\hat{T}\right) + \sum_{i=1}^{n} w_i \hat{V}_i \,, \quad (3.3.1)$$

  where $f\left(\hat{T}\right)$ is the estimated variance contributed by the first stage, $\hat{T}$ is the vector of the estimated totals of the PSUs, $\hat{V}_i$ is an unbiased estimator of the second-stage variance of the estimated total of PSU *i*, and w_i is the first-stage sampling weight of PSU *i*. The formula applies only to invariant and independent second-stage designs, i.e. to cases where the design that is used for sampling within PSU *i* does not depend on which other PSUs have been selected in the first-stage sample, and where the subsampling in PSU *i* is independent of subsampling in any other PSU.

  Rao (1975) extended the above method to the case where the second-stage design is not invariant:

$$\hat{V}\left(\hat{Y}\right) = f\left(\hat{T}\right) + \sum_{i=1}^{n} \left(w_i^2 - b_{is}\right) \hat{V}_i \,, \quad (3.3.2)$$

  where b_{is} are coefficients that depend on the sample s of selected PSUs, on the sampling design and on the first and second-order inclusion probabilities of PSUs. The weights w_i might also depend on the selected sample s.

The latter formula shrinks to a single term when (1) the estimator of the total is Horvitz-Thompson, (2) the variance estimator is unbiased and (3) the first-stage inclusion probabilities are small. *This simplification should not be made without checking whether the ignored term is indeed negligible.*

Actually, in a multi-stage design, if the ratio of selected clusters at the first stage to the total number of clusters in the population is small, then stages other than the first add little to the standard error and so the variance estimation methods may take account only of the first stage of sampling design. For a mathematical elaboration, see Kish (1965) and Cochran (1977). The variance estimation process gets easier if clusters are selected with probabilities proportional to their size and when we assume (without important errors) that they have been selected with replacement (Nikolaidis, 2008).

- The selection in the sample of one PSU per stratum makes it impossible to estimate the variance of any statistic. This can happen with very small strata, or with strata suffering from severe non-response problems.

  In these cases, an unbiased estimator of the variance is not available, not even for linear statistics. It is possible to use an estimator that tends towards an overestimate of the variance. This is the collapsed stratum estimator, which is applicable only to problems of estimating the variance of linear statistics.

  One solution is to collapse strata so as to have at least two PSUs in each stratum. So in order to estimate the variance of estimator $\hat{Y}$, we combine the original H strata into G groups of at least two strata each. Assume that H is even and that each group contains precisely two of the original strata. The estimator of the population total can be expressed as $\hat{Y} = \sum_{g=1}^{G} \hat{Y}_g = \sum_{g=1}^{G} \left(\hat{Y}_{g1} + \hat{Y}_{g2} \right)$, where $\hat{Y}_{gh} (h = 1, 2)$ denotes the estimator of the total of stratum h in the group g. If we consider $g1$ and $g2$ as independent selections from group g, then the estimator of the variance of $\hat{Y}$ is $\hat{V}_{cs}\left(\hat{Y}\right) = \sum_{g=1}^{G} \left(\hat{Y}_{g1} - \hat{Y}_{g2} \right)^2$. The estimator overestimates the true variance by $B\left(\hat{V}_{cs}\left(\hat{Y}\right)\right) = \sum_{g=1}^{G} \left(\mu_{g1} - \mu_{g2} \right)^2$, where $\mu_{gh} = E\left(\hat{Y}_{gh}\right)$. For more details, see Cochran (1977), Kish (1965), etc. and for the implications on variance estimation, see Wolter (2007) (Section 2.5).

  For non-linear estimators, the variance can be estimated by a combination of collapsed stratum and Taylor series methodology. See Wolter (2007) (chapter 6).

- Another difficulty arises with the calculation of confidence intervals, for which it is generally assumed that the statistic follows a normal distribution. However, when dealing with non-linear statistics, this assumption can be questioned. Bootstrap or empirical likelihood (Hartley and Rao, 1962; Owen, 2001) methods are appropriate for this type of situation.

  Analytical (exact and approximate) methods can be used for linear statistics and simple non-linear statistics (ratios, ratios of two ratios) under most of the commonly used sampling designs (see *Appendix 7.4*). Exact methods can be used under simple random sampling, stratified random sampling and cluster sampling, while approximate methods can be used for multi-stage sampling designs.

- **Linearisation methods**

These are used to obtain a variance estimator in the event of a complex form of the parameter.

Variance estimation based on linearisation methods consists of finding a linear approximation to a non-linear statistic (such as ratio estimate, regression coefficient, correlation coefficient), working out a variance estimator for the linear approximation and finally using it as an estimator for the variance of the non-linear statistic.

- **Taylor linearisation (TS)**

  The Taylor linearisation method (Tepping, 1968; Woodruff, 1971; Wolter, 2007) is a well-established method of obtaining variance estimators for non-linear statistics (defined either in an explicit or an implicit way) which are smooth (differentiable). It consists of approximating a non-linear statistic with a linear function of observations by using first-order Taylor series expansions.

  Taylor series expansion requires the assumption that all higher-order terms are of negligible size. Some underestimation of variance is to be expected, at least for moderate-sized samples, because higher-order terms are neglected. If all higher-order terms are of negligible size, then the variance approximation works well and can be used; otherwise, serious biases in the estimates may result. Dippo and Wolter (1984) apply second-order Taylor series approximations to several estimates and show that this reduces the bias, but increases the variance of the variance estimate and complicates computations considerably. Underestimation of the variance from Taylor series may compensate for some of the overestimation that results when variance is computed assuming sampling with replacement for a without-replacement sampling design.

  For linearisation, the function of interest may not be expressible as a smooth function of population totals or means. In order to accommodate statistics with a more complex mathematical expression, generalised linearisation frameworks have been developed:

- **Linearisation based on estimating equations** (Binder, 1983; Kovacevic and Binder, 1997)

  In this technique, population parameters are expressed as solutions to appropriate population estimating equations. Sample estimates of these parameters are obtained by solving sample estimating equations which involve design weights and, possibly, calibrated weights based on auxiliary information. Variance estimates can then be obtained either by linearisation or by Newton Raphson optimisation.

  This method can, for example, treat quantiles as well as measures of income inequality, such as the Gini coefficient, which the Taylor method cannot. The method is applicable to a wide range of complex sampling designs and is also less computationally intensive than replication alternatives.

- **Linearisation based on influence functions** (Deville, 1999)

  If the parameter of interest is a non-linear function of population totals, the statistic of interest (the estimate of the parameter) is computed as the same non-linear function of the estimated population totals. Each estimated total is a weighted sum of the sample observations of the corresponding variable. The sampling weights can be viewed as the values of a measure defined on the real multidimensional space with as many dimensions as there are variables of the survey. The estimator of the parameter of interest is then a function, say T, of this measure. A linearised variable is defined with

its value on each sample point equal to the value of the influence function (a special form of derivative) of T on the sample point. If the sample size is large and some additional conditions hold (Deville, 1999), then the variance of the statistic of interest is approximated by the variance of the weighted sum of the linearised values of the variable, which is easy to compute. This linearisation covers more non-linear functions than Taylor linearisation without involving more calculations. Influence functions can be used for linearisation of non-smooth statistics for which Taylor series expansions can no longer be used. In fact, the derivation rules for influence functions are similar to the rules for computing the derivative of a function in a standard differential calculus.

Practical applications of linearisation based on influence functions can be found in EU-SILC (Eurostat, 2005; Osier, 2009).

- **Jackknife linearisation**

  The idea of jackknife linearisation is to replace repeated resampling of a statistic with analytic differentiation. The result is a formula that is simple to calculate and which in large samples is a good approximation to the traditional jackknife calculation (Canty and Davison, 1999).

Formulae for ***linearisation***-based variance estimates can be found in the literature for the most common indicators on poverty and income dispersion (Verma and Betti, 2011). Thus, variance estimation based on linearisation should be feasible for nearly all statistics used in practice and under most of the commonly used sampling designs.

A recommended application of ***linearisation*** *methods is when the number of sample members becomes a random variable, for example in the case of unplanned domains or by using clusters with different sizes.* Then the mean becomes a ratio of variables. Moreover the majority of statistics are non-linear, and because there are no exact expressions for calculating their variances, it becomes necessary to use approximations such as Taylor series linearisation and replication methods. *In order to obtain an adequate approximation for the variance of the ratio estimator by means of a Taylor series, the sample size (the denominator of the ratio) should not be subject to great variation, which would however be the case when e.g. clusters have widely different sizes.* The difficulty in keeping the denominator variability under control increases when the estimation is directed at subclasses, since it is then impossible to control the number of units belonging to each subclass.

The ***linearisation*** approach relies on the assumption that the sample-to-sample variation of a non-linear statistic around its expected value is small enough to be considered linear. The latter assumption is particularly correct with large samples. Although this is unlikely to be a problem with national samples comprised of thousands of elements, *we should be cautious when dealing with breakdown estimates, especially when such estimates refer to small domains (Osier, 2009).*

With the ***linearisation*** approach to variance estimation, a separate formula for the linearised estimate must be developed for each type of estimator (requiring additional programming efforts). In this respect, the linearisation approach differs from the ***replication methods*** which do not require derivation of a variance formula for each estimator. This is because in replication methods, the approximation is a function of the sample, and not of the estimate. Replication methods for estimating variance for very complex functions are therefore easier than linearisation methods. On the other hand, some sampling designs may not satisfy restrictions required by replication methods. The linearisation method is applicable to any sampling design. Linearisation and replication approaches do not produce identical estimates

of standard error, but empirical investigations have shown that for many statistics, the differences are negligible (Kish and Frankel, 1974).

- **Replication methods**

The replication approach is based on the originally derived sample (full sample), from which a (usually) large number of smaller samples (sub-samples or replicate samples) is drawn. From each replicate sample the statistic of interest is computed. Replicate estimates are determined using the same estimation method as the original estimate. The variability of these replicate estimates is used to derive the variance of the statistic of interest (of the full sample).

In the case of calibration estimators the replicate weights should be produced in a similar way to full sampling weights. All the weighting adjustment processes performed on the full sampling weights should also be carried out for each replicate weight.

- **Jackknife (JK)**

Jackknife is used in statistical inference to estimate bias and standard error in a statistic, when a random sample of observations is used. The basic idea behind the jackknife estimator lies in systematically re-computing the statistic leaving out one observation (or group of observations such as in delete-a-group jackknife) at a time from the sample set and reweighting the remaining units. Using this new set of 'observations' for the statistic, an estimate for the bias (estimator bias) can be calculated, as well as an estimate for the variance of the statistic.

A general description of the variance estimation method is given below.

Assume a sample s of n elements obtained by a non-stratified sampling design. Let also the population parameter θ be estimated by $\hat{\theta}$, an estimator based on data from the full sample s.

The jackknife technique starts by partitioning the sample into A random groups of equal size. We assume that for any given s each group is a simple random sample (without replacement) from s. Next, for each group ($a = 1, \ldots, A$), we calculate $\hat{\theta}_{(a)}$, an estimator of θ, based only on the data that remain after omitting group a. For $a = 1, \ldots, A$, we then define

$$\hat{\theta}_a = A\hat{\theta} - (A-1)\hat{\theta}_{(a)} \qquad (3.3.3)$$

sometimes called the a^{th} pseudovalue. The jackknife estimator of θ (an alternative to the estimator $\hat{\theta}$) is given by

$$\hat{\theta}_{JK} = \frac{1}{A}\sum_{a=1}^{A}\hat{\theta}_a \qquad (3.3.4)$$

and the jackknife variance estimator is defined as

$$\hat{V}_{JK} = \frac{1}{A(A-1)}\sum_{a=1}^{A}\left(\hat{\theta}_a - \hat{\theta}_{JK}\right)^2. \qquad (3.3.5)$$

In practice $\hat{V}_{JK}$ is used as an estimator of $V(\hat{\theta})$ as well as of $V(\hat{\theta}_{JK})$. Alternatively, the variance of the statistic is estimated by

$$\hat{V} = \frac{1}{A(A-1)} \sum_{a=1}^{A} \left(\hat{\theta}_a - \hat{\theta}\right)^2 . \qquad (3.3.6)$$

Denoting $\hat{\theta}_{(\cdot)}$ the mean of the A values $\hat{\theta}_{(a)}$, the jackknife variance estimator can also be written as

$$\hat{V}_{JK} = \frac{A-1}{A} \sum_{a=1}^{A} \left(\hat{\theta}_{(a)} - \hat{\theta}_{(\cdot)}\right)^2 . \qquad (3.3.7)$$

The jackknife variants are JK1 for non-stratified sampling designs, JK2 for stratified designs with two PSUs per stratum, and JKn for stratified designs with two or more PSUs per stratum. The number of replicates depends on the variant selected, the number of strata and the number of PSUs by strata.

Note that, in stratified designs, jackknife is not an independent process in each of the strata, but is done sequentially for all PSUs in all strata, taken one by one.

Jackknife is easier to apply to many kinds of sampling designs, including complex sampling schemes (Shao and Tu, 1995), such as multi-stage sampling with varying sampling weights, than the ***bootstrap*** method.

Jackknife can accommodate most estimators likely to occur in survey practice. However, ***delete-one or groups jackknife*** variance estimators do not work for complex non-smooth statistics such as poverty measures (except for the Gini coefficient — Berger, 2008) as they lead to inconsistent variance estimators (Miller, 1974; Wolter, 2007). ***Delete-one or groups jackknife*** *should therefore not be used for complex non-smooth statistics (except for the Gini coefficient*). In this case, linearisation based on estimating equations or influence functions can be used instead, followed by e.g. analytical methods. However, the ***delete-a-group jackknife*** might also perform well for non-smooth statistics (e.g. quantiles, rank statistics) as the number of sample elements in clusters increases. Additionally, the deficiency of the delete-one or groups jackknife can be restricted by using a more general jackknife, called the ***delete-d jackknife***, with the number of deleted observations, d, depending on a smoothness measure of statistics. In particular, for sample quantiles, the delete-d jackknife with d satisfying $\frac{\sqrt{n}}{d} \to 0$ (n is the number of PSUs) and $n-d \to \infty$ (as $n \to \infty$) leads to consistent variance estimators in the case of independent and identically distributed observations (Shao and Wu, 1989).

Nor should ***delete-one jackknife*** *be used in stratified sampling* (Wolter, 2007, pp. 172-173).

Existing jackknife variance estimators used with sample surveys can seriously overestimate the true variance under one-stage stratified sampling without replacement with unequal probabilities. In this case, Berger (2007) describes ***a generalised jackknife variance estimator.*** However, it has many disadvantages (i.e. it cannot be implemented in standard statistical packages, it is computationally intensive, the exact joint inclusion probabilities are difficult to calculate). Further, Berger (2007) proposes

a jackknife estimator which does not require exact joint inclusion probabilities and is always positive, unlike the generalised jackknife variance estimator. The jackknife proposed by Berger (2007) is unbiased.

Berger and Skinner (2005) propose a ***delete-one weighted jackknife*** for general unequal probability sampling designs.

Kott (2001) shows that a ***delete-a-group jackknife*** variance estimator underestimates variance when the number of replicates is higher than the number of PSUs and that this is a frequent case in social surveys, where generally only one or two PSUs are selected for each stratum. To deal with this, Kott proposes an ***extended delete-a-group jackknife*** which is implemented in EVER software. ISTAT has carried out simulation studies to investigate the performance of the methods in the presence of imputation in social surveys.

For stratified multi-stage with replacement designs, Rao *et al* (1992) propose a ***customary delete-cluster jackknife***. For stratified multi-stage without replacement designs, this ***customary delete-cluster jackknife*** is biased when there is a low variation between PSUs or high variation within PSUs.

For two-stage self-weighted designs, Escobar & Berger (2010) propose a ***jackknife estimator*** which involves deleting PSUs as well as units. The estimator is design consistent and asymptotically unbiased.

- **Bootstrap**

Bootstrapping is a statistical method for estimating the sampling distribution of an estimator by sampling with replacement from a population reconstructed from the sample using appropriate weights: see Canty and Davison (1999). It is most often used to derive robust estimates of standard errors and confidence intervals of population parameters such as mean, median, proportion, ratio, correlation coefficient or regression coefficient.

The number of replicates, and the number of PSUs sampled in each replicate, can be chosen by the analyst, although there are practical recommendations for both these quantities. Efron and Tibshirani (1986) report that the number of replicates in the range of 50 to 200 is adequate in most situations. The precision of the bootstrap is higher if the number of replicates is increased.

Assume a sample s drawn from a population U by a sampling design without replacement. Let also the population parameter θ be estimated by $\hat{\theta}$. A brief description of the bootstrap technique is as follows:

i. Using the sample data, construct an artificial population U^*, assumed to mimic the real but unknown population U.

ii. Draw A independent 'bootstrap samples' from U^* by a design identical to the one by which s was drawn from U. Independence implies that each bootstrap sample must be replaced into U^* before the next one is drawn. For each bootstrap sample, calculate an estimate $\hat{\theta}_{(a)}$ $(a=1,\ldots,A)$ in the same way as $\hat{\theta}$ was calculated.

iii. The observed distribution of $\hat{\theta}_{(1)},\ldots,\hat{\theta}_{(a)}$ is considered an estimate of the sampling distribution of the estimator $\hat{\theta}$, and $V(\hat{\theta})$ is estimated by

$$\hat{V}_{BS} = \frac{1}{A-1}\sum_{a=1}^{A}\left(\hat{\theta}_{(a)} - \hat{\theta}_{(\cdot)}\right)^2, \qquad (3.3.8)$$ where

$$\hat{\theta}_{(\cdot)} = \frac{1}{A}\sum_{a=1}^{A}\hat{\theta}_{(a)}\,. \qquad (3.3.9)$$

The ***bootstrap*** method is quite flexible as it can handle most estimators likely to occur in survey sampling practice, including nondifferentiable statistics (unlike for instance ***jackknife***) and new, novel statistics (more easily than ***TS*** method) (Wolter, 2007).

However, Wolter (2007) notes that ***bootstrap*** has not been adequately tested for large-scale, complex surveys and cannot give any recommendation on its use in this environment.

Bootstrap can be used for constructing hypothesis tests. It is often used as a robust alternative to inference based on parametric assumptions when those assumptions are in doubt, or where parametric inference is impossible or requires very complicated formulae for calculation of standard errors. The bootstrap method does not depend on any specific properties of the sample statistic and can therefore be used universally in a general computational algorithm. But the ordinary ***Monte-Carlo non parametric bootstrap*** can lead to biased variance estimates when samples are drawn with unequal probabilities or without replacement and large sampling fractions (Davison and Sardy, 2004; Münnich *et al*, 2011a); techniques for addressing this issue are available (see Canty & Davison, 1999).

In planning a sample survey, a pilot sample is typically used to determine the required sample size to achieve a specified level of accuracy for statistical inference. Mak (2004) proposes a practical procedure based on ***bootstrap*** for estimating simultaneously the variances of a statistic for all sample sizes based on a single observed pilot sample. For an observed sample of size n_0, it is well known that bootstrap can be used to estimate numerically the variance of a sample statistic computed from the sample. To study the variances of the statistic for other sample sizes, we can in principle generate bootstrap samples of size n for a range of values of n, and then calculate the bootstrap variance estimate for each n. This, however, will be computationally demanding and inefficient. By contrast, the method proposed requires bootstrap samples to be generated for only two selected values n_1 and n_2 of n. Estimates of the variances of the statistic with small biases can then be computed for any other values of n. It is proved theoretically that these biases decrease rapidly to zero as n_1 and n_2 increase.

- **Balanced repeated replication (BRR) or balanced half-samples (BHS)**

This method is suitable for a stratified sampling design with two sampled elements in each stratum or for cluster designs where each cluster has exactly two final stage units per cluster. The aim is to select a set of samples from the family of 2^H half-samples (where H stands for the number of strata or clusters), compute an estimate for each one and then use them for the variance estimator such that the selection satisfies the 'balance' property.

A general description of the method is given below.

We consider a sample s, where $s = \bigcup_{h=1}^{H} s_h$ and each s_h consists of exactly two elements drawn from stratum h. A half-sample is a set consisting of exactly one of the two elements from each s_h. Therefore, there are 2^H possible half-samples. The basic idea of this method is to select a set of half-samples from the set of all 2^H half-samples to estimate the variance. Balanced repeated replication uses the variability among A replicate half-samples that are selected in a balanced way.

Let the elements of each s_h be denoted as $h1$ and $h2$ and $r_{ah} = \begin{cases} 1 & \text{if half-sample } a \text{ contains element } h1 \\ -1 & \text{if half-sample } a \text{ contains element } h2 \end{cases}$. The set of A replicate half-samples is balanced if $\sum_{a=1}^{A} r_{ah} r_{al} = 0$ for all $l, h, l \neq h$.

Let also $y_h(r_a) = \begin{cases} y_{h1} \text{ if } & r_{ah} = 1 \\ y_{h2} \text{ if } & r_{ah} = -1 \end{cases}$ denote the value of the element of s_h included in half-sample a.

Finally, let $\hat{\theta}(r_a)$ be the estimate of interest, calculated in the same way as $\hat{\theta}$ but using only the observations in half-sample a, e.g. $\hat{\theta}(r_a) = \sum_{h=1}^{H} \left(\frac{N_h}{N} \right) y_h(r_a)$ if θ is the population mean. Then the BRR estimator of the variance of $\hat{\theta}$ is given by

$$\hat{V}_{BRR} = \frac{1}{A} \sum_{a=1}^{A} \left(\hat{\theta}(r_a) - \hat{\theta} \right)^2 . \quad (3.3.10)$$

BRR is a flexible method in terms of the kinds of estimators that can be accommodated. But it is restricted in terms of sampling design, as the standard BRR works when two units (or two first-stage clusters) are sampled from each stratum and data imputation is not used. However, by pairing adjacent selections in a random sampling design, BRR can also be applied to non-stratified designs. By more complicated balancing schemes or by collapsing schemes, BRR can also accommodate three-or-more-per-stratum designs and one-per-stratum designs (Wolter, 2007).

BRR is popular in the United States for variance estimation for non-linear survey estimators under stratified multi-stage sampling design. Survey agencies such as the U.S. Census Bureau, the U.S. Bureau of Labor Statistics and Westat have computer software for computing BRR variance estimates.

- **Random groups method (RG)**

The random group method of variance estimation amounts to selecting two or more samples from the population, using the same sampling design for each sample, constructing a separate estimate of the population parameter of interest from each sample and an estimate from the combination of all samples, and then computing the sample variance among the several estimates.

- **Independent random groups method**

Mutual independence of the various samples arises when one sample is replaced into the population before selecting the next sample. We assume that

we draw A independent samples $s_1,\ldots,s_A$ of equal size. The first sample s_1 is drawn by sampling from the whole population and is then replaced into the population. The second sample s_2 is then drawn by the same design that produced s_1 and independently of s_1. Then s_2 is replaced into the population. A third sample s_3 is drawn, by the sampling design that produced s_1 and s_2 and so on until A samples have been drawn. What is important is the replacement of each sample s_a before the next sample s_{a+1} is drawn.

For each $a=1,\ldots,A$, an estimator $\hat{\theta}_{(a)}$ of θ is calculated on data from s_a only. The same estimator formula applies throughout. The average of the $\hat{\theta}_{(a)}$, denoted as $\hat{\theta}_{(\cdot)}$, is used for estimating θ:

$$\hat{\theta}_{(\cdot)}=\frac{1}{A}\sum_{a=1}^{A}\hat{\theta}_{(a)}, \qquad (3.3.11)$$

whereas the independent random groups variance estimator is given by

$$\hat{V}^{*}_{IRG}=\frac{1}{A(A-1)}\sum_{a=1}^{A}\left(\hat{\theta}_{(a)}-\hat{\theta}_{(\cdot)}\right)^2. \qquad (3.3.12)$$

- **Dependent random groups method**

The method of dependent random groups is used for samples that do not meet the requirements of independent random groups. We assume that we first draw one large sample from the whole population by a probability sampling design. A random mechanism is then used to divide it into a number of disjoint subsamples, the random groups. These will not be independent but are treated as if they were.

Let s be the sample drawn from the population U, which we call full sample. We divide s into A disjoint random groups $s_1,\ldots,s_A$ such that $s=\bigcup_{a=1}^{A}s_a$. We assume that s is of a fixed size n and we also assume for simplicity that the groups are of equal size. Let $\hat{\theta}_1,\ldots,\hat{\theta}_{(a)},\ldots,\hat{\theta}_{(A)}$ be estimators of θ, where $\hat{\theta}_{(a)}$ is based only on data from the group s_a, where $a=1,\ldots,A$. We now consider the estimator $\hat{\theta}_{(\cdot)}$, formed by averaging:

$$\hat{\theta}_{(\cdot)}=\frac{1}{A}\sum_{a=1}^{A}\hat{\theta}_{(a)} \qquad (3.3.13)$$

and the estimator $\hat{\theta}$ based on data from the full sample s, disregarding the division into random groups. The variance estimator is given by

$$\hat{V}^{*}_{DRG}=\frac{1}{A(A-1)}\sum_{a=1}^{A}\left(\hat{\theta}_{(a)}-\hat{\theta}_{(\cdot)}\right)^2. \qquad (3.3.14)$$

In some cases $\hat{\theta}$ and $\hat{\theta}_{(\cdot)}$ can be so defined that they will be identical. An alternative variance estimator is given by

$$\hat{V} = \frac{1}{A(A-1)} \sum_{a=1}^{A} \left(\hat{\theta}_{(a)} - \hat{\theta} \right)^2 \quad . \qquad (3.3.15)$$

The **random group** method is a flexible method which can accommodate almost any estimator and almost any sampling design.

The general characteristics of replication methods are as follows:

- Even though ***replication methods*** have strong theoretical foundations, their purpose is rather to provide users with a sort of universal 'recipe' which could fit any type of sampling design and any type of statistics of interest (linear and more complex statistics). For a given design, the same analysis procedure is used for almost all statistics, regardless of their complexity. They can be applied to complex sampling schemes, such as multi-stage sampling. Rao *et al* (1992) describe a ***jackknife***, a ***bootstrap*** and a ***balanced repeated replication*** for such sampling schemes. A ***bootstrap*** for multi-stage sampling is also illustrated in Preston (2009).
- Another property of ***replication methods*** is that they enable users of secondary survey data to estimate standard errors without knowing the detailed sampling design. The replicate weights created by the survey methodologists can quite simply be included in the data file and be used by users of secondary survey data (e.g. researchers) to estimate the variance. This is especially useful when there are confidentiality issues involving sample units and there is need to prevent dissemination of any information that identifies the sample units. However, the release of replicate weights with the public use data files may still raise confidentiality issues. See Section 4.2 for a possible solution to this problem.
- ***Replication methods*** need computational power to perform all calculations on each replicate sample. They require more extensive computation than for instance ***Taylor series linearisation.***

- **Generalised Variance Functions (GVF)**

GVFs model the relationship between the variance or the relative variance of an estimator and its expectation. The model parameters can be determined from a set of variance estimates obtained through direct computations (using analytical or replication methods) and then used to estimate the variance of any other statistic of interest. As such, GVFs cannot entirely substitute analytical or replication methods, but they are used after the application of these methods on a set of estimates to approximate standard errors for a wide variety of estimates of target population characteristics.

GVFs are somewhat less flexible than the other methods, being designed primarily for multi-stage sample surveys of households. Although GVFs have been used in other applications, these have not been generally as successful (Wolter, 2007).

GVFs are applicable primarily to estimated proportions or to estimates of the total number of individuals in certain domains or sub-populations. There have been a few attempts, not entirely successful, to develop GVF techniques for quantitative characteristics (Wolter, 2007).

Variance estimation based on GVFs is mainly empirical: there is no theoretical evidence to guide the choice of a model. This approach is particularly suitable for handling variance estimation when the publication load is extremely heavy, with up to thousands of statistics for which variance estimates are needed. The approach is also interesting for data users,

since it enables them to estimate accuracy without any information regarding sampling design. For more information, see Section 4.2.

Evaluation criteria of variance estimation methods and related recommendations

Many estimation methods lead to adequate estimators of variance under the commonly used sampling designs. Even so, there are basic criteria which can help any survey practitioner to choose a variance estimation method:

- ***applicability of methods (to the sampling design used and the type of statistics)***

  The complexity of a variance estimation method derives from two aspects — the complexity of the statistic under study (linear statistics, non-linear but smooth statistics and non-smooth statistics) and the complexity of the sampling design. Variance estimation methods can be classified according to their fitness to deal with complex statistics and complex sampling designs, as follows:

  - methods that can be used for complex statistics: e.g. Taylor linearisation;
  - methods that can be used for complex designs: e.g. jackknife method;
  - methods that can be used both for complex statistics and complex sampling designs: e.g. jackknife method.

  The Task Force mapped the different variance estimation methods in relation to the main categories of the identified sampling designs and to the main types of statistics. *The purpose of the exercise was to lay down some recommendations on good and bad practices of using certain methods for certain sampling designs and types of statistics. The list of recommended and not recommended methods does not claim to be exhaustive. The result of this exercise is presented in Appendix 7.4.* References to the literature are also provided.

  Experience from household surveys clearly shows that for a given sampling design there is no unique method but rather several ways to handle estimates of standard errors. The choice of variance estimation method is often a matter of taste: there are 'schools' which give prominence to analytical methods, while other 'schools' consider that replication methods are better tailored to statistical production. The national approaches used to compute estimates of standard errors vary greatly from one country to another. For instance, in the EU-SILC, some countries (e.g. France, Italy) use analytical variance estimation methods, while others (e.g. Luxembourg, Spain) seem to have a preference for replication methods like bootstrap or jackknife.[28] Thus, for national sampling designs which can be regarded as 'similar' (multi-stage sample selection, non-response adjustments, calibration to external sources, etc.), the variance estimation methods used at national level are quite different from one country to another, and yet they produce estimates which are 'acceptable' from a statistical point of view as they have been released in the national quality reports.

- ***accuracy considerations*;**

  The most common accuracy criteria, also mentioned by Wolter (2007), are ***confidence interval coverage probabilities*** (the proportion of the times the interval contains the true value of the parameter over repeated samples), ***mean square error (MSE)*** (which

[28] Source: National EU-SILC Quality Reports.

is also referred to as ***stability***, see U.S. Census Bureau, 1993) and ***unbiasedness***. Cochran (1977) states that an unbiased estimator has the property that its expected value (the average value of the estimate, taken over all possible samples of given size *n*) is exactly equal to the true unknown population value.

These criteria interact in a way which seems to be unpredictable and which does not allow any generalisation about which method is best from the point of view of all criteria considered at the same time. Different studies seek to compare the accuracy of different variance estimation methods in terms of these criteria. Indeed, different variance estimation methods turn out to be the best, given different accuracy criteria. Since the most important purpose of a variance estimator will usually be to construct confidence intervals for the parameter of interest θ or for testing statistical hypotheses about θ, Wolter (2007) suggests that the most relevant criterion of accuracy will usually be the confidence interval coverage probability. Moreover, the bias criterion, and to some extent the MSE criterion, do not lead to any definitive conclusions about the different variance estimators. This is because the biases of the RG, BRR, JK and TS estimators of variance are, in almost all circumstances, identical, at least to a first-order approximation. Thus, we have to look to second- and higher-order terms to distinguish between the estimators. Since the square of the bias is one component of MSE, this difficulty also carries over to the MSE criterion of accuracy. The second component of the MSE, the variance, is under control as the survey methodologist can choose from a range of strategies about the number of random groups, partial versus full balancing etc. So the best estimator of variance is not obvious in terms of the bias and MSE criteria.

Frankel (1971) and Kish and Frankel (1974) make a Monte Carlo comparison of the performance of three methods (JK, BRR and TS) for a two-per-stratum single-stage cluster sampling design of households. The types of estimates examined were: means (that were ratio estimators), differences of means, regression coefficients and correlation coefficients. Overall, the studies (which are described in Wolter (2007) and U.S. Census Bureau (1993)) show that:

- ***BRR*** is clearly best in terms of the confidence interval criterion, while ***TS*** seems to be the worst;
- the three methods, however, performed in the reverse order in terms of the stability of variance estimator; the MSE of the ***TS*** variance estimator was smallest;
- ***TS*** and ***JK*** may have smaller biases than ***BRR***, but the patterns are not very clear or consistent.

A study undertaken by Bean (1975) (described in Wolter (2007)) and using a sampling design involving two PSUs per stratum selected by probability proportional to size with replacement sampling, shows that:

- ***BRR*** tends to offer the best confidence intervals;
- ***TS*** tends to have the smallest MSE;
- no estimator of variance consistently and generally has the smallest bias.

Wolter (2007) mentions that Mulry and Wolter (1981) and Dippo and Wolter (1984) come to similar conclusions that:

- ***BRR*** and ***RG*** are better in terms of confidence intervals;
- ***TS*** tends to have good properties in terms of MSE;

- actual confidence interval coverage probabilities tend to be too low in all cases.

Wolter (2007) mentions that with adequate replication, ***bootstrap*** should have statistical properties similar to the other replication methods such as ***BRR*** and ***JK***.

Rao and Wu (1985) make an asymptotic second-order comparison of ***JK***, ***BRR*** and ***TS*** for any stratified multi-stage design in which the primary sampling units (PSUs) are selected with replacement. When the design consists of two sampled PSUs per stratum, the TS variance estimator is shown to be identical (in second-order asymptotic expansions) to the BRR variance estimator and to the JK variance estimator for non-linear estimates such as ratio, correlation and regression coefficients. These results suggest that for two PSUs per stratum designs with a large number of strata, there is not much to choose between TS, BRR and JK variance estimators in terms of statistical criteria. It follows that practical considerations, such as available computing resources and computing costs, should dictate the choice of variance estimator. The results of the study are described in U.S. Census Bureau (1993).

For non-linear statistics that can be expressed as functions of estimated totals, Krewski and Rao (1981) establish asymptotic consistency of ***TS***, ***JK*** and ***BRR*** variance estimators as the number of strata increases (U.S. Census Bureau, 1993).

Wolter (2007) notes that the bias and the MSE of the ***RG*** method will depend on both the number and size of the random groups. Generally speaking, he found that the variance of the variance estimator declines as the number of random groups increases, while the bias increases. However it was somewhat unclear what the net effect of these competing forces is in the MSE. These remarks also apply to the bias and MSE of the ***BRR***, ***JK*** and ***TS*** variance estimators.

To sum up, with respect to accuracy, different studies show very good results for BRR when it comes to confidence interval coverage probabilities (the most relevant accuracy criterion). Some studies show very good results for TS when it comes to stability (MSE). However, it is not obvious which is the best variance estimation method in terms of the stability and bias criteria.

*As regards **GVFs** there is very little theoretical justification and the estimators of variance are surely biased. However, survey practitioners who have used these methods feel that they bring some additional stability (lower variance) to variance estimates. The GVF method is clearly inferior to the other methods in terms of confidence interval criterion* (Wolter, 2007). In conclusion, GVFs seek not to provide the best variance estimators possible, but to provide users with a sort of 'black-box' from which they can get a variance estimate for any survey statistic.

- ***administrative considerations***: cost, timeliness and simplicity (Wolter, 2007).

The ***cost*** of calculating accurate variance estimates for each statistic may indeed turn out to be formidable when the publication load is quite heavy (with hundreds, perhaps thousands of statistics at stake). If so, cost-effective methods, though less accurate, are highly desirable as a means of handling such a situation.

Timeliness obviously refers to the amount of time needed to produce the variance estimates, which should be set in accordance with the survey deadlines.

Simplicity refers to the need for simple methods applicable to (although possibly not optimal for any of) the multitude of parameters that may need to be estimated from a survey's data. It also refers to the need for methods which are (a) simple enough to

program and (b) simple enough to be understood by stakeholders of the survey, e.g. its main users.

*Wolter (2007) mentions that **GVFs** cannot be recommended for any but the very largest sample surveys, where administrative considerations prevail. If we are dealing with hundreds, perhaps thousands of indicators (also considering breakdowns) for which variance estimates are wanted, then given the time constraints, working out variance estimates one by one becomes unfeasible using direct variance computations (computation from the microdata, with analytical or replication methods).*

Summary

For a given sampling design and type of statistic there is no one unique method but rather several methods of estimating standard errors. There is consequently also a broad range of variability in the methods used by countries to compute estimates of standard errors. *The recommendation is to use variance estimation methods which are appropriate to the sampling design and type of estimator. Appendix 7.4 presents some recommendations on good and bad practices of using certain methods for certain sampling designs and types of statistics. The list of recommended and not recommended methods does not claim to be exhaustive.* This appendix has been devised to help the survey manager choose the appropriate method, from the applicability point of view.

Other criteria for the choice of methods are accuracy (confidence interval coverage probabilities, unbiasedness and stability) and administrative considerations (time, cost, simplicity). *With respect to accuracy, different studies show very good results for BRR when it comes to confidence interval coverage probabilities (the most relevant accuracy criterion). Some studies show very good results for TS when it comes to stability (MSE). However, it is not obvious which is the best variance estimation method in terms of the stability and bias criteria. There is very little theoretical justification for GVFs and the estimators of variance are surely biased. However, survey practitioners who have used these methods feel that they bring some additional stability to variance estimates. The GVF method is clearly inferior to the other methods in terms of confidence interval criterion. With respect to administrative considerations, GVFs are suitable for the very largest sample surveys with hundreds, perhaps thousands of indicators (also considering breakdowns) for which variance estimates are wanted.*

3.4 Some recommended variance estimation methods to account for different sources of variability

Yves Berger (University of Southampton) and Ralf Münnich (University of Trier)

This section presents some recommended variance estimation methods which can be used to estimate or incorporate the different sources of variability (presented in Section 3.2) in the variance as a whole.

Estimating variance in the case of indirect sampling

Indirect sampling is applicable to situations that involve two populations U_A and U_B that are linked by some relation, and where we want to produce estimates for one of them, say U_B. Suppose that we have a sampling frame for U_A only. We select a sample from U_A in order to obtain a sample from U_B using the links between the two populations. For example consider U_A to be individuals and U_B households; then a sample of individuals may be selected from a population register, and a sample of households is obtained by taking all households that have at least one of their current members in the selected sample of individuals. The selected households are weighted by the Generalised Weight Share Method (GWSM) (Lavallée, 2007; Lavallée and Caron, 2001). Variance estimation for indirect sampling of households considers individuals (and not households) as the ultimate sampling units. However, to account for the unequal probabilities of selection of the households, the study variable is adjusted by household size. For example, suppose a sample s_A of m_A units is selected from the population U_A of M_A units using some sampling design. Let $\pi_j^{(A)}$ be the selection probability of unit j. Let the population U_B contain M_B units. This population is divided into N clusters, where each cluster i contains $M_{B,i}$ units. We are interested in estimating the total $Y_B = \sum_{i=1}^{N} \sum_{k=1}^{M_{B,i}} y_{ik}$ for some characteristic y over population U_B. With GWSM, we make the following assumptions:

1. There is a link between each unit j of population U_A and at least one unit k of cluster i of population U_B.
2. Each cluster i of U_B has at least one link with a unit j of U_A.
3. There can be zero, one or more links for a unit k of cluster i of population U_B.

By using GWSM we assign an estimation weight w_{ik} to each unit k of an interviewed cluster i. To estimate the total Y_B belonging to population U_B, we can use the estimator

$$\hat{Y}_B = \sum_{i=1}^{n} \sum_{k=1}^{M_{B,i}} w_{ik} y_{ik} \quad , \quad (3.4.1)$$

where n is the number of selected clusters and w_{ik} is the weight attached to unit k of cluster i.

With the GWSM, the estimation process uses the sample s_A together with the links between U_A and U_B to estimate the total Y_B. The links are in fact used as a bridge to go from population U_A to population U_B and vice versa.

The GWSM allocates each sampled unit a final weight established from an average of weights calculated within each cluster i entering into $\hat{Y}$. An *initial weight* that corresponds to the inverse of the selection probability is first obtained for unit k of cluster i of $\hat{Y}$ having a non-zero link with a unit j belonging to s_A. An initial weight of zero is assigned to units not having a link. The *final weight* is obtained by calculating the ratio of the sum of the initial weights for the cluster over the total number of links for that cluster. This final weight is finally assigned to all units within the cluster. Note that allocating the same estimation weight to all units has the considerable advantage of ensuring consistency of estimates for units and clusters.

In equation (3.4.1) we assign *the final weight* $w_{ik} = w_i$ for all $k \in i$, which is

$$w_i = \frac{\sum_{k=1}^{M_{B,i}} w'_{ik}}{\sum_{k=1}^{M_{B,i}} L_{ik}} \quad , (3.4.2)$$

where L_{ik} represents the number of links between the units of U_A and the unit k of cluster i of U_B, and the *initial weights* are

$$w'_{ik} = \sum_{j=1}^{M_A} L_{j,ik} \frac{t_j}{\pi_j^{(A)}}. \quad (3.4.3)$$

In the above equation, $L_{j,ik} = 1$ if there is a link between unit j of population U_A and k of cluster i of population U_B and zero otherwise, and $t_j = 1$ if $j \in s_A$ and zero otherwise.

Now let $z_{ik} = \frac{Y_i}{L_i}$ for all $k \in i$, with $Y_i = \sum_{k=1}^{M_{B,i}} y_{ik}$ and L_i being the number of links present in cluster i, $L_i = \sum_{k=1}^{M_{B,i}} L_{ik}$.

Then

$$\hat{Y}_B = \sum_{j=1}^{M_A} \frac{t_j}{\pi_j^{(A)}} \sum_{i=1}^{N} \sum_{k=1}^{M_{B,i}} L_{j,ik} z_{ik} = \sum_{j=1}^{M_A} \frac{t_j}{\pi_j^{(A)}} Z_j \quad (3.4.4)$$

and the variance of $\hat{Y}$ is directly given by

$$V\left(\hat{Y}_B\right) = \sum_{j=1}^{M_A} \sum_{j'=1}^{M_A} \frac{\pi_{jj'}^{(A)} - \pi_j^{(A)} \pi_{j'}^{(A)}}{\pi_j^{(A)} \pi_{j'}^{(A)}} Z_j Z_{j'} \quad . \quad (3.4.5)$$

where $\pi_{jj'}^{(A)}$ is the joint probability of selecting units j and j'. See Särndal, Swensson and Wretman (1992) for how to calculate $\pi_{jj'}^{A}$ under various sampling designs.

Accounting for the variability caused by systematic sampling and implicit stratification

Since a systematic sample can be regarded as a random selection of one cluster, it is not possible to give an unbiased, or even consistent, estimator of the design variance (Wolter, 2007).

- When there is no particular 'structure' in the sampling frame, that is, when it appears that units have been randomly ordered, then the standard variance estimator under simple random sampling can be used if we accept that random ordering is part of sampling design, because in this situation the sampling design is a simple random sampling design. Let us assume we want to estimate the population total of a study variable y. The following estimator of the variance of the population total can be used:

$$\hat{V}_{1,i}\left(\hat{Y}\right)=N^2\left(\frac{1}{n}-\frac{1}{N}\right)s^2\ , \quad (3.4.6)$$

where s^2 is the sample variance of the study variable y. Let $y_{i,j}$ denote the value of y for the j^{th} individual ($j=1\cdots n$) of the i^{th} systematic sample. Let us assume the following super-population model:

$$(M_1): y_{i,j}=\alpha+\varepsilon_{i,j}\ , (3.4.7)$$

where $\varepsilon_{i,j}$ are random variables of null expectation, with variance σ^2 (independent of i and j) and uncorrelated among themselves. We thus get (Ardilly and Tillé, 2005):

$$E_{M_1}\left(\hat{V}_{1,i}\left(\hat{Y}\right)-V\left(\hat{Y}\right)\right)=0\ , \quad (3.4.8)$$

where E_{M_1} is the expectation under the model (3.4.7) and V is the exact variance of the Horvitz-Thompson estimator of the population total (see *Appendix 7.1*).

- If the frame is sorted according to an auxiliary variable correlated to y (**implicit stratification** present), as specified by the following super-population model:

$$(M_2): y_{i,j}=\alpha(i+jg)+\beta+\varepsilon_{i,j}\ , (3.4.9)$$

where $\varepsilon_{i,j}$ are random variables of null expectation, with variance σ^2 (independent of i and j), uncorrelated among themselves and $g=N/n$, this gives rise to (Ardilly and Tillé, 2005):

$$E_{M_2}\left(\hat{V}_{1,i}\left(\hat{Y}\right)-V\left(\hat{Y}\right)\right)=\frac{N^2}{n}\frac{\alpha^2}{12}\ , (3.4.10)$$

where E_{M_2} is the expectation under the model (3.4.9) and V is the exact variance of the Horvitz-Thompson estimator of the population total. Thus, when $\alpha>0$, the naive variance estimator under simple random sampling overestimates the exact variance.

- When ***implicit stratification*** is present, the following variance estimator (Wolter, 2007; Nikolaidis, 2010a) can be used. It means considering a systematic sample as a stratified

simple random sample with two units selected from each successive stratum. This yields an estimator based on non-overlapping differences:

$$\hat{V}_{2,i}\left(\hat{Y}\right)=N^2\left(\frac{1}{n}-\frac{1}{N}\right)\sum_{j=1}^{n/2}\frac{\left(y_{i,2j}-y_{i,2j-1}\right)^2}{n} \quad . (3.4.11)$$

Under the model (3.4.7) (population in random order), we get:

$$E_{M_1}\left(\hat{V}_{2,i}\left(\hat{Y}\right)-V\left(\hat{Y}\right)\right)=0 \quad . (3.4.12)$$

Under the model (3.4.9) (implicit stratification), we get (Ardilly and Tillé, 2005):

$$E_{M_2}\left(\hat{V}_{2,i}\left(\hat{Y}\right)-V\left(\hat{Y}\right)\right)\cong-\frac{\alpha^2}{12}\left(\frac{N}{n}\right)^2. \quad (3.4.13)$$

Thus, under implicit stratification, $\hat{V}_2$ underestimates the true variance. If we now combine the estimators $\hat{V}_1$ and $\hat{V}_2$ by defining

$$\hat{V}_{comb,i}\left(\hat{Y}\right)=\frac{1}{n+1}\hat{V}_{1,i}\left(\hat{Y}\right)+\frac{n}{n+1}\hat{V}_{2,i}\left(\hat{Y}\right), \quad (3.4.14)$$

$$\text{we get: } E\left(\hat{V}_{comb,i}\left(\hat{Y}\right)-V\left(\hat{Y}\right)\right)=0 \text{ . (3.4.15)}$$

- When ***implicit stratification*** is present, an alternative estimator based on overlapping differences (Wolter, 2007) that seeks to increase the number of 'degrees of freedom', can be used:

$$\hat{V}_{3,i}\left(\hat{Y}\right)=N^2\left(\frac{1}{n}-\frac{1}{N}\right)\sum_{j=2}^{n}\frac{\left(y_{i,j}-y_{i,j-1}\right)^2}{2(n-1)} \quad . (3.4.16)$$

Under the model (3.4.7) (population in random order), we also have:

$$E_{M_1}\left(\hat{V}_{3,i}\left(\hat{Y}\right)-V\left(\hat{Y}\right)\right)=0 \quad . (3.4.17)$$

Under the model (3.4.9) (implicit stratification), the formulae (3.4.13), (3.4.14) and (3.4.15) can be applied to $\hat{V}_3$ instead of $\hat{V}_2$.

The estimator (3.4.16) is used in the variance estimation software POULPE (Caron, 1998; Ardilly and Osier, 2007).

- When ***implicit stratification*** is present, the variance estimator proposed by Berger (2005) can be used. It takes systematic sampling into account by using the order of the units in the population. This produces a variance estimator with reduced bias under systematic sampling for a given (non-random) order of the population. This estimator can be used with any given population order. However, we need to know the order of the units in the population. Simulation based on the IBGE[29] (Brazil) household surveys shows that this estimator has less bias than classical estimators.

[29] The Brazilian Institute of Geography and Statistics.

- A jackknife variance estimator for systematic sampling requires pairing of the sampled clusters (PSUs), with adjacent clusters, in the systematic selection order, being paired. A replicate is formed by deleting one cluster from the sample, doubling the weight of its complementary pair member, and recalculating the estimator $\hat{\theta}$. In the case of stratified sampling, the procedure is carried out n_h times in each stratum h $(h=1,2,\ldots,H)$, by dropping each cluster in turn. For implicit stratification, the variance estimator for $\hat{\theta}$ is given by Burke and Rust (1995):

$$\hat{V}_{JK}(\hat{\theta}) = \sum_{h=1}^{H} \frac{1}{2} \sum_{j=1}^{n_h} \left(\hat{\theta}_{(hA)} - \hat{\theta}\right)^2 \quad . (3.4.18)$$

  $\hat{\theta}_{(hA)}$ is the estimate obtained at the replication A in stratum h after deleting one cluster, doubling the weight of the complementary pair member, and recalculating the estimator $\hat{\theta}$.

- Similarly, there are many (biased) variance estimators which we can use in the event of unequal probability sampling (Wolter, 2007). By treating the sample as a stratified random sample with two units selected per stratum, we obtain the following estimators:

$$\hat{V}_{4,i}\left(\hat{Y}\right) = \frac{1}{n} \sum_{j=1}^{n/2} \frac{\left(\frac{y_{i,2j}}{p_{i,2j}} - \frac{y_{i,2j-1}}{p_{i,2j-1}}\right)^2}{n} \quad (3.4.19)$$

$$\hat{V}_{5,i}\left(\hat{Y}\right) = \frac{1}{n} \sum_{j=2}^{n} \frac{\left(\frac{y_{i,j}}{p_{i,j}} - \frac{y_{i,j-1}}{p_{i,j-1}}\right)^2}{2(n-1)} \quad , (3.4.20)$$

  where $p_{i,j}$ is the inclusion probability of the unit j of the i^{th} systematic sample.

- To deal with systematic sampling with unequal probabilities, we can use approximate variance formulae, wherein double inclusion probabilities are not used. These approximate variance formulae are defined in the next section.

Accounting for the variability caused by using unequal probability sampling

Calculating the probabilities of selection of the order two, i.e. the probability π_{ij} that two distinct units i and j be selected in the sample, is difficult for certain sampling designs. For simple random sampling without replacement of size n from a population of size N, we have: $\pi_{ij} = n(n\text{-}1) / N(N\text{-}1)$. On the other hand, when units are selected with unequal probabilities (e.g. probability proportional to size) and without replacement, it is not generally possible to fix any formula for the double inclusion probabilities. To overcome this, approximate variance formulae can be used, with the double inclusion probabilities not being used. The

joint inclusion probabilities π_{ij} are approximated in terms of first-order inclusion probabilities π_i.

Some variance estimators free of joint inclusion probabilities are:

$$\hat{V}_6\left(\hat{\bar{Y}}\right)=\frac{n}{n-1}\sum_{i=1}^{n}\left(1-\pi_i\right)\left(\frac{y_i}{\pi_i}-B\right)^2 \quad (3.4.21) \quad \text{(Hájek, 1964)}, \quad B=\frac{\sum_{i=1}^{n}\frac{y_i}{\pi_i}(1-\pi_i)}{\sum_{i=1}^{n}(1-\pi_i)}$$

$$\hat{V}_7\left(\hat{\bar{Y}}\right)=\frac{n}{n-1}\sum_{i=1}^{n}\left(1-\pi_i\right)\left(\frac{y_i}{\pi_i}-A\right)^2 \quad (3.4.22) \quad \text{(Rosen, 1991)}, \quad A=\frac{\sum_{i=1}^{n}y_i\frac{1-\pi_i}{\pi_i^2}\cdot\log(1-\pi_i)}{\sum_{i=1}^{n}\frac{1-\pi_i}{\pi_i}\cdot\log(1-\pi_i)}$$

$$\hat{V}_8\left(\hat{\bar{Y}}\right)=\frac{1}{1-\sum_{i=1}^{n}a_i^2}\sum_{i=1}^{n}\left(1-\pi_i\right)\left(\frac{y_i}{\pi_i}-\sum_{i=1}^{n}\frac{a_i y_i}{\pi_i}\right)^2 \quad (3.4.23) \text{ (Deville, 1999)}, \quad a_i=\frac{1-\pi_i}{\sum_{i=1}^{n}(1-\pi_i)}.$$

Accounting for the variability caused by calibration

Calibration consists of computing weights that incorporate specified auxiliary information and which are constrained by calibration equations. It is now common practice in household surveys to calibrate sampling weights to auxiliary data sources, thereby improving the accuracy of estimates.

A calibration estimator uses calibrated weights which are as close as possible, according to a given distance function, to the original sampling design weights, while also complying with a set of constraints, the calibration equations. For every distance function there is a corresponding set of calibrated weights and a calibration estimator.

Deville *et al* (1993) present four different 'methods' corresponding to four different distance functions, i.e. the linear, raking, logit and truncated linear methods. The estimators based on the linear calibration method, i.e. with quadratic distance function and linear calibration function, are generalised regression (GREG) estimators.

A major result of calibration theory states that the variance of a calibration estimator is (asymptotically) equal to that of the estimator based on the non-calibrated weights, but where the study variable has been replaced by the residuals u_i of its regression on the calibration variables (Deville and Särndal, 1992). The following formula can be used:

$$\hat{V}\left(\hat{Y}_w\right)=\hat{V}\left(\sum_{i\in s}w_i y_i\right)\approx\hat{V}\left(\sum_{i\in s}d_i u_i\right), \quad (3.4.24)$$

where w_i is the sampling weight of i after calibration, d_i is the sampling weight of i before calibration and u_i is the residual of i from the regression of the study variable y on the calibration variables.

Thus, variance decreases strongly when calibration variables are more explanatory.

An important issue with the approximation of (3.4.24) is that it applies when the original statistic is linear (e.g. total, mean). If the statistic is non-linear it has to be linearised first; (3.4.24) can then be applied to the linearised estimator. A non-linear, smooth statistic (e.g. a ratio of totals) can be linearised with Taylor series approximation. A non-linear, non-smooth statistic (e.g. Gini coefficient) can be linearised using influence functions or estimating equations.

Using calibration to compensate for non-response can be generally recommended only if the response probability of statistical units can be modelled or if the rate of non-response amounts to just a few percent (Särndal, 2007 and Bethlehem and Schouten, 2004). Although these conditions are rarely met nowadays, this kind of calibration use is still quite common and is accepted since stopping it would result in even greater bias than is currently the case.

For regression estimation we recommend the 'g-weighted' variance estimator (Särndal *et al*, 1989):

$$\hat{V}\left(\sum_{i\in s} w_i u_i\right). \quad (3.4.25)$$

The estimator (3.4.25) uses an adjustment of the variance estimator based on first-order Taylor linearisation; it better accounts for the dispersion of the GREG weights.

The form of equation (3.4.25) and the corresponding estimator of variance depend on the details of the sampling design (Wolter, 2007). For example, for stratified random sampling, the estimator of the variance is

$$\hat{V}(\hat{Y}_{GREG}) = \sum_{h=1}^{H}\left(1-\frac{n_h}{N_h}\right)\frac{n_h}{n_h-1}\sum_{i\in s_h}\left(w_{hi}u_{hi}-\frac{1}{n_h}\sum_{i\in s_h} w_{hi}u_{hi}\right)^2. \quad (3.4.26)$$

For simple random sampling without replacement, the estimator is

$$\hat{V}(\hat{Y}_{GREG}) = \left(1-\frac{n}{N}\right)\frac{n}{n-1}\sum_{i\in s}\left(w_i u_i-\frac{1}{n}\sum_{i\in s} w_i u_i\right)^2. \quad (3.4.27)$$

For two-stage sampling (when primary sampling units are selected with probabilities proportional to size and secondary sampling units are selected by simple random sampling without replacement), the estimator is

$$\hat{V}(\hat{Y}_{GREG}) = \sum_{h=1}^{H}\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}\left(\sum_{j=1}^{m_{hi}} w_{hij}u_{hij}-\frac{1}{n_h}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}u_{hij}\right)^2. \quad (3.4.28)$$

Most general-purpose software packages do not take the impact of calibration into account in variance estimation. However, there are specific programs which can estimate variance correctly (see Section 3.5). A way of doing it would be for the user to first compute the regression residuals and then substitute these for the study variable.

Accounting for the variability caused by unit non-response

The main point is to build a variance estimator $\hat{V}\left(\hat{\theta}\right)$ that takes into account unit non-response.

- Unit non-response is generally viewed as an additional phase of sampling. So the overall variance $V\left(\hat{\theta}\right)$, taking into account both sampling design and unit non-response, can be split into two components, i.e. the first-phase variance $V_S\left(\hat{\theta}\right)$ (sampling variance) caused by selecting a sample s from a target population U, and the second-phase variance $V_{NR}\left(\hat{\theta}\right)$ (non-response variance) caused by having a subset r of respondents from s:

$$V\left(\hat{\theta}\right)=V_S\left(\hat{\theta}\right)+V_{NR}\left(\hat{\theta}\right). \quad (3.4.29)$$

  Thus, unit non-response introduces extra variability in the form of an additional variance component $V_{NR}\left(\hat{\theta}\right)$. The estimation of $V_{NR}\left(\hat{\theta}\right)$ relies on further assumptions concerning the non-response mechanism (which is unknown). For example, the variance estimation software POULPE, developed by the French NSI (INSEE), treats unit non-response either as a Poisson sampling or a post-stratified sampling phase (Ardilly and Osier, 2007). Särndal and Lundström (2005) proposed an estimator for $V_{NR}\left(\hat{\theta}\right)$ in the case of calibration estimators.

- Under the assumption that the set of respondents is a random sample from the original sample and the values of the study variable for the non-respondents are missing at random, the net sample size (the number of respondents) can be used in variance estimation formulae instead of the gross sample size (the number of units in the original sample).

  In simple random sampling, the gross sample size can be simply replaced by the net sample size. In stratified random sampling, the gross sample size in every stratum can be replaced by the corresponding net sample size.

- If the values of the study variables for the non-respondents are not missing at random and the probability of response is related to the study variable, response homogeneity groups can be formed, within which net sample size can be used for variance estimation instead of gross sample size. Logistic regression models can be used to estimate the response probability for an individual (respondent or non-respondent) based on the individual's characteristics (also called explanatory variables). Individuals are then divided into classes based on the size of their predicted response probability.

- *Fay's approach* (Fay, 1991; Shao and Steel, 1999) is another recommended approach. It has clear practical and theoretical advantages over the classic two-phase approach (Rao, 1990; Särndal, 1990; Deville and Särndal, 1994): the variance estimator is simple, robust, unbiased under the real response mechanism and can be calculated using standard variance estimation packages. It is therefore not necessary to assume a response mechanism such as missing at random (MAR) or missing completely at random (MCAR) (Rubin, 1976). The usual assumption is that the finite population can be divided into J imputation cells. An additional assumption for the design-based approach is that in each imputation cell, the response probability for a given variable is a constant and the response statuses for different units are independent; imputation is carried out within each imputation cell and independently across the imputation cells. An additional assumption for the model-based approach is that in each imputation cell, the response mechanism is unconfounded in the sense that whether or not a unit responds does not depend on the variable being imputed (but may depend on the covariates used for imputation). Imputation is carried out independently across the imputation cells, and

within an imputation cell imputation is performed according to a model that relates the variable being imputed to the covariates used for imputation.

Accounting for the variability caused by imputation

- The Rao and Shao adjusted jackknife method (Rao and Shao, 1992) adjusts the imputed values for each jackknife pseudo-replicate. This method produces consistent variance estimators for smooth statistics. The Rao and Shao method is valid under simple random sampling. Berger and Rao (2006) propose a modified jackknife estimator which takes imputation and unequal probabilities into account.
- The bootstrap method can incorporate imputation for both smooth and non-smooth statistics. A bootstrap for imputed survey data is given in Shao and Sitter (1996). They propose a bootstrap method for stratified multi-stage designs which avoids the adjusted imputed values used in the jackknife method of Rao and Shao (1992). The idea is to re-impute the bootstrap data set in the same way as the original data set is imputed. This method therefore requires much more computation than the jackknife, although bootstrap provides an approximation to the entire distribution of the statistic.

  Saigo *et al* (2001) propose a modified bootstrap that does not require rescaling so that Shao and Sitter's procedure can be applied to cases where random imputation is applied and the first-stage stratum sample sizes are very small. This gives a unified method that works irrespective of the imputation method (random or non-random), the stratum size (small or large) or the type of estimator (smooth or non-smooth).
- The standard balanced repeated replication (BRR) method does not account for the increase in variance due to imputation. An adjusted BRR method can adjust the imputed values for each replication (pseudo-replicated data set). For several popular single-imputation methods, this adjusted BRR method produces consistent variance estimators for functions of estimated totals and sample quantiles under certain regularity conditions (for the adjusted BRR method see e.g. Shao *et al* (1998)).
- Multiple imputation methods (Rubin, 1987; Davison and Sardy, 2007) offer opportunities to derive variance estimators taking imputation into account. In multiple imputation each missing value is replaced, instead of a single value, by a set of plausible values that reflect the uncertainty about what is the right value to impute. The incorporation of imputation can be easily derived based on the variability of the estimates among the multiple imputed data sets.

  The procedure is described by Wolter (2007) on the basis of Rubin (1987). For the procedure developed by Rubin (1987) the imputations must be 'proper', which essentially means that they are drawn from Bayesian posterior distributions. In national statistical institutes the methods used for imputation seldom satisfy the requirement of being 'proper'.[30] Bjørnstad (2007) has given alternative combination rules for other methods of imputation. In this example we consider that missing values are imputed using hot-deck imputation.

  - Make D independent hot-deck imputations for each missing item.
  - Construct (conceptually) D complete data sets, each consisting of all reported data plus one set of the imputed data.

[30] The variability in non-proper imputations is too small, and the between-imputation component must be given a larger weight in the variance estimate.

- Estimate the population total, say $\hat{Y}_d$, using each complete data set *d=1,…,D.*
- Estimate the variance of the estimated total, say $\hat{V}(\hat{Y}_d)$, using each complete data set and a variance estimation method.
- Estimate the variability between the complete data sets as an allowance for the imputation variance.
- Estimate the total variance as the sum of the within-data-set variance (the average of the $\hat{V}(\hat{Y}_d)$) and the between-data-set variance.
- Let *nr* be the non-response rate.

The multiple imputation estimator of the variance is given by:

$$\hat{V}_{MI}(\hat{\bar{Y}}) = \frac{1}{D}\sum_{d=1}^{D}\hat{V}(\hat{Y}_d) + \left(\frac{1}{1-nr} + D^{-1}\right)\frac{1}{D-1}\sum_{d=1}^{D}(\hat{Y}_d - \hat{\bar{Y}})^2 , \quad (3.4.30)$$

where $\hat{\bar{Y}} = \sum_{d=1}^{D}\frac{\hat{Y}_d}{D}$.

The second term in (3.4.30) is the between-data-set variance, which makes an allowance for the imputation variance.

In the case of 'proper' imputation methods, the term 1/(1-*nr*) is replaced by 1.

- Kim and Fuller (2004) propose fractional hot-deck imputation, which replaces each missing observation with a set of imputed values and assigns a weight to each imputed value. For example, three imputed values might be assigned to each non-respondent, with each entry allocated a weight of one-third of the non-respondent's original weight. The paper shows that fractional hot-deck imputation is an effective imputation procedure under a model in which observations in an imputation cell are independently and identically distributed. The paper suggests a consistent replication variance estimation procedure for estimators computed by fractional imputation. Simulations show that fractional imputation and the suggested variance estimator are superior to multiple imputation estimators in general, and much superior to multiple imputation for estimating the variance of a domain mean.
- Analytical methods for incorporating imputation, under the assumption of simple random sampling, are presented in Eurostat (2002) — for mean and hot-deck imputation, as well as for ratio imputation. *To obtain better variance estimators, random imputation methods such as hot-deck methods are recommended (Ardilly, 2006) as they introduce extra variability due to the random component of the imputation model.*
- Deville and Särndal (1994) researched the issue of variance estimation for the regression imputed Horvitz-Thompson estimator under the classical two-phase approach (Rao, 1990; Särndal, 1990). By assuming that the non-response mechanism is missing at random (MAR) or missing completely at random (MCAR) (Rubin, 1976), it was possible to divide the overall variance (taking both the sampling design and the imputation process into account) into a sampling variance and an imputation variance. The latter can be estimated from sample data. This framework underlies the software SEVANI (Beaumont and Mitchell, 2002), developed by Statistics Canada in order to estimate the variance due to non-response and imputation.

Accounting for the variability caused by coverage (frame) errors

- **Over-coverage:**

  The methodology of domain estimation can be used, where the target population has to be defined as a domain of the frame population.

  In order to quantify the loss of precision due to over-coverage, we can calculate the ratio R between the variance of the standard Horvitz-Thompson estimator[31] under simple random sampling without replacement — for sample size n and assuming that a proportion P $(0 < P \leq 1)$ of the frame units is eligible — to that assuming no over-coverage:

$$R = \frac{1}{P} \times \left(1 + \frac{Q}{CV_y^2} \right), \qquad (3.4.31)$$

  where $Q = 1 - P$ and CV_y is the coefficient of variation of the study variable y,

  $\frac{1}{P}$ represents the increase of variance due to an average reduction of sample size: basically, the lower the value of P, the more important the over-coverage and the greater the increase in variance,

  $1 + \frac{Q}{CV_y^2}$ represents the increase in variance due to the random size of the final sample

  (Cochran, 1977; Särndal *et al*, 1992; Ardilly and Tillé, 2005).

- **Multiple listings**: Since multiple listings can be viewed as a particular kind of over-coverage (all duplications of population units form a set of non-eligible units), the amount of variability they create can be estimated using (3.4.31). Since multiple frame units have higher selection probabilities, an alternative option to taking multiple listings into account in variance calculations would be to use estimators for unequal probability designs. However, this option requires the (unequal) probabilities of selection to be known for each element in the population, which is unlikely to happen in practice.

Accounting for the variability caused by measurement errors/ processing errors/ substitution errors

Measurement errors arise because response values which are collected through a fixed survey procedure may differ from true values. Processing and substitution errors can be treated the same way as measurement errors (see Section 3.2).

A simple model for response value is given by (Särndal *et al*, 1992):

$$(M_3): y_i = \mu_i + \varepsilon_i, \quad (3.4.32)$$

where y_i is the (random) response given by respondent i to a study variable y, μ_i is the expectation of y_i under (M_3) and ε_i is a random component, of mean 0 and variance σ_i^2. The model assumes that the responses given by respondents i and $j, (i \neq j)$ are not

[31] See Appendix 7.1.

independent, and that the covariance between y_i and y_j $(i \neq j)$ is equal to σ_{ij}. To sum up, the model (M_3) satisfies the following assumptions:

$$(M_3): \begin{cases} E_{M_3}(y_i) = \mu_i \\ V_{M_3}(y_i) = \sigma_i^2 \\ Cov_{M_3}(y_i, y_j) = \sigma_{ij} \ (i \neq j) \end{cases} . \quad (3.4.33)$$

For simplicity, let us assume that the model (M_3) is unbiased, that is, the expected value μ_i is equal to the true value of the study variable, say Y_i, for all i. The joint variance $V(\hat{Y}_{HT})$ of the Horvitz-Thompson estimator $\hat{Y}_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i}$ of the total $Y = \sum_U Y_i$ under the sampling design $p = \{p(s), s \in S_0\}$ and the measurement model (M_3) can be written as:

$$\begin{aligned} V(\hat{Y}_{HT}) &= V_p(\hat{Y}_{HT}) + \sum_U \frac{\sigma_i^2}{\pi_i} + \sum_{k \neq l} \frac{\pi_{kl}}{\pi_k \pi_l} \sigma_{kl} \\ &= V_p(\hat{Y}_{HT}) + V_m(\hat{Y}_{HT}) \end{aligned} \quad (3.4.34)$$

$V_p(\hat{Y}_{HT})$ is the sampling variance,

$V_m(\hat{Y}_{HT}) = \sum_U \frac{\sigma_i^2}{\pi_i} + \sum_{k \neq l} \frac{\pi_{kl}}{\pi_k \pi_l} \sigma_{kl} = V_{m1}(\hat{Y}_{HT}) + V_{m2}(\hat{Y}_{HT})$ is the measurement variance,

$V_{m1}(\hat{Y}_{HT}) = \sum_U \frac{\sigma_i^2}{\pi_i}$ is the simple measurement variance,

$V_{m2}(\hat{Y}_{HT}) = \sum_{k \neq l} \frac{\pi_{kl}}{\pi_k \pi_l} \sigma_{kl}$ is the correlated measurement variance.

In practice, it is difficult to estimate the additional variance due to measurement errors as repeated measurements of the variable are needed. Each new set of measurements must be conducted under the same conditions as the others and, above all, must be uncorrelated with the other sets. This latter assumption is an important one.[32]

Let us assume that we select a simple random sub-sample r, of size n_r from the original sample s, of size n. For each element in r, we observe the study variable y a second time, under the same conditions as for the first round of observations. Then, the following provides unbiased variance estimators for $V_{m1}(\hat{Y}_{HT})$ and $V_{m2}(\hat{Y}_{HT})$ (Särndal *et al*, 1992):

$$\hat{V}_{m1}(\hat{Y}_{HT}) = \frac{n}{2m} \sum_{i \in r} \left(\frac{z_i}{\pi_i} \right)^2 \quad (3.4.35)$$

$$\hat{V}_{m2}(\hat{Y}_{HT}) = \frac{n(n-1)}{2m(m-1)} \left[\left(\sum_{i \in r} \frac{z_i}{\pi_i} \right)^2 - \sum_{i \in r} \left(\frac{z_i}{\pi_i} \right)^2 \right], \quad (3.4.36)$$

[32] Unfortunately, recall effects may produce this kind of unwanted correlation over time.

where for all i in r, z_i is the difference between the two measurements of the study variable y: $z_i = y_{i,1} - y_{i,2}$.

Summary

This section presents recommended variance estimation methods which can be used to estimate or incorporate different sources of variability in the total variance. The following sources of variability are accounted for in the methods: indirect sampling, implicit stratification, unequal probability sampling, calibration, unit non-response, imputation, coverage errors and measurement errors. Processing and substitution errors can be treated in the same way as measurement errors.

3.5 Software tools for variance estimation: presentation

Kari Djerf (Statistics Finland)

There are many software packages available which can calculate variance estimates for linear and non-linear statistics under simple and complex sampling designs. In this section, we present the pros and cons of the software tools most frequently used for variance estimation. Given that software markets evolve quite rapidly, readers are advised to keep up to date with any changes that may affect such markets. Some outdated software tools (like the ones written for the DOS operating system) are excluded from this review.

For multi-stage sampling designs, most of the software packages determine the overall sampling variance by calculating the variance of the estimated PSU totals between the primary sampling units (PSU): this is known as ultimate cluster approximation (Kalton, 1979). The variance estimator thus determined underestimates the true variance (but overestimates the first-stage variance), though the underestimation is small when the sampling fraction at first stage is low.

To deal with non-linear statistics, many software tools offer the option of using either the Taylor linearisation method or sample re-use approximation (replication methods).

However, as mentioned in Section 3.3, the most common sample re-use method (replication method), namely the delete-one jackknife, should not be used in stratified sampling (see Wolter, 2007, pp. 172-173); nor should the delete-one or groups jackknife be used with non-smooth statistics (e.g. median), except for the Gini coefficient. See Section 3.3 for more information.

On the other hand, calibrated weights (Deville and Särndal, 1992) are often used in official statistics. Most general-purpose software products do not contain any proper variance estimator that takes the impact of calibration into account in variance estimation. However, there are dedicated software products capable of properly estimating variance. Some of these dedicated software products are presented in this section, under sub-section ‘Special sampling variance software for GREG’. In addition some R programs are available.

Comprehensive (commercial) statistical packages

Three widely used commercial software packages include modules which can be used for complex survey data analysis:

- SAS (v. 9.2.3 with STAT module);
- IBM SPSS (v. 19 with Complex Samples module);
- STATA (v. 11).

Currently, SPSS offers only the Taylor linearisation method for variance approximation, while SAS and STATA also offer sample re-use (replication) methods: delete-one jackknife and balanced repeated replication. Bootstrap weights can also be provided for SAS and STATA. Multiple imputation is available in both SAS (proc MIANALYZE) and STATA but not directly in the survey sampling procedures.

Other general statistical packages

- **R language**

Various authors have helped write programs with the R language for free; these programs are available for various platforms and operating systems. New procedures are envisaged since the R-community is developing new methods rapidly. A good reference on available packages in R for official statistics and survey methodology is 'CRAN Task View: Official Statistics & Survey Methodology' (http://cran.r-project.org/web/views/OfficialStatistics.html). There are several packages available for variance estimation:

- *Package sampling* (http://cran.r-project.org/web/packages/sampling/index.html). There is a function 'calibev' for calibration estimator and its variance estimation;
- *Package survey* (http://cran.r-project.org/web/packages/survey/index.html) enables a complex survey design to be specified. The resulting object can be used to estimate (Horvitz-Thompson) totals, means, ratios and quantiles for domains or the whole survey sample, and to apply regression models. The package also implements calibration. Variance estimation for means, totals and ratios can be done either by Taylor linearisation or replication (balanced repeated replication, jackknife, bootstrap or user-defined). Calibration, if applied, is also taken into account. The package (version 3.22) by and large includes modules that are similar to those in SAS and STATA;
- *Package EVER* (http://cran.r-project.org/web/packages/EVER/index.html) provides an estimation of variance for complex designs by delete-a-group jackknife replication for (Horvitz-Thompson) totals, means, absolute and relative frequency distributions, contingency tables, ratios, quantiles and regression coefficients, even for domains;
- *Package Laeken* (http://cran.r-project.org/web/packages/laeken/index.html) provides functions for estimating certain Laeken indicators (at-risk-of-poverty rate, income quintile share ratio, relative median risk-of-poverty gap, Gini coefficient), including their variance for domains and strata based on bootstrap resampling;
- *Package simFrame* (http://cran.r-project.org/web/packages/simFrame/index.html) allows comparison (user-defined) to be made of point and variance estimators in a simulation environment.

The biggest problem with R programs, i.e. capacity limitation with large data files, has been reduced by the recent release of R 3.0.0 (codename: 'Masked Marvel'). This release includes some major updates such as the introduction of big vectors to R, which eliminates some big

data restrictions in the core R engine by allowing R to make better use of the memory available on 64-bit systems. See for more information R-announce mailing list (https://stat.ethz.ch/pipermail/r-announce/2013/000561.html). There are also several packages available for dealing with the problem of capacity limitation (for example: *ff*, *bigmemory*, *biglm*). For more information see 'CRAN Task View: High-Performance and Parallel Computing with R' (http://cran.r-project.org/web/views/HighPerformanceComputing.html).

- **S-Plus**

S-Plus was actually a commercial predecessor of the R programming language. There is a library of programs for survey sampling, and these are currently by and large included in the R software modules.

- **MicrOsiris**

A general-purpose statistical software, Osiris used to be the grandfather when it came to including complex sampling design analysis in comprehensive statistical software. The current new development MicrOsiris is based on Osiris IV, though the sample survey analysis part is taken from IVEware (see below), and therefore contains complex sampling design analysis and imputation. MicrOsiris can be downloaded for free from the website (http://www.microsiris.com/).

Stand-alone software

- **SUDAAN**

SUDAAN is one of the oldest software products for complex sample data analysis. There are two versions of the software: a stand-alone and an SAS callable version (both licensed). The current version (10.0.1) calculates sampling variance with Taylor linearisation, delete-one jackknife, jackknife with multiple weights and balanced repeated replication (BRR). SUDAAN is capable of analysing multiply imputed data sets. More information can be viewed at http://www.rti.org/sudaan/.

- **WesVar**

WesVar (v.5.1) is stand-alone Windows software for survey data analysis. Its variance approximation is based on sample re-use techniques: delete-one, delete-two and delete-n jackknife, balanced repeated replication (BRR), and Fay's BRR. WesVar is capable of analysing multiply imputed data sets and can be used to create different types of weights. WesVar can be downloaded for free from the Westat website:
http://www.westat.com/westat/expertise/information_systems/wesvar/index.cfm

- **IVEware/Srcware**

IVEware is a free imputation and survey data software package developed at the University of Michigan. It exists as SAS procedures and stand-alone software (name Srcware), see http://www.isr.umich.edu/src/smp/ive/.

Variance approximation is based on Taylor linearisation. As the name suggests, it can handle different types of imputations properly.

- **Epi Info**

Epi Info is free stand-alone Windows software geared to aiding epidemiological research. It also contains some survey data analysis. Variance approximation is based on Taylor linearisation. The current (v. 3.5.1) or older versions can be downloaded from the website http://wwwn.cdc.gov/epiinfo/.

Special Sampling Variance Software for GREG

- **BASCULA**

BASCULA (v. 4) is a software tool for weighting and sampling variance and was originally developed for the data collection system BLAISE. BLAISE (and BASCULA) are licensed products. Variance estimation in BASCULA also accounts for calibration. Information can be viewed on the Statistics Netherlands homepage (http://www.cbs.nl/en-GB/menu/informatie/onderzoekers/blaise-software/blaise-voor-windows/productinformatie/bascula-info.htm).

- **CALJACK**

CALJACK is an SAS macro program for GREG variance estimation using jackknife. Further information on the current version and licence policy can be obtained from Statistics Canada (e-mail: Pierre.Lavallee@statcan.ca).

- **CLAN**

CLAN (v. 3.4.3) is a SAS macro program for sampling variance estimation. It can also calibrate weights and estimate variances with the GREG estimator. CLAN can be ordered for free from Statistics Sweden (e-mail: claes.andersson@scb.se).

- **g-Calib**

g-Calib is a calibration and sampling variance estimation macro program for the SPSS environment. Please contact Mr Camille Vanderhoeft for additional information by e-mail: camille.vanderhoeft@economie.fgov.be.

- **GENESEES**

GENESEES (Generalised Software for Sampling Estimates and Errors in Surveys) is a calibration and sampling variance estimation software package written as an SAS macro. Sampling variances are approximated with Taylor linearisation, and it also provides the GREG variance estimator. It can be requested for free from the ISTAT homepage (http://www.istat.it/it/strumenti/metodi-e-software/software/genesees (Italian version only).

- **GES**

Generalised Estimation System (GES) is an add-on software package for SAS. It can handle various sampling designs and also the GREG estimation. Further information on the current version and the licence policy can be obtained from Statistics Canada (e-mail: Laurie.Reedman@statcan.ca).

- **POULPE**

POULPE is an SAS macro program for sampling variance estimation. It is very exact on applied formulae but quite demanding to use. POULPE takes into account the impact of calibration on variance estimation. Further information on the current version and the licence policy can be obtained from INSEE (e-mail: Nathalie.Caron@INSEE.fr or Olivier.Sautory@INSEE.fr).

- **SEVANI**

The System for Estimation of Variance due to Non-response and Imputation (SEVANI) is an SAS-based prototype system developed by Statistics Canada (Beaumont and Mitchell, 2002). Variance estimation is based on the quasi-multi-phase framework. In this framework, a non-response model is required and an imputation model can also be used. Two types of non-response treatment methods can be dealt with: non-response weighting adjustment and imputation. If imputation is chosen, SEVANI requires one of the following four imputation methods to be used: deterministic linear regression imputation, random linear regression imputation, auxiliary value imputation or nearest-neighbour imputation.

- **ReGenesees**

ReGenesees (R evolved GENESEES) is a fully-fledged R system for design-based and model-assisted analysis of complex sample surveys. It handles multi-stage, stratified, clustered, unequally weighted survey designs. Sampling variance estimation for non-linear (smooth) estimators is done by Taylor series linearisation. Sampling variance estimation for multi-stage designs can be done under ultimate cluster approximation or by means of an actual multi-stage computation. Estimates, standard errors, confidence intervals and design effects are provided for: totals, means, absolute and relative frequency distributions (marginal or joint), ratios and quantiles (variance via the Woodruff method). ReGenesees also handles complex estimators, i.e. any user-defined estimator that can be expressed as an analytic function of Horvitz-Thompson or calibration estimators of totals or means, by automatically linearising them. All the above analyses can be carried out for arbitrary sub-populations. ReGenesees is available at JOINUP — the European Commission open source software repository https://joinup.ec.europa.eu/software/regenesees/description and at http://www.istat.it/it/strumenti/metodi-e-software/software/regenesees (Italian version only). Further information can be found at: http://www1.unece.org/stat/platform/display/msis/ReGenesees. For other information please contact ISTAT (e-mail: zardetto@istat.it).

Please also take a look at *Appendix 7.5*, which contains structured information on the appropriateness of some software tools to sampling designs and on their capacity to take into account different sources of variability in the overall variance estimation.

Summary

There are many software packages which can calculate variance estimates for linear and non-linear statistics under simple and complex sampling designs. For multi-stage sampling designs, most of them determine the overall sampling variance by ultimate cluster approximation. To deal with non-linear statistics, many software tools offer the option of using either Taylor linearisation or replication methods.

Available software tools are:

- comprehensive (commercial) statistical packages — SAS, SPSS, STATA;
- other general statistical packages – R, S-Plus, MicrOsiris;
- stand-alone software — SUDAAN, WesVar, IVEware/Srcware, Epi Info (SUDAAN, WesVar and IVEware are capable of analysing multiply imputed data sets);
- special sampling variance software for GREG — BASCULA, CALJACK, CLAN, g-Calib, GENESEES, GES, POULPE, SEVANI, ReGenesees (these are dedicated software tools for calibration; some R packages are also available). Most general-purpose software products do not contain any proper variance estimator that takes the impact of calibration into account in variance estimation.

3.6 Some examples of methods and tools used for variance estimation

Statistics Latvia uses two-stage sampling design for most household surveys in Latvia. Stratified systematic sampling of areas (PSUs) is used at the first stage; and PSUs are selected with probability proportional to size. Simple random sampling of dwellings (secondary sampling units — SSUs) is used at the second stage.

Self-made procedures[33] in R language have been used at Statistics Latvia for variance estimation since 2012. The procedures are based on the paper by Osier (2012). Variance can be estimated for population parameter estimates like total and the ratio of two totals. Procedures for other types of population parameters are under development. The process of variance estimation is split into four main steps: 1) extra variables are computed if domain estimation is considered; 2) non-linear parameters (for example, the ratio of two totals) are linearised; 3) the residuals of linear regression are computed if calibration of weights has been used to estimate population parameters; 4) variance estimates are computed using the 'ultimate cluster estimate' (Hansen, Hurwitz, & Madow, 1953, p. 257). Several estimates of precision measures are available in the output of the procedure – variance, absolute and relative standard error, the coefficient of variation, absolute and relative margin of error, confidence interval, and design effect.

SUDAAN software used to be used for variance estimation, until it was decided to completely discard the use of SUDAAN in the production of statistics. The main reasons were — SUDAAN is closed-source software (R is an open-source software with great customisation and integration possibilities); SUDAAN is a pay ware (R is a free ware); it is good for the efficiency of statistical production to minimise the number of software items used in production (R is used also in other production steps, e.g. calibration of weights).

ISTAT has been using both analytical and replication methods. GENESEES software, developed by ISTAT, is the main tool for variance estimation. Other references are: Moretti and Rinaldelli (2005), Rinaldelli (2006), Falorsi *et al* (2008).

The Luxembourg NSI (STATEC) has, in collaboration with CEPS/INSTEAD, been applying the bootstrap method to yield variance estimates for the main EU-SILC target indicators. For more information, see the EU-SILC national quality report for Luxembourg (available on CIRCA). In particular, this approach takes into account the impact of imputation on variance.

[33] Available at https://github.com/djhurio/vardpoor.

INSEE uses POULPE, which is an SAS macro-based application for variance estimation in complex designs. POULPE can deal with the following sampling plans (Ardilly and Osier, 2007):

- The one-phase multi-stage plans, with one of the following at each stage:
    - simple random sampling without replacement;
    - balanced simple random sampling;
    - sampling with unequal inclusion probabilities (probability proportional-to-size sampling);
    - systematic sampling with equal inclusion probabilities.
- The two-phase multi-stage plans, where the second phase is by either Poisson sampling or post-stratified sampling.
- The three-phase multi-stage plans, where the second phase is by post-stratified sampling and the third by Poisson sampling.

In particular, the impact of unit non-response on variance estimates can be included in the calculations by viewing a sample of respondents as the outcome of an additional phase of selection. POULPE can also take into account the impact of weight adjustments to external data sources.

POULPE was also used in Eurostat for EU-SILC (Osier, 2009) first wave, but was found to be too demanding as it required a lot of metadata for the design and calibration to be redone at Eurostat. The rotating design was further complicating the processing by adding a second, third and fourth phase to the sampling design, which was practically untraceable. Then, EU-SILC developed ad-hoc jackknife macros in SAS. The use of jackknife for EU-SILC was a feasibility study. See more details on the use of jackknife macros in Section 4.1. The next step was to test methods other than jackknife, i.e. bootstrap and linearisation. Comparative experiments were carried out on a limited number of countries, and the results of different methods are similar. The present choice is to work with linearisation (ultimate cluster approximation), which was discussed at the Net-SILC2 workshop on accuracy and validated by the SILC Working Group. This approach yields acceptable results given the administrative considerations.

3.7 Sampling over time and sample coordination

Martin Axelson (Statistics Sweden) and Ioannis Nikolaidis (EL.STAT)

It is not uncommon for national statistical institutes to conduct continuing surveys, in the sense that the same population is sampled repeatedly over time. Such surveys are typically conducted for one or more of the following reasons (e.g. Duncan and Kalton, 1987):

- to provide estimates of parameters at specific time points;
- to provide estimates of parameters defined as averages over a period of time;
- to provide estimates of net change between two time points, i.e. to estimate the difference, the ratio, or some other measure of change, between parameters at different time points;
- to provide estimates of gross change, i.e. aggregates of change at the element level between time points. Gross change is often referred to as flows when the variable under study is categorical (see Section 3.7.3).

Considering the number of surveys done by statistical institutes every year, it is not surprising that many different sampling methods have been developed to address the above objectives. Although different, all sample surveys that use probability sampling methods for sampling over time belong to one of the following two classes:

- surveys in which samples selected at different time points are statistically independent;
- surveys in which samples selected at different time points are statistically dependent.

Surveys belonging to the second class share the common feature that samples are coordinated *by design* over time. Typically, samples over time are positively coordinated, in the sense that a part of the sample selected at time t is retained in the sample at time point $t+1$, though it has to be noted that negative coordination may also take place. Examples of surveys performed by EU countries in which positive sample coordination is used are the LFS, the ICT and the EU-SILC.

Repeated and periodic surveys have many similarities to multinational and domain comparisons, but designs over time usually differ from the others in two special respects. Firstly, they are designed for and selected from the same population, which tends to retain its characteristics and structures; and secondly, those similarities permit and often encourage designs of overlapping samples, where covariance tends to reduce the variance of comparisons (Kish, 1994). Surveys are sometimes repeated at irregular intervals. However, periodic surveys, repeated at regular intervals, are becoming more common. The periods may be long, as for ten-yearly censuses or for annual surveys, or they may be short, as for the quarterly or monthly surveys (e.g. the LFS in many countries).

Sample sizes and sampling methods may vary for different waves, except that they need to be fixed for the sample overlaps (of the same ultimate sampling units, but also of the same areas being used as primary sampling units). *However, it is recommended that survey methods be kept similar for comparisons between waves. The variances of changes and sums in periodic surveys should take into account any overlapping correlations;* the positive correlation between periods increases the variance of the sum of estimates over time but reduces the variance of changes.

There is abundant literature on the design and analysis of surveys over time and sample coordination (e.g. Duncan and Kalton, 1987; Binder, 1998; Kish, 1998; Nordberg, 2000; and Steel and McLaren, 2009). Terms like repeated surveys, panel surveys, rotating panel surveys and split panel surveys appear frequently in the survey literature when positive sample coordination is being discussed. Unfortunately, not all authors assign the same meaning to the terms, a fact that complicates an already complex issue even further.

Sampling over time makes it possible to analyse changes in variables of interest. *Apart from design issues, we should also consider the frequency of sampling, the spread of surveyed units over time and the application of overlapping or non-overlapping samples over time.* Surveys using sampling over time can be classified into:

- repeated surveys;
- longitudinal surveys:
    - panel surveys (fixed panel, fixed panel plus ‘births’);
    - repeated panel surveys;
    - rotating panel surveys;
    - split panel surveys;

- rolling samples.

The key factors which influence sampling design over time are the estimates of the main variables to be produced and the type of analyses to be performed. Of course, the interaction between sampling over time and design features, such as stratification and cluster sampling, also need to be decided (Steel and McLaren, 2009).

The ***repeated surveys*** serve for comparison and for the production of time series. The analysis and creation of such time series involves seasonal adjustment, business cycle and time trend estimation. High-quality surveys are based on probability sampling methods that yield estimates of population characteristics and make it possible to analyse relationships between variables. Sampling frequency depends on the purpose of the survey. A repeated survey enables population change estimation and cross-sectional estimation. *It is recommended that the population frame be updated to incorporate population changes as soon as possible. The sample should be updated to give the new units a chance of selection and to remove defunct units that may affect standard errors.* In a repeated survey there is no need for any sample overlap on different occasions.

Repeated surveys may be irregular, but ***periodic surveys*** are repeated at regular intervals and are becoming more common. The periods may be long, as for annual surveys (e.g. Structural Business Surveys), or they may be short, as for quarterly or monthly surveys (e.g. survey for compiling monthly industrial production indices, survey for compiling quarterly turnover indices, etc.).

A **longitudinal survey** is a survey that collects data from the same sample elements on multiple occasions over time. An initial sample is selected, and at each occasion or wave an attempt is made to include units of the initial sample. A longitudinal survey may be used to provide estimates of changes at aggregate levels, but these estimates refer to the population at the time of the initial sample selection unless the sample has been updated to make it representative of the current population. The main purpose of a longitudinal survey is to yield estimates and analyses of changes at the unit level (gross changes).

Five broad types of longitudinal survey designs can be identified (Lynn, 2009):

- A ***panel survey*** is a kind of longitudinal survey in which an initial sample is selected and interviewed for several time periods. Panel surveys are needed to detect dynamics of gross (micro) changes of units (individuals, households etc.).

  A panel survey may be fixed panel or fixed panel plus 'births'. For a ***fixed panel survey***, statistical data are collected from the same units at multiple time periods. These panel surveys have problems of attrition and mortality, since no additions to the sample are made after initial sample selection. On the other hand, in the ***fixed panel plus 'births' surveys,*** at each wave of data collection, a sample of units which are born since the previous wave is added. This type of panel survey may be used when there are significant 'births' in the population during the life of the survey and there is a desire to represent the cross-sectional population at the time of each wave as well as the longitudinal population of wave 1 'survivors'. Most household panel surveys have this design.

- A ***repeated panel*** takes the form of a series of panel surveys which may or may not overlap in time. Typically, each panel is designed to represent an equivalent population, i.e. the same population definition applied at a different point in time.

- In a ***rotating panel survey***, predetermined proportions of sample units are replaced on each occasion. This type of survey enables gradual changes of sampling units and

partial overlapping between successive fieldwork occasions. Typically, each unit will remain in the sample for the same number of waves. Rotating panel designs are often used when the main objectives are cross-sectional estimates and short-term estimates of net and gross change. Labour Force Surveys have a rotating panel design in many countries. It is often applied at dwelling level, which means that people and households are not followed when they leave a selected dwelling. People and households moving into a selected dwelling are included in the survey. This approach is suitable when the main objective is to provide unbiased aggregate estimates. However, any overlapping sample can also be used to analyse change at the micro-level. The resulting sample of individuals for whom longitudinal data are available may be biased due to people who move permanently out or are temporarily absent from the selected households. The total time period and the time interval between observations are determined by the rotation pattern.

- An alternative to a rotating panel survey is a ***split panel survey***, which involves a panel survey supplemented for each reference period by an independent sample (Kish, 1987). This approach permits longitudinal analysis from the panel survey for more periods than would be possible in a rotating panel design. It also enables cross-sectional estimates to be obtained from the entire sample (Kish, 1998).

This is only a broad typology which does not fully describe the range of possible designs. For example, each panel in a repeated panel design may or may not include an additional regular sample of births.

It is important to distinguish between **longitudinal surveys** and **longitudinal data**. Longitudinal surveys are a source of longitudinal data, as the resultant data include items that refer to different points in time. But there are other ways of obtaining longitudinal data, including diary methods and the use of retrospective recall within a single survey instrument (Lynn, 2009). The OECD Glossary of Statistical Terms (OECD) refers to **panel data** as synonymous to **longitudinal data**.

When estimates are focused on population totals, an independent sample may be used for each reference period, which is often the case when the interval between surveys is quite large (e.g. annual surveys or surveys every two years). In this case, a non-overlapping design exists, according to which sampling units are changed deliberately for each time period. For monthly or quarterly surveys, the sample is often designed with considerable overlap between successive periods. Some overlaps may be built into the samples to gain advantages from positive correlations between periods. Standard errors of the estimate of changes over time are minimised by using complete overlap of samples (Kish, 1965). Respondent load, attrition, non-response and generally declining response rate usually lead to some degree of replacement or rotation of the sample from one period to the next (Steel and McLaren, 2009).

Rolling samples are samples that have been deliberately designed to cover (roll over) the entire population in several or even many periodic surveys, and are taken by moving to different primary sampling units (PSUs) in each wave (Kish, 1990). One example of a rolling sample of households is where the 52 weekly samples are designed to cover the whole country instead of being confined within the same sample of PSUs.

To sum up, rolling samples are a separate category from longitudinal surveys. While rolling samples are taken by moving to different PSUs each wave, longitudinal surveys include units of the previous sample in the new sample. On the other hand, repeated surveys may or may not be rolling samples and may or may not be longitudinal surveys.

Any discussion of design and analysis of surveys over time would most likely benefit from having a clear, unambiguous terminology. There are many important aspects that need to be considered. *In particular, a good terminology should clearly address the fact that different sampling designs allow for sample coordination at different levels.* For example, under one-stage element sampling, sample coordination may be only at the element level. However, if one-stage cluster sampling is used to select the initial sample, then for the following time point it is possible to achieve sample coordination at (a) the cluster level, (b) the element level or (c) a combination of (a) and (b). Although (a) — (c) all result in sample coordination, they require different mathematical treatment of the data to obtain valid point and variance estimates.

An inventory of the rotation schemes used in the LFS, ICT and EU-SILC household surveys shows that:

- The EU-SILC is generally characterised by a rotation pattern with a periodicity of one year. The sample at a given year consists of four rotation groups which have been in the survey for 1-4 years. Any particular rotation group remains in the survey for four years; each year, one of the four rotation groups from the previous year is dropped and a new one is added. Between year *t* and *t+1* the sample overlap is 75 %; the overlap between year *t* and year *t+2* is 50 %; it is reduced to 25 % from year *t* to year *t+3*, and to zero for longer intervals. The rotation seems to be done either within the PSUs selected in the previous survey waves or is totally independent of the PSUs of previous waves.
- The LFS is characterised by a rotation pattern with a periodicity of one quarter. There are six rotation schemes, which means a wider diversity of sampling designs used by countries and a more complex design compared to the EU-SILC. The panel component is meant to introduce more efficiency into measuring changes in indicators. Increasing the quarter-on-quarter overlap of the sample would enhance the precision of estimates of changes between consecutive quarters and estimates of quarter-on-quarter flows. In general, rotation seems to affect the ultimate sampling units within the previously selected PSUs.
- For ICT, it seems that there are ten countries whose sample is affected by rotation procedures:[34] the ICT sample is either embedded in the LFS, embedded in the EU-SILC or uses a panel that is independent of other surveys.

Statistics estimated from the EU-SILC, and in particular the indicators on social exclusion and poverty, are usually estimated for cross-sectional samples and published annually. Simply comparing point estimates tables might lead to an over-interpretation of the data because any observed changes might be due only to sampling variances. It is therefore useful to provide variance estimates for changes in (cross-sectional) point estimates. This requires accounting for covariance between cross-sectional estimates among consecutive measuring points — something that is introduced through the rotating sampling scheme. Furthermore, if estimates are non-linear, which most indicators on social exclusion and poverty are, then linearisation techniques have to be used for variance estimation (Münnich *et al*, 2011b).

The consideration of covariance in surveys based on panels is a research topic. The project 'Advanced Methodology for European Laeken Indicators',[35] which was funded under the Seventh Framework Programme of the European Commission, worked on advanced recommendations for variance estimators of the Laeken indicators. They take into account not

[34] According to the quality reports for 2008.

[35] See http://ameli.surveystatistics.net.

only the various countries' different sampling schemes but also practical problems, such as data peculiarities which may lead to the inappropriate use of indicator outcomes. This would enable better use of indicators in policy making.

Although the consideration of covariance in surveys based on panels is a research topic, there is a strong need to improve variance estimation under rotation procedures and to get countries to provide the best possible sampling variance approximation.

3.7.1 Variance estimation for annual averages

Ioannis Nikolaidis (EL.STAT)

One of the goals of a quarterly survey is to calculate annual averages, e.g. the total unemployed persons over one year. If $\hat{Y}_q$ is the estimated total of unemployed at quarter $q(q=1\ldots4)$, the annual total is given by:

$$\hat{Y}_{tot} = \frac{1}{4}\cdot\sum_{q=1}^{4}\hat{Y}_q \quad . \quad (3.7.1.1)$$

The variance of $\hat{Y}_{tot}$ is given by:

$$V\left(\hat{Y}_{tot}\right)=\left(\frac{1}{4}\right)^2\cdot\left[\sum_{q=1}^{4}V\left(\hat{Y}_q\right)+2\sum_{q=1}^{3}\sum_{q'>q}Cov\left(\hat{Y}_q,\hat{Y}_{q'}\right)\right]. \quad (3.7.1.2)$$

A rotation group is often replaced by a group selected from the same geographical 'sector', thus causing covariance as well (Place, 2008). For example, rotation may take place within PSUs, as in the case of the Australian Labour Force Survey (cf. Steel and McLaren, 2009). In order to estimate the difference $D_{q,q+1}$ of parameters of variable y between two consecutive waves q and $q+1$, we have to take into account not only the covariance caused by the 'panel' component (that is, the sub-sample which remains between q and $q+1$) but also that caused by the renewal of any rotation group.

An additional complication arises from the fact that rotation groups are often created by dividing the first wave sample into a fixed number of sub-groups of equal size. These sub-groups are generally not independent of each other. For example, assuming that a simple random sample is selected at first wave and then divided into two rotation groups **1** and **2** by taking a simple random sub-sample from the whole sample, then the covariance between the sample means $\bar{y}_1$ and $\bar{y}_2$ of a variable y over the rotation groups **1** and **2** can be written as (Tam, 1984; Ardilly and Tillé, 2005):

$$Cov\left(\bar{y}_1,\bar{y}_2\right)=-\frac{S_y^2}{N}, \quad (3.7.1.3)$$

where N is the population size and S_y^2 is the variance of the study variable y over the population (see *Appendix 7.1*). More generally, if another variable z is measured on the rotation group **2**, we obtain:

$$Cov(\bar{y}_1, \bar{z}_2) = -\frac{S_{yz}}{N}, \quad (3.7.1.4)$$

where S_{yz} is the covariance between y and z over the total population:

$$S_{yz} = \frac{1}{N-1}\sum_{i \in U}(y_i - \bar{Y})(z_i - \bar{Z}). \quad (3.7.1.5)$$

Thus, if we ignore the covariance term (3.7.1.3) or (3.7.1.4) in variance calculations, the variance of the sample means $\bar{y}$ over the whole sample **1** + **2** will be overestimated (Steel and McLaren, 2009).

However, some authors (Kish, 1965) assume that the rotation groups are independent because they are disjoint. Also, in practice it is usually assumed that the different rotation groups are nearly independent, and covariance is computed through correlation on the overlapping samples with the additional hypothesis that design effects are roughly the same for the three variables y_q, y_{q+1} and $y_{q+1} - y_q$ (Place, 2008). Thus, ***a simple method*** for calculating the covariance $Cov(\hat{Y}_q, \hat{Y}_{q'})$ consists of assuming that the non-overlapping parts between q and q' are independent. Hence, we obtain (Steel and McLaren, 2009; Salonen, 2008):

$$\hat{Cov}(\hat{Y}_q, \hat{Y}_{q'}) \approx o_{q,q'} \cdot \sqrt{\hat{V}(\hat{Y}_q)} \cdot \sqrt{\hat{V}(\hat{Y}_{q'})} \cdot \hat{\rho}(\hat{Y}_q, \hat{Y}_{q'}), \quad (3.7.1.6)$$

where

$\hat{\rho}(\hat{Y}_q, \hat{Y}_{q'})$ is the correlation coefficient between $\hat{Y}_q$ and $\hat{Y}_{q'}$ [36],

$o_{q,q'}$ is the proportion of overlapping units between q and q'.

For an annual average estimate $\hat{Y}_{tot}$, ***an alternative method*** of variance estimation is the Inflation Coefficient approximation (Salonen, 2008). The Inflation Coefficient can be defined by $IC = \sqrt{\frac{resp_T}{resp_R}}$, (3.7.1.7)

where

$resp_T$ = theoretical respondent group over four quarters;

$resp_R$ = real respondent group over four quarters (influence of overlapping).

The variance of the average over four quarters can be given as

$$\hat{V}(\hat{Y}_{tot\_IC}) = IC^2 \cdot \hat{V}(\hat{Y}_{tot\_GREG}), \quad (3.7.1.8)$$

where $\hat{V}(\hat{Y}_{tot\_GREG})$ is the variance of the generalised regression estimator (GREG) over four quarters, estimated with the assumption that the quarters are independent (there is no sample overlap between quarters). See Section 3.4 for estimating variance for the GREG estimator.

[36] In general, $\rho(\hat{Y}_q, \hat{Y}_{q'})$ is not the same as the correlation at the element level between the study variable y at time q and $q+1$.

3.7.2 Variance estimation for estimators of net change

Martin Axelson (Statistics Sweden) and Ioannis Nikolaidis (EL.STAT)

This section discusses variance estimation for estimators of net change, with some attention being given to problems encountered when samples are coordinated over time.

Estimation of net change

A finite population parameter is typically defined in terms of a numerical expression in order to summarise values of some study variable(s) for elements in a target population. *Given that the composition of a target population typically varies over time, just like the values of study variable(s) at element level, the time dimension should be taken into account when defining finite population parameters. Clear, unambiguous definitions of the reference period, at both object level and variable level, are therefore necessary so that parameters are uniquely defined. This becomes even more important when discussing estimation of change over time.*

Let U_t denote a target population at time t. That is, U_t consists of all elements of the same type, which at time t qualify for inclusion in the target population according to some predefined criteria. Let y denote a study variable of interest, and let $y_{t,k}$ denote the value of y which at time t is associated with element k. The parameter of interest at time t is $Y_t = \sum_{k \in U_t} y_{t,k}$. At time point $t+1$, the parameter of interest is $Y_{t+1} = \sum_{k \in U_{t+1}} y_{t+1,k}$. The greater part of the discussion in this section will focus on estimating $D = Y_{t+1} - Y_t$, as this is the measure of net change that is perhaps most often encountered in practice. The following, somewhat simplified, setup will be considered:

- Time point t:
  - A probability sample s_t of elements is selected from U_t, according to the sampling design $p_t(\cdot \mid s_{t-1})$. The notation $p_t(\cdot \mid s_{t-1})$ is used to indicate that the choice of sampling design at time t may be conditioned by the sample selected at time $t-1$.
  - An estimator for Y_t is given by $\hat{Y}_t$. The estimator $\hat{Y}_t$:
    (a) appropriately reflects the sampling design used to select s_t;
    (b) is non-response-adjusted;
    (c) may incorporate auxiliary information.
- Time point $t+1$:
  - A probability sample s_{t+1} of elements is selected from U_{t+1}, according to the sampling design $p_{t+1}(\cdot \mid s_t)$.
  - An estimator for Y_{t+1} is given by $\hat{Y}_{t+1}$. The estimator $\hat{Y}_{t+1}$:
    (a) appropriately reflects the sampling design used to select s_{t+1};
    (b) is non-response-adjusted;
    (c) may incorporate auxiliary information.

Let us consider $\hat{D} = \hat{Y}_{t+1} - \hat{Y}_t$ as an estimator of $D = Y_{t+1} - Y_t$. If $\hat{Y}_t$ and $\hat{Y}_{t+1}$ are approximately unbiased estimators for Y_t and Y_{t+1}, then $\hat{D}$ is an approximately unbiased estimator for D.

Variance estimation for estimators of net change

The variance of $\hat{D}$ is given by:

$$V(\hat{D}) = V(\hat{Y}_t) + V(\hat{Y}_{t+1}) - 2Cov(\hat{Y}_t, \hat{Y}_{t+1}), \quad (3.7.2.1)$$

where $V(\hat{Y}_t)$ and $V(\hat{Y}_{t+1})$ denote the unconditional variances of $\hat{Y}_t$ and $\hat{Y}_{t+1}$ respectively, and $Cov(\hat{Y}_t, \hat{Y}_{t+1})$ denotes the unconditional covariance between $\hat{Y}_t$ and $\hat{Y}_{t+1}$. It is important to realise that under the setup considered above, $V(\hat{Y}_t)$ depends not only on the sampling design $p_t(\cdot \mid s_{t-1})$ but also on the sampling designs $p_{t-h}(\cdot \mid s_{t-(h+1)})$, $h = 1,2,...$ Analogously, $V(\hat{Y}_{t+1})$ and $Cov(\hat{Y}_t, \hat{Y}_{t+1})$ depend not only on the sampling design $p_{t+1}(\cdot \mid s_t)$ but also on the sampling designs $p_{t-h}(\cdot \mid s_{t-(h+1)})$, $h = 0,1,2,...$. An alternative expression for $V(\hat{D})$ is given by:

$$V(\hat{D}) = V(\hat{Y}_t) + V(\hat{Y}_{t+1}) - 2[V(\hat{Y}_t)V(\hat{Y}_{t+1})]^{1/2}\rho(\hat{Y}_t, \hat{Y}_{t+1}), \quad (3.7.2.2)$$

where $\rho(\hat{Y}_t, \hat{Y}_{t+1})$ denotes the unconditional correlation between $\hat{Y}_t$ and $\hat{Y}_{t+1}$, i.e.

$$\rho(\hat{Y}_t, \hat{Y}_{t+1}) = \frac{Cov(\hat{Y}_t, \hat{Y}_{t+1})}{\sqrt{V(\hat{Y}_t)V(\hat{Y}_{t+1})}}. \quad (3.7.2.3)$$

Clearly, when s_t and s_{t+1} are statistically independent, i.e. when $p_{t+1}(s_{t+1} \mid s_t) = p_{t+1}(s_{t+1})$ for all possible s_t, then the same is true for $\hat{Y}_t$ and $\hat{Y}_{t+1}$, and therefore the variance is given by:

$$V(\hat{D}) = V(\hat{Y}_t) + V(\hat{Y}_{t+1}). \quad (3.7.2.4)$$

Typically, s_t and s_{t+1} are positively coordinated. One major reason for this is of course that positive sample coordination implies that a large part of the elements included in s_t will be retained in s_{t+1}, which in turn implies that $\hat{Y}_t$ and $\hat{Y}_{t+1}$ has to be positively correlated. When this is the case, it follows that:

$$V(\hat{D}) = V(\hat{Y}_t) + V(\hat{Y}_{t+1}) - 2Cov(\hat{Y}_t, \hat{Y}_{t+1}) \leq V(\hat{Y}_t) + V(\hat{Y}_{t+1}). \quad (3.7.2.5)$$

That is, under positive sample coordination such that $Cov(\hat{Y}_t, \hat{Y}_{t+1}) > 0$, $\hat{D}$ is more efficient than it would be under independent sampling at time points t and $t+1$. To see the extent of the efficiency gain when $\hat{Y}_t$ and $\hat{Y}_{t+1}$ are positively correlated, consider a situation in which $V(\hat{Y}_t) \approx V(\hat{Y}_{t+1})$. Then it follows that:

$$V(\hat{Y}_t) + V(\hat{Y}_{t+1}) \approx 2V(\hat{Y}_t) \quad (3.7.2.6)$$

and

$$V(\hat{Y}_t)+V(\hat{Y}_{t+1})-2Cov(\hat{Y}_t,\hat{Y}_{t+1})\approx 2V(\hat{Y}_t)[1-\rho(\hat{Y}_t,\hat{Y}_{t+1})]\ . \quad (3.7.2.7)$$

Therefore,

$$\frac{V(\hat{Y}_t)+V(\hat{Y}_{t+1})-2Cov(\hat{Y}_t,\hat{Y}_{t+1})}{V(\hat{Y}_t)+V(\hat{Y}_{t+1})}\approx 1-\rho(\hat{Y}_t,\hat{Y}_{t+1})\cdot \quad (3.7.2.8)$$

Clearly, a considerable gain in efficiency is achieved if $\hat{Y}_t$ and $\hat{Y}_{t+1}$ are positively correlated.

Under independence, an estimator for the variance of $\hat{D}$ is given by

$$\hat{V}(\hat{D})=\hat{V}(\hat{Y}_t)+\hat{V}(\hat{Y}_{t+1})\ . \quad (3.7.2.9)$$

Deriving appropriate estimators for $V(\hat{Y}_t)$ and $V(\hat{Y}_{t+1})$ under independence is normally not a major problem. However, when $\hat{Y}_t$ and $\hat{Y}_{t+1}$ are statistically dependent due to sample coordination, then estimating $V(\hat{Y}_t)$ and $V(\hat{Y}_{t+1})$ may become far from trivial. Nevertheless, in what follows it is assumed that working estimators for $V(\hat{Y}_t)$ and $V(\hat{Y}_{t+1})$ are readily available. Hence, we need to find a working estimator for the covariance term $Cov(\hat{Y}_t,\hat{Y}_{t+1})$.

Under certain choices of sampling designs $p_{t+1-h}(\cdot\,|\,s_{t-h})$ $h=0,1,2,\ldots$, estimation of $Cov(\hat{Y}_t,\hat{Y}_{t+1})$ is straightforward, but in general covariance estimation under sample coordination is not an easy task. Unfortunately, there is no universally applicable method which can be used to solve this problem. However, Berger (2004) proposes an ingenious approach which can be used under a broad class of sampling designs used for positive sample coordination at the element level. The approach discussed by Berger (2004) is to use:

$$\hat{Cov}(\hat{Y}_t,\hat{Y}_{t+1})=\sqrt{\hat{V}(\hat{Y}_t)\hat{V}(\hat{Y}_{t+1})}\,\hat{\rho}(\hat{Y}_t,\hat{Y}_{t+1})\ , \quad (3.7.2.10)$$

where $\hat{\rho}(\hat{Y}_t,\hat{Y}_{t+1})$ is an estimator for $\rho(\hat{Y}_t,\hat{Y}_{t+1})$, as an estimator for $Cov(\hat{Y}_t,\hat{Y}_{t+1})$. This approach has the advantage that using

$$\hat{V}[(\hat{Y}_t,\hat{Y}_{t+1})']=\begin{bmatrix}\hat{V}(\hat{Y}_t) & \sqrt{\hat{V}(\hat{Y}_t)\hat{V}(\hat{Y}_{t+1})}\,\hat{\rho}(\hat{Y}_t,\hat{Y}_{t+1})\\ \sqrt{\hat{V}(\hat{Y}_t)\hat{V}(\hat{Y}_{t+1})}\,\hat{\rho}(\hat{Y}_t,\hat{Y}_{t+1}) & \hat{V}(\hat{Y}_{t+1})\end{bmatrix} \quad (3.7.2.11)$$

guarantees non-negative estimates of the covariance matrix.

$$V[(\hat{Y}_t,\hat{Y}_{t+1})']=\begin{bmatrix}V(\hat{Y}_t) & Cov(\hat{Y}_t,\hat{Y}_{t+1}))\\ Cov(\hat{Y}_t,\hat{Y}_{t+1}) & V(\hat{Y}_{t+1})\end{bmatrix} \quad (3.7.2.12)$$

That is, any estimate produced by $\hat{V}[(\hat{Y}_t,\hat{Y}_{t+1})']$ is a positive semi-definite matrix.

However, great caution should be exercised when this approach is used. For $\hat{Cov}(\hat{Y}_t,\hat{Y}_{t+1})$ *to be a valid estimator,* $\hat{\rho}(\hat{Y}_t,\hat{Y}_{t+1})$ *should be a valid estimator for* $\rho(\hat{Y}_t,\hat{Y}_{t+1})$. *In particular, the estimator for* $\rho(\hat{Y}_t,\hat{Y}_{t+1})$ *should be such that it properly reflects the correlation between* $\hat{Y}_t$ *and* $\hat{Y}_{t+1}$. *For example,* $\rho(\hat{Y}_t,\hat{Y}_{t+1})$ *is not generally the same as the correlation at the element*

level between the study variable y *at time* t *and* $t+1$*. Hence, using an estimate of the correlation at the element level between the study variable* y *at time* t *and* $t+1$ *instead of* $\hat{\rho}(\hat{Y}_t, \hat{Y}_{t+1})$ *in the above formula will not produce a valid general covariance estimator. In practice,* $\rho(\hat{Y}_t, \hat{Y}_{t+1})$ *is sometimes estimated from the sample overlap, i.e, from* $s_t \cap s_{t+1}$. Uncritical use of such an approach may very well result in an estimator for $\rho(\hat{Y}_t, \hat{Y}_{t+1})$ which has a positive bias. When this is the case, the estimator

$$\hat{V}(\hat{D}) = \hat{V}(\hat{Y}_t) + \hat{V}(\hat{Y}_{t+1}) - 2[\hat{V}(\hat{Y}_t)\hat{V}(\hat{Y}_{t+1})]^{1/2}\,\hat{\rho}(\hat{Y}_t, \hat{Y}_{t+1}) \qquad (3.7.2.13)$$

will have a negative bias. Berger (2004) argues that for the large positive correlations $\rho(\hat{Y}_t, \hat{Y}_{t+1})$ that are often encountered in practice, even a small positive bias in $\hat{\rho}(\hat{Y}_t, \hat{Y}_{t+1})$ may lead to a severe negative bias in $\hat{V}(\hat{D})$. Hence, if the approach outlined above is to be used, special care should be taken to ensure that $\hat{\rho}(\hat{Y}_t, \hat{Y}_{t+1})$ is a valid estimator for $\rho(\hat{Y}_t, \hat{Y}_{t+1})$.

Note also that the above discussion can be extended to cover more complex measures of net change. Consider a more general setting, in which the study variable has a vector value rather than a scalar one, and the parameter of interest at each time point is defined as a function of the population total of the study variable. Let $\mathrm{y} = (y_1, ..., y_Q)'$ denote the study variable of interest, let $\theta_t = f(\mathrm{Y}_t)$ (where f is a rational function and $\mathrm{Y}_t = \sum_{k \in U_t} \mathrm{y}_{t,k}$ denotes the parameter of interest at time t) and let $\theta_{t+1} = f(\mathrm{Y}_{t+1})$. Then for any measure of change defined as $g(\theta_t, \theta_{t+1})$, where g is a rational function, an estimator is given by $g(\hat{\theta}_t, \hat{\theta}_{t+1})$, where $\hat{\theta}_t = f(\hat{\mathrm{Y}}_t)$ and $\hat{\theta}_{t+1} = f(\hat{\mathrm{Y}}_{t+1})$. Since f and g are both rational functions, it follows that $g(\hat{\theta}_t, \hat{\theta}_{t+1}) = g(f(\hat{Y}_t), f(\hat{Y}_{t+1})) = h(\hat{Y})$, where h is in itself a rational function and $\hat{\mathrm{Y}} = (\hat{\mathrm{Y}}_t', \hat{\mathrm{Y}}_{t+1}')'$. Hence, using the first-order Taylor linearisation (e.g. Andersson and Nordberg, 1994), the variance of $g(\hat{\theta}_t, \hat{\theta}_{t+1})$ may be approximated by an expression to give:

$$V[g(\hat{\theta}_t, \hat{\theta}_{t+1})] \approx \nabla(\mathrm{Y})' V(\hat{\mathrm{Y}}) \nabla(\mathrm{Y}) \quad , \qquad (3.7.2.14)$$

where $\nabla(\mathrm{Y})$ is the gradient of $h(\hat{\mathrm{Y}})$ at $\mathrm{Y} = (\mathrm{Y}_t', \mathrm{Y}_{t+1}')'$ and $V(\hat{\mathrm{Y}})$ is the covariance matrix of $\hat{\mathrm{Y}}$. If $\hat{V}(\hat{\mathrm{Y}})$ is an estimator for $V(\hat{\mathrm{Y}})$, then an estimator for $V[g(\hat{\theta}_t, \hat{\theta}_{t+1})]$ is given by

$$\hat{V}[g(\hat{\theta}_t, \hat{\theta}_{t+1})] \approx \nabla(\hat{\mathrm{Y}})' \hat{V}(\hat{\mathrm{Y}}) \nabla(\hat{\mathrm{Y}}) \quad , \qquad (3.7.2.15)$$

where $\nabla(\hat{\mathrm{Y}})$ is an estimator for $\nabla(\mathrm{Y})$.

Berger and Priam (2010) propose an estimator for the covariance between estimates at two different waves. This estimator is valid under stratified, two-stage sampling designs, which involve unequal probabilities and non-response. The proposed estimator can also handle a wide range of measures of change. Berger and Priam (2010) show how their proposed estimator can be used to estimate correlation between complex estimators of change. The estimator is based on a multivariate linear regression approach to estimate covariance. This is not a model-based estimator, as it is valid even if the model does not fit the data. Berger and Priam (2010) show that the regression approach gives a design-consistent estimator for the correlation when finite population corrections are negligible. The multivariate regression approach is simple to apply as it can be easily implemented in most statistical software tools.

Not surprisingly, the estimator proposed by Berger and Priam (2010) is equivalent to Tam's (1984) estimator under simple random sampling design. However, for more complex designs, the proposed estimator is more accurate than (3.7.1.3), as (3.7.1.3) is based on the assumption that the sample is selected with simple random sampling.

The following section gives some practical guidance on how to estimate variance of the difference of estimates from two different time points, for **stratified multi-stage sampling**. If $\hat{D} = \hat{Y}_2 - \hat{Y}_1$ is the net change of estimates $\hat{Y}_2$ and $\hat{Y}_1$ of a variable y, between two time periods, then the problem is how to estimate variance of $\hat{D}$.

Large-scale surveys often employ stratified multi-stage designs with a large number of strata H and relatively few primary sampling units (clusters: e.g. areas, blocks, municipalities or local government areas) are selected within each stratum h.

At time period t, let n_t be the number of selected PSUs and let w_{htij} (>0) stand for a survey weight attached to a sample's ultimate element j ($j = 1,...,m_{hti}$), belonging to the selected cluster (PSU) i and stratum h. Then w_{htij} is the product of three factors: a) the inverse of the inclusion probabilities of the ultimate sampling units, b) the inverse of the response rate r_{ht} in the stratum h and c) a factor k_{htij}, which makes weighted sample estimates conform to external total values (values from known totals from censuses, administrative sources, population projections, etc.).

At time period t, let y_{htij} be the value of the characteristic y of the ultimate unit j, belonging to the hi primary sampling unit (cluster). Moreover, Y_t stands for the population total at time point t. We assume that the strata are response homogeneity groups. The form of the estimator on the basis of the two-stage design is thus:

$$\hat{Y}_t = \sum_{h=1}^{H} \sum_{i=1}^{n_{ht}} \sum_{j=1}^{m_{hti}} w_{htij} \cdot y_{htij} \text{ . (3.7.2.16)}$$

Let s_t be the sampled ultimate units at time t. Let us also consider estimating the change in totals between two quarters. Then the basic estimator will be given by:

$$\hat{Y}_2 - \hat{Y}_1 = \sum_{(hij) \in s_2} w_{h2ij} \cdot y_{h2ij} - \sum_{(hij) \in s_1} w_{h1ij} \cdot y_{h1ij} \text{ , (3.7.2.17)}$$

where subscripts 1 and 2 refer to the two quarters and the variance of this change is given by:

$$V(\hat{Y}_2 - \hat{Y}_1) = V(\hat{Y}_2) + V(\hat{Y}_1) - 2 \cdot Cov\left(\sum_h \sum_{i \in s_h^*} \sum_j w_{h2ij} \cdot y_{h2ij}, \sum_h \sum_{i \in s_h^*} \sum_j w_{h1ij} \cdot y_{h1ij}\right) \text{, (3.7.2.18)}$$

where s_h^* is the sub-sample of clusters in stratum h for which data are available in both quarters. Essentially, s_h^* denotes n_c PSUs, which are common in both quarters. So, the third component is calculated on the basis of the dwelling data taken from the common PSUs

surveyed in both quarters. The first two components are obtained in a simple manner from the variance estimates of level at each quarter, as follows (Rao, 1988):

$$\hat{V}\left(\hat{Y}_t\right)=\sum_{h=1}^{H}\frac{n_{ht}}{n_{ht}-1}\sum_{i=1}^{n_{ht}}\left(\hat{Y}_{hti}-\frac{\sum_{i=1}^{n_{ht}}\hat{Y}_{hti}}{n_{ht}}\right)^2 \quad (3.7.2.19) \text{ or}$$

$$\hat{V}\left(\hat{Y}_t\right)=\sum_{h=1}^{H}\frac{n_{ht}}{n_{ht}-1}\left(\sum_{i=1}^{n_{ht}}\hat{Y}_{hti}^2-\frac{\left(\sum_{i=1}^{n_{ht}}\hat{Y}_{hti}\right)^2}{n_{ht}}\right), \quad (3.7.2.20)$$

where

$\hat{Y}_{hti}$ is the estimator of the total of the characteristic y for a PSU i at stratum h. That is

$\hat{Y}_{hti}=\sum_{j=1}^{m_{hti}}w_{htij}y_{htij}$ (3.7.2.21).

We have assumed here that the first-stage sampling fraction is small and that the first-stage sample is drawn with replacement, so that the variance estimation depends only on the first-stage sampling.

The covariance $\hat{Cov}(\hat{Y}_2,\hat{Y}_1)$, which is the third component of the relation (3.7.2.18), is calculated on the basis of the selected data from the common PSUs surveyed in both quarters, as follows:

$$\hat{Cov}\left(\hat{Y}_2,\hat{Y}_1\right)=\sum_{h=1}^{H}\frac{n_{hc}}{n_{hc}-1}\sum_{i=1}^{n_{hc}}\left(\hat{Y}_{h2i}-\frac{\sum_{i=1}^{n_{hc}}\hat{Y}_{h2i}}{n_{hc}}\right)\cdot\left(\hat{Y}_{h1i}-\frac{\sum_{i=1}^{n_{hc}}\hat{Y}_{h1i}}{n_{hc}}\right) \quad (3.7.2.22) \text{ or equivalently}$$

$$\hat{Cov}\left(\hat{Y}_2,\hat{Y}_1\right)=\sum_{h=1}^{H}\frac{n_{hc}}{n_{hc}-1}\left(\sum_{i=1}^{n_{hc}}\hat{Y}_{h2i}\cdot\hat{Y}_{h1i}-\frac{\left(\sum_{i=1}^{n_{hc}}\hat{Y}_{h2i}\right)\cdot\left(\sum_{i=1}^{n_{hc}}\hat{Y}_{h1i}\right)}{n_{hc}}\right) \quad (3.7.2.23).$$

In order to reduce the cost of quarterly surveys, any ultimate units (e.g. individuals) newly selected in the sample for all time points (e.g. quarters) are taken from the same PSUs. But again, there is only partial overlapping in the ultimate units, between time points. In this case the variance of $\hat{D}=\hat{Y}_2-\hat{Y}_1$ is given by:

$$\hat{V}\left(\hat{D}\right)=\sum_{h=1}^{H}\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}\left(\hat{D}_{hi}-\frac{\sum_{i=1}^{n_h}\hat{D}_{hi}}{n_h}\right)^2 \quad (3.7.2.24) \text{ or}$$

$$\hat{V}\left(\hat{D}\right)=\sum_{h=1}^{H}\frac{n_h}{n_h-1}\left(\sum_{i=1}^{n_h}\hat{D}_{hi}^2-\frac{\left(\sum_{i=1}^{n_h}\hat{D}_{hi}\right)^2}{n_h}\right), \quad (3.7.2.25)$$

where

$\hat{D}_{hi}=\hat{Y}_{h2i}-\hat{Y}_{h1i}$ is the estimator of the change of the total in PSU i at stratum h.

Let x_{hij} be the value of the characteristic x for an ultimate unit j, belonging to the hi primary sampling unit (cluster). Moreover, X stands for the total of x in the population. The form of the estimator $\hat{R}$ on the basis of the two-stage design will then be represented by:

$$\hat{R}_t=\frac{\hat{Y}_t}{\hat{X}_t}=\frac{\sum_{h=1}^{H}\sum_{i=1}^{n_{ht}}\sum_{j=1}^{m_{hti}}w_{htij}\cdot y_{htij}}{\sum_{h=1}^{H}\sum_{i=1}^{n_{ht}}\sum_{j=1}^{m_{hti}}w_{htij}\cdot x_{htij}}. \quad (3.7.2.26)$$

Let $\hat{D}=\hat{R}_2-\hat{R}_1$ be the estimate of the net change of ratios between the two time periods. Then the problem addressed here is the estimation of the variance.

One possible approach to obtaining a design-based variance estimate of $\hat{D}$ is Taylor linearisation. According to this approach the variance of $\hat{D}$ is given by (Roberts and Kovacevic, 1999):

$$V(\hat{D})=V\left(\hat{R}_2-\hat{R}_1\right)\cong V\left(\hat{Z}_2-\hat{Z}_1\right)=V\left(\hat{Z}_2\right)+V\left(\hat{Z}_1\right)-2Cov\left(\hat{Z}_2,\hat{Z}_1\right), \quad (3.7.2.27)$$

where

$$\hat{Z}_2-\hat{Z}_1=\sum_{hij\in S_2}w_{h2ij}z_{h2ij}-\sum_{hij\in S_1}w_{h1ij}z_{h1ij} \text{ and}$$

$$z_{htij}=\frac{y_{htij}-\hat{R}_t x_{htij}}{\hat{X}_t},\ t=1,2.$$

The quantities $\hat{V}\left(\hat{Z}_t\right)$ and $\hat{Cov}\left(\hat{Z}_2,\hat{Z}_1\right)$ are calculated by applying the above formulae for linear statistics.

A mass of papers and books have been written over recent decades on the subject of ***replication methods*** for variance estimation. However, there has been virtually no discussion on replication methods for variance estimation of estimators of net change under sample coordination. One exception is Canty and Davidson (1999), who discuss replication-based variance estimation within the context of the LFS.

3.7.3 Estimation of gross change

Martin Axelson (Statistics Sweden)

Introduction to gross change and flows

This section discusses the concept of gross change. Gross change refers to aggregates of change between time points at element level. When a study variable is categorical, the parameter of interest is typically a table that reflects how elements have transferred or 'flowed' between categories over time. Consequently, gross change is typically referred to as flows when the study variable is categorical.

Let U_t and U_{t+1} denote the target populations at times t and $t+1$ respectively, and let $U = U_t \cup U_{t+1}$. Following Nordberg (2000), let $U = U_d \cup U_p \cup U_b$, where $U_p = U_t \cap U_{t+1}$, $U_d = U_t - U_p$, and $U_b = U_{t+1} - U_p$. Moreover, let $U_{t,C} \subseteq U_t$ denote a domain of interest at time t, and let $U_{t+1,C} \subseteq U_{t+1}$ denote the corresponding domain of interest at time $t+1$. For example, U_t and U_{t+1} may denote the resident population at time t and $t+1$ respectively, and $U_{t,C}$ and $U_{t+1,C}$ may be the subset of people who are unemployed at the corresponding time points. When estimation of flows is being discussed, the parameter of interest is typically a table based on some function of the vector:

$$\mathrm{N}_p = \left(N_{p,\bar{C}\bar{C}} \quad N_{p,C\bar{C}} \quad N_{p,\bar{C}C} \quad N_{p,CC}\right)'$$

where $N_{p,\bar{C}\bar{C}}$ denotes the number of elements in the set $U_p \cap U_{t,\bar{C}} \cap U_{t+1,\bar{C}}$, i.e.

$$N_{p,\bar{C}\bar{C}} = \sum_{k \in U_p \cap U_{t,\bar{C}} \cap U_{t+1,\bar{C}}} 1 \quad ,$$

$U_{t,\bar{C}}$ and $U_{t+1,\bar{C}}$ denote the subsets of people who are employed at the corresponding time points, and $N_{p,\bar{C}C}$, $N_{p,C\bar{C}}$, and $N_{p,CC}$ denote the sizes of the corresponding sets $U_p \cap U_{t,\bar{C}} \cap U_{t+1,C}$, $U_p \cap U_{t,C} \cap U_{t+1,\bar{C}}$ and $U_p \cap U_{t,C} \cap U_{t+1,C}$. Sometimes, the parameter of interest is a table based on a function of the vector

$$\mathrm{N} = \left(\mathrm{N}_p' \quad N_{d,\bar{C}} \quad N_{d,C} \quad N_{b,\bar{C}} \quad N_{b,C}\right)'$$

where $N_{d,\bar{C}}$, $N_{d,C}$, $N_{b,\bar{C}}$, and $N_{b,C}$ are the sizes of the corresponding sets $U_d \cap U_{t,\bar{C}}$, $U_d \cap U_{t,C}$, $U_b \cap U_{t,\bar{C}}$ and $U_b \cap U_{t,C}$. Hence, each cell in N corresponds to the size of a particular domain.

Let

$$y_{t,k} = \begin{cases} 1 & \text{if } k \in U_{t,C} \\ 0 & \text{otherwise} \end{cases}$$

$$z_{t,k} = \begin{cases} 1 & \text{if } k \in U_t \\ 0 & \text{otherwise} \end{cases}$$

and let $y_{t+1,k}$ and $z_{t+1,k}$ be analogously defined for time $t+1$. Then, using the variables $y_{t,k}$, $z_{t,k}$, $y_{t+1,k}$, and $z_{t+1,k}$, each element in N may be expressed as a sum over the elements in U for an appropriately defined summation variable. For example, let $I_{k,C\bar{C}} = z_{t,k} y_{t,k} z_{t+1,k} (1 - y_{t+1,k})$. Since

$$I_{k,C\bar{C}} = \begin{cases} 1 & \text{if } k \in U_p, y_{t,k} = 1, \text{and } y_{t+1,k} = 0 \\ 0 & \text{otherwise} \end{cases}$$

it follows that

$$N_{p,C\bar{C}} = \sum_{k \in U} I_{k,C\bar{C}} .$$

In practice, the matrices N and N_p are often of larger dimensions, due to the fact that U_t and U_{t+1} are partitioned into more domains of interest. In the above notation, the variables y_t and y_{t+1} are indicator vectors rather than scalar indicator variables.

Estimating gross change under sample coordination

As is evident from the vector N, estimating gross change typically gets reduced to estimating domain sizes. A pre-requisite for design-based estimation of gross change is that the values of both y_t and y_{t+1} are recorded for a subset of elements in $U = U_t \cup U_{t+1}$. In theory, this may be achieved in many different ways. For example, one possibility would be to select a probability sample s from U at time $t+1$, and then observe $y_{t,k}$ and $y_{t+1,k}$ for all responding elements. However, when sampling over time is considered, typically y_t is recorded for the responding elements at time t and y_{t+1} is recorded for the responding elements at time $t+1$. Hence, as there must be an overlap between the two response sets for estimation of gross change to take place, positive sample coordination is often used to guarantee estimation of gross change.

For example, by using the positive sample coordination property of certain surveys like EU-SILC or EU-LFS, we are able to use the sample data to study transitions from one status to another:

- Longitudinal poverty rates are based on union and/or intersection of an individual's poverty status at a series of cross-sections (Eurostat, 2010a). The 'persistent at-risk-of-

poverty rate' is actually the main EU-SILC longitudinal indicator. For a four-year panel, it is defined as the share of persons who are at risk of poverty at the fourth wave of the panel and at two of the three preceding waves. The at-risk-of-poverty threshold at wave i (i={1,2,3,4}) is set at 60 % of the median income at wave i.

- Although collection of longitudinal data is not laid down in the EU Regulation, many national LFS samples have a panel dimension which makes it possible to estimate flows into and out of employment, unemployment, the labour market, etc.

General considerations on variance estimation for gross change

Let $\hat{N}_p$ denote an estimator for N_p. Even though the construction of $\hat{N}_p$ may be far from trivial, it is assumed that $\hat{N}_p$:

(a) appropriately reflects the sampling design used to select $s_t \cup s_{t+1}$;
(b) is non-response-adjusted;
(c) may incorporate auxiliary information.

Let $V(\hat{N}_p)$ denote the covariance matrix of $\hat{N}_p$ and let $\hat{V}(\hat{N}_p)$ denote an estimator for $V(\hat{N}_p)$. Depending on the method used for sample coordination and the construction of $\hat{N}_p$, different ways, such as analytical methods or replication methods, can be considered for estimation of $V(\hat{N}_p)$. Whether or not an existing software tool can be used depends on the choice of $\hat{V}(\hat{N}_p)$. *However, for $\hat{V}(\hat{N}_p)$ to be considered as a valid estimator for $V(\hat{N}_p)$, it should properly reflect the implications of points (a) — (c) above.*

Precision requirements for flow estimators can also be expressed using precision thresholds. For instance, we may want a standard error of 0.5 percentage points for EU-SILC's persistent at-risk-of-poverty rate. Precision requirements for estimators of flows can alternatively be expressed in terms of the minimum sample size to be achieved between any pair of consecutive waves. See *Appendix 7.6* for more details.

Summary

It is usual for NSIs to conduct continuing surveys, where the same population is sampled repeatedly over time. A possible classification of the surveys that use sample over time is 1) repeated surveys, 2) longitudinal surveys (panel surveys, rotating panel surveys, repeated panel surveys, split panel surveys) and 3) rolling samples. Rolling samples are a separate category from longitudinal surveys. While rolling samples are taken by moving to different PSUs each wave, longitudinal surveys include units of the previous sample in the new sample. On the other hand, repeated surveys may or may not be rolling samples and may or may not be longitudinal surveys.

Such surveys are typically conducted to meet one or more of the following objectives:

- provide estimates of parameters at specific time points;
- provide estimates of parameters defined as averages over a period of time;
- provide estimates of net change between two time points;

- provide estimates of gross change.

It is recommended to take into account covariance effects in the estimation of variance for the averages over a period of time and for the net change between two time points (estimates of change). In general, covariance estimation under sample coordination is not straightforward. Covariance estimation in surveys based on panels is a research topic. In practice it is usually assumed that the non-overlapping parts are nearly independent, and covariance is computed through correlation of the overlapping samples. However, this assumption has to be assessed.

This chapter proposes an analytical method to compute variance for annual averages. The method can be applied by using the indications from this chapter and Section 3.4, in cases where the sampling design is simple random sampling without replacement, stratified random sampling or two-stage sampling (when primary sampling units are selected with probabilities proportional to size and secondary sampling units are selected by simple random sampling without replacement). The handbook also proposes an analytical method to compute variance for estimators of net change for stratified multi-stage sampling. The sampling designs are rotating panel designs with at least some common PSUs between successive periods.

Estimation of gross changes typically gets reduced to estimation of domain sizes, while variance for gross changes uses the variance estimation methodology used for domains.

4. Computing standard errors for national and European statistics

This chapter identifies possible approaches to increasing the availability of variance of EU statistics, lists the pros and the cons of the various approaches, and recommends the integrated approach — applying replication methods and generalised variance functions. These methods, which were generally described in Section 3.3, are presented in this chapter as possible solutions for the integrated approach.

4.1 Recommendations for improving computation of standard errors for national and European statistics

Denisa Camelia Florescu and Jean-Marc Museux (Eurostat)

A common current practice is for NSIs to compute standard errors for a limited list of indicators and national population breakdowns and then report the figures to Eurostat. Nevertheless, standard errors are needed for all relevant indicators and breakdowns, and not just for those for which they have been computed. This means we need to develop appropriate methodologies for variance estimation in order to make information on standard errors more accessible to data users. There are three important constraints, however. The first is that the diversity of results that are of interest to data users is often so large that individual variance computations statistic-by-statistic may be time-intensive. This issue is even more acute at EU level, with thirty or so countries being systematically handled. The variance estimation method therefore has to be fast. The second constraint is that data users do not always have statistical expertise, so the approach has to be made as easy and straightforward as possible for non-statisticians. Finally, confidentiality issues generally place restrictions on the variables which are available in public microdata files. In particular, design-related variables like stratum or primary sampling unit (PSU) codes are often removed from the files as their disclosure power is generally considered to be high. As a result, data users are unable to perform variance calculations by taking the whole sampling design into account, and this could lead to severely biased estimates. In extreme cases, users have no access to microdata, the access being restricted to the parameters of generalised variance functions that model the relationship between the variance or the relative variance of an estimator and its expectation. On the other hand, the estimation method has to be as 'accurate' as possible in the sense that it has to reflect most of the sampling design components.

Seeking to boost the availability of standard errors for EU statistics requires extra efforts, and these may be shared by Eurostat and the NSIs. This chapter assesses the delegation by NSIs to Eurostat of variance estimation tasks. Such delegation of tasks centralises some of the work and depends on the accuracy with which Eurostat is able to reproduce the actual standard errors of the NSIs.

Three main approaches are presented and discussed: decentralised, fully centralised and integrated. These approaches are possible with aggregated data and/or microdata transmission from the NSIs to Eurostat.

The decentralised approach

This is the most common approach; it is linked to the transmission of aggregated data. NSIs compute standard errors for all relevant indicators and national breakdowns and report them to Eurostat. NSIs use their own methods and tools for variance estimation. There is neither a common method nor a tool.

In practice, Eurostat defines the indicators and breakdowns for which point and variance estimates have to be calculated by NSIs. Then, the national partners have to come up with the estimates, by taking into account at least:

- the main sampling design components (stratification, clustering, etc.);
- non-response adjustments;
- other weighting effects (especially weight calibration (Deville and Särndal, 1992)).

Eurostat then computes the point and variance estimates for the ESS statistics, by considering countries as technical strata.

There is a requirement at Eurostat level to monitor national processes. In this respect, a decentralised approach requires sound statistical expertise in the NSIs in order to carry out variance calculations and to provide Eurostat with the requisite information to enable it to monitor compliance of process and output.

A decentralised approach should rely on NSIs using suitable methods and tools for the different sampling designs (guidance on some suitable and unsuitable methods is provided by the matrix in Appendix 7.4) and on ensuring that such methods are actually followed by each NSI (guidance provided by the metadata template in Appendix 7.3).

The advantages of this approach are:

- it meets the requirement of a standard delivery of an aggregated table;
- Eurostat involvement is minimal.

The weak points are:

- The use of different variance estimation methods leads in principle to negligible differences in the results. However, the results lack comparability if the methods and tools do not account for exactly the same sources of variability, which is difficult for Eurostat to monitor.
- This strategy yields standard errors for a limited list of indicators and breakdowns. It does not meet the need for standard errors for all relevant indicators and breakdowns and is not flexible: if Eurostat needs estimates of standard errors for extra/unforeseen indicators and breakdowns, the only possibility is to ask the countries to provide them.

A decentralised approach in which the NSIs calculate and transmit standard errors for all relevant indicators and breakdowns may impose a considerable burden on NSIs, especially, or disproportionally so, for smaller ones.

The fully centralised approach

Under this approach, Eurostat sets up a common methodology for variance estimation and computes standard errors for all indicators and breakdowns on the basis of the sampling design information provided by NSIs. Even though countries use different sampling designs, a

common variance estimation method is needed at EU level, to take into account the different national-level strategies.

We might consider using replication methods (bootstrap, jackknife, balanced repeated replication, etc.) as standardised methods. These methods are flexible enough for most of the commonly used sampling designs. They are also able to take into account the complex structure of a sample in order to yield estimates for the whole population and diverse sub-populations (breakdowns).

This approach is linked to the transmission of microdata.

EU-SILC developed ad-hoc jackknife macros in SAS that required NSIs to send Eurostat not only the file containing the microdata but also a file with the following additional variables at record level (Ghellini *et al*, 2009):

- the stratum to which the ultimate sampling unit belongs;
- the primary, the secondary, etc. sampling units to which the ultimate sampling unit belongs;

  Self-representing primary sampling units are treated as primary strata, and their secondary sampling units are treated as primary sampling units.
- where systematic sampling is used at any sampling stage, the order of selection of the primary, the secondary, etc. sampling units;

  This information allows the effect of implicit stratification on the overall variance to be taken into account.
- the final sampling weight of the units used in the estimation and adjusted for non-response and calibration.

This file had to make it possible to identify the stratum to which each primary, secondary etc. sampling unit belongs, the primary, secondary etc. sampling unit to which each household belongs, and the household to which each individual belongs.

Design weights could also be transmitted although they are not necessary for the macros; they serve to enable the impact of non-response and calibration on weights to be assessed.

Eurostat was then able to calculate replicate weights, replicate estimates and variance.

In EU-SILC, the SAS macros took into account the effect of implicit stratification on the variance. However, calibration and imputation effects were not fully taken into account and the method was not validated in-depth. Eurostat could have re-calculated the imputed value for each replication using random imputation methods.[37] The variability of the imputed values between replications would have served to incorporate the variance due to imputation in the overall variance. Implementing this approach was found nevertheless highly time-consuming.

The use of jackknife for EU-SILC was a feasibility study. The subsequent step was to test other methods than jackknife, i.e. bootstrap and linearisation. Comparative experiments were carried out on a limited number of countries; the results of different methods are similar. The present choice is to work with linearisation (ultimate cluster approximation), which was discussed at the Net-SILC2 workshop on accuracy and was validated by the SILC Working Group. This approach provides acceptable results given the administrative considerations.

[37] This approach would have resulted in duplication of the NSIs' imputation work. However, the random imputation by Eurostat would have been merely to incorporate the variability of imputation into the whole variance, and the imputed values would have not been used for any other purpose, e.g. the estimation of point estimates.

Experience with EU-SILC demonstrated that the quality reports collect descriptions of (for instance) sampling designs which vary greatly from one NSI to another. Many clarifications and exchanges were needed with NSIs on the number of stages of the sampling design, whether there are self-representing PSUs, whether systematic sampling allows for implicit stratification, etc. *If the fully centralised approach were to be implemented, the metadata template (Appendix 7.3) would be particularly useful and is recommended as a means of collecting clear and detailed information on the sampling designs.*

The advantages of this approach are:

- Contrary to the decentralised approach, this one achieves full flexibility in terms of estimating standard errors for all relevant indicators and breakdowns, including extra/unforeseen indicators and breakdowns, without the need to ask the NSIs to provide them.

  For replication methods, domain estimation can be viewed as a particular case of national estimation where the target variable takes value 0 for units outside the domain. So any breakdown needed can be handled by using the same method as long as national weights are available.

- It drastically reduces the burden on NSIs.
- It enables Eurostat to fully control the estimation of standard errors needed for compliance assessment.
- It facilitates full harmonisation of the way standard errors are computed.

On the other hand, there are ***weak points*** too:

- A main one is that Eurostat is now burdened with the preparation and use of full design/estimation information (calibration, rotation of sample, imputation) and with the likely difficulties related to unavailable information required for weighting and the complexity concerning rotation schemes (derivation of approximated variance that takes the panel covariance into account). The information needed by Eurostat can be technically complex to acquire and requires sampling expertise and knowledge of details which are country-specific. This approach requires very sound statistical expertise from Eurostat and is burdensome, in terms of personnel and computing power. The average number of records in the LFS (i.e. ultimate sampling units) per quarter in 2007 was almost 1.4 million. If we take into account stratification (32 countries*country wise strata) and differences between the various element and cluster sampling designs, then any overall variance estimation experiment becomes very cumbersome. This approach is clearly not feasible for the LFS, which is a continuous short-term survey.
- NSIs will compute and publish their own precision estimates and continue to use their own methods (at least in the short run), irrespective of whether Eurostat publishes precision estimates under a fully centralised approach. This is because NSIs have their own data requests from data users. A comparability problem between precision computed by NSIs and Eurostat (using different methods) therefore arises if the methods and tools do not account for exactly the same sources of variability.

The integrated approach

Under this approach, NSIs compute certain required statistics and report them to Eurostat, which uses them to compute standard error estimates for any relevant indicator and breakdown. These statistics can be as follows:

- Using a common replication method, NSIs report replicate weights and full sampling weights to Eurostat. On the basis of this information and the file containing the microdata, Eurostat then calculates the replicate estimates and the overall variance. This is done from the variability of the replicate estimates around the estimate based on full sampling weights, over all replications.

  This option is linked to the transmission of microdata.

  Replicate weights are already calculated by NSIs by taking into account the main sampling design features (stratification, multi-stage selection, calibration, etc.), so sample structure variables are not needed at Eurostat.

  See Section 4.2 for more information.

- Under the use of generalised variance functions, NSIs first calculate a set of standard errors using direct methods (analytical or replication). They then use the results to estimate parameters of generalised variance functions. NSIs transmit such parameters to Eurostat, which uses them to calculate standard errors for all indicators and any breakdowns needed.

  This option is linked to the transmission of aggregated data.

  See Section 4.2 for more information.

The integrated approach requires sound statistical expertise at Eurostat; nonetheless, the success of this task also relies on the statistical expertise in NSIs and on the quality of the estimation of national statistics, which is closely linked to the actual national sampling design.

Advantages of this approach are:

- It has good flexibility, and allows Eurostat (and other data users) to estimate standard errors for all relevant indicators and breakdowns.

  For replication methods, standard errors can be estimated for extra/unforeseen indicators and breakdowns, without any need to ask the NSIs for them.

  For replication methods, domain estimation can be viewed as a particular case of national estimation, where the target variable takes the value 0 for units outside the domain. So as long as national weights are available, any breakdown needed can be handled by using the same method.

- The replication methods take account of the sampling design while simultaneously enabling users of secondary survey data (e.g. Eurostat) to estimate standard errors without knowing the detailed sampling design.

- It enables sampling design information to be integrated at source.

- The approach based on the use of a common replication method supports the comparability of national standard error estimates, assuming that the common method/tool used by the NSIs accounts for exactly the same sources of variability.

Disadvantages of this approach are:

- NSIs will still compute and publish their own precision estimates, using their own methods, even if Eurostat publishes precision estimates under an integrated approach, using the common method. This is because NSIs have their own data requests to deal with from data users. This disadvantage is more likely to occur with generalised variance functions than with replication methods. (In the latter case, it is assumed that, in the long run, NSIs will change their own methods with the common replication method.) A comparability problem therefore arises between precision computations by NSIs and Eurostat (using different methods) if the methods and tools do not account for exactly the same sources of variability.
- Under replication methods, the burden on NSIs and expertise requirements of NSIs are no less than with the decentralised approach. However, the same replicate weights can be used to estimate variance for any indicator and domain needed, including extra/unforeseen indicators and domains.
- The use of generalised variance functions reduces but does not eliminate the need for NSIs to calculate variances using direct methods. However, the parameters of the GVFs can be carried over from one data collection to another with similar features (in terms of sampling design, survey variables, etc.).

  There may also be problems with the validity of national calculations of statistics required by Eurostat in connection with generalised variance functions (i.e. the validity of the parameters of generalised variance functions).

In the long run, the objective and a main challenge are for Eurostat and NSIs to converge and use the same (replication) method. This will allow Eurostat (and other data users) to estimate standard errors for all relevant indicators and breakdowns, including for extra/unforeseen ones. It will also prevent comparability problems when different national methods and tools do not account for exactly the same sources of variability. Guidelines and training sessions can be organised at Eurostat level to train the people in charge at country level.

Unlike the decentralised approach, the integrated approach meets the objective of increasing the availability of standard errors for Eurostat. As against the fully centralised approach, the integrated approach shares the burden between NSIs and Eurostat. *The integrated approach tends to be the most feasible and is recommended.*

Summary

The portfolio for policy-making indicators is becoming broader and more detailed with time. The need to provide standard errors for them is increasing.

A customary practice is for NSIs to transmit standard errors for national estimates to Eurostat for a limited list of indicators and breakdowns; Eurostat then computes standard errors for European estimates for the same indicators and breakdowns. This is incompatible with the requirement described above because of ever changing needs.

There are three main approaches (options) that enable standard errors to be computed and disseminated for national and European estimates for all relevant indicators and breakdowns.

Under the **decentralised approach**, the option is to ask NSIs to estimate and report standard errors for national estimates, for all relevant indicators and breakdowns. This is very burdensome for NSIs since the needs may change over a short period of time and lead to

duplication of efforts (the stovepipe approach). Furthermore, standard errors computed by NSIs using different methods and tools may raise concerns about the comparability of results if the methods and tools do not account for exactly the same sources of variability.

In any case, a decentralised approach should rely on NSIs using suitable methods and tools for the different sampling designs (guidance on some suitable and unsuitable methods is provided by the matrix in Appendix 7.4) and on ensuring that such methods are actually followed by each NSI (guidance provided by the metadata template in Appendix 7.3).

Under a **fully centralised approach**, Eurostat would develop a methodology and regularly estimate standard errors for national and European estimates based on information provided by countries. The promising option for Eurostat is the use of replication methods based on design information provided in microdata files and external contextual information (totals for calibration, etc.). However, Eurostat would need to put in a considerable amount of work and expertise to develop a methodology and regularly estimate standard errors on the basis of information provided by countries. This is not very feasible, especially for short-term surveys like the LFS.

However, if the fully centralised approach with the use of replication methods is applied, the metadata template (Appendix 7.3) would be particularly useful and is recommended for collecting clear and detailed information on the sampling designs.

An **integrated approach**, *where the burden is shared between Eurostat and NSIs, tends to be the most feasible option and is therefore recommended*. A major drawback is however the burden for the NSIs having to estimate certain statistics required by Eurostat which might not be produced with current methods and tools used by NSIs. However, guidelines and training sessions can be organised to train those in charge at country level. In the long run, the objective is for Eurostat and NSIs to converge and use the same method.

4.2 Possible methods for implementing the integrated approach of variance estimation

Loredana Di Consiglio, Stefano Falorsi (ISTAT) and Ioannis Nikolaidis (EL.STAT)

This section describes methods which can be used to implement the integrated approach presented in the previous section. The purpose is to increase the availability of standard errors for all relevant indicators and breakdowns, in order to meet the needs of Eurostat (and other data users).

Generalised Variance Functions

Generalised Variance Functions (GVFs) attempt to model the relative variance[38] $\mathrm{Re}l\,\mathrm{var}(\hat{Y})$ of a survey estimator $\hat{Y}$ as a function F of its expectation $Y = E(\hat{Y})$ (Wolter, 2007):

$$\mathrm{Re}l\,\mathrm{var}\left(\hat{Y}\right) = F\left(Y;\alpha,\beta,\gamma\ldots\right). \qquad (4.2.1)$$

[38] The relative variance of an estimator is defined as the ratio between its variance and the square of its expectation. Put another way, the relative variance is equal to the square of the coefficient of variation (relative standard error) of the estimator.

For each planned domain level, once the model parameters $\alpha, \beta, \gamma ...$ have been determined from a subset of estimated statistics of interest and their variances, the expression (4.2.1) can be used to estimate the variance of any other statistic of interest of the same type (totals, means or ratios) in the same (or similar) context (sampling design).

Hence, GVFs provide a quick and easy way of estimating variance of a statistic without resorting to direct variance computations which, as far as analytical methods are concerned, is more difficult and more complex. Besides, direct variance computations (analytical and replication methods) usually need microdata files with variables related to the sampling design (e.g. stratum identification codes, calibration variables), which data users generally do not have access to. Thus, GVFs might turn out to be an efficient strategy for making information on standard errors more accessible to data users: data producers can provide the parameters for a predefined set of GVFs (e.g. for each planned domain) from which data users are able to estimate variance for any statistic they may be interested in.

There are other benefits of using GVFs instead of direct variance computations:

- although there is no theoretical evidence for this, GVFs would generate variance estimates that are generally more stable than variances which are estimated one statistic at a time;
- GVFs can be carried over from one data collection to another with similar features (in terms of sampling design, survey variables, etc.);
- in particular, GVFs may be useful when dealing with repeated surveys (e.g. EU-SILC, LFS), as direct computations may provide variance estimates for the first wave of data, while GVFs based on first-wave results can be valuable instruments for estimating variances from second wave onwards;
- presentation of variance estimates statistic-by-statistic would considerably increase the size of survey reports, making them less easy to read. On the other hand, a summary table with estimated parameters from different variance models would be a more tractable option (Swanepoel and Stoker, 2000).

In many of its practical applications, GVFs model the relative variance of an estimator[39] as a decreasing function of its expectation:

$$\mathrm{Re}\,l\,\mathrm{var}\left(\hat{Y}\right) = \alpha + \frac{\beta}{Y}\ , \quad (4.2.2)$$

where the model parameters α and β are unknown and have to be estimated from a set of variance estimates $(\mathrm{Re}\hat{l}\,\mathrm{var}(\hat{Y}), \hat{Y})$ obtained through direct computations. Ordinary Least Squares (OLS) may be used as a natural fitting methodology. In order to smooth out the effect of outliers, Weighted Least Squares (WLS) may also be used, with weights being taken as a decreasing function of the relative variance.

The U.S. Census Bureau has been using this model for its Current Population Survey (CPS) since 1947. Another major survey in the United States that uses (4.2.2) is the National Health Interview Survey (NHIS) (Valliant, 1987).

A distinguishing feature of variance estimation based on GVFs is that the approach is mainly empirical: there is no irrefutable scientific evidence to guide the choice of a variance model.

[39] In most cases, GVFs deal with estimators of population and sub-population totals.

Nevertheless, the use of model (4.2.2) can be justified under certain assumptions. Let $\hat{Y}$ be an estimator of the total number Y of individuals who fall into a certain category (e.g. the total number of individuals aged 50-64, the total number of unemployed people, the total number of individuals who live in a certain geographical region). And let $P = \frac{Y}{N}$ be the proportion of the total population, of size N, in the category. Assuming an arbitrary sample selection leading to a sample of size *n*, the relative variance of the estimator $\hat{Y}$ of the population size Y is given by:

$$\mathrm{Re}\,l\,\mathrm{var}\left(\hat{Y}\right) = Deff \cdot \frac{1-P}{nP} = -\frac{Deff}{n} + \frac{N \cdot Deff}{nY} = \alpha + \frac{\beta}{Y}\ , \qquad (4.2.3)$$

where $\alpha = -\frac{Deff}{n}$ and $\beta = \frac{N \cdot Deff}{n}$ and $Deff$ is the 'design effect' (Kish, 1965).

Let V be the group of statistics to which the variance model is fitted. Assuming that the design effect $Deff$ and *n* vary little from one statistic to another in group *V*, then the model (4.2.2) should work on the basis of relationship (4.2.3). This rule is often applied by practitioners (Ghangurde, 1981).

Alternative variance models may also be considered (Wolter, 2007):

$$\mathrm{Re}\,l\,\mathrm{var}\left(\hat{Y}\right) = \alpha + \frac{\beta}{Y} + \frac{\gamma}{Y^2} \qquad (4.2.4)$$

$$\mathrm{Re}\,l\,\mathrm{var}\left(\hat{Y}\right) = \left(\alpha + \beta Y\right)^{-1} \qquad (4.2.5)$$

$$\mathrm{Re}\,l\,\mathrm{var}\left(\hat{Y}\right) = \left(\alpha + \beta Y + \gamma Y^2\right)^{-1} \qquad (4.2.6)$$

$$\mathrm{Re}\,l\,\mathrm{var}\left(\hat{Y}\right) = \alpha Y^{-\beta}. \qquad (4.2.7)$$

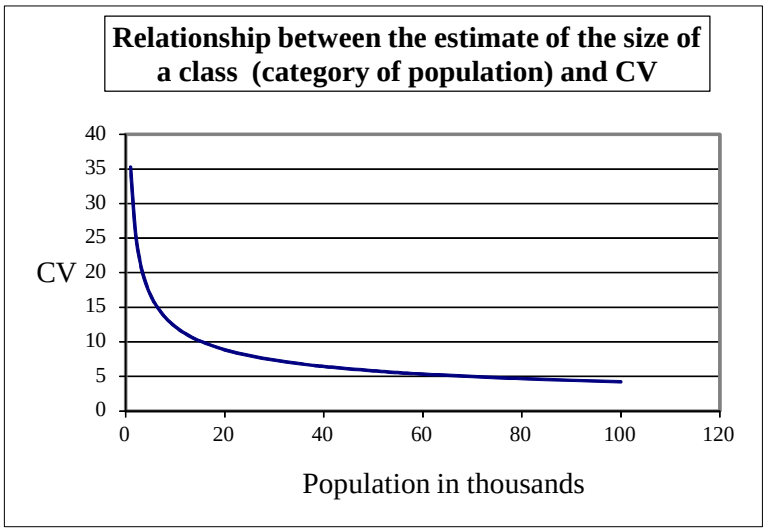
Model (4.2.7) is the one that the Hellenic Statistical Authority (ELSTAT) has been using to generate variance estimates for the Greek Labour Force Survey.

More precisely, the relationship between the estimated number $\hat{Y}$ of people (in thousands) who fall into the 'employment/unemployment' category (e.g. total number of unemployed people, total number of employed men aged 20-24) and the estimated coefficient of variation $\hat{CV}(\hat{Y})$ is given by:

$$\hat{CV}(\hat{Y}) = \frac{35.264864}{\hat{Y}^{0.461818}}. \qquad (4.2.8)$$

For instance, if we estimate the total number of unemployed females at 150 000, the relationship (4.2.8) leads to an estimated coefficient of variation of 3.49 %.

Figure 4.2.1: Relationship between estimate of the size of a class and CV



Moreover, by applying the logarithm on both sides in (4.2.7), we get the following equivalent relationship:

$$\log\left[\mathrm{Re}\,l\,\mathrm{var}\left(\hat{Y}\right)\right] = \alpha' - \beta\log(Y), \quad (4.2.9)$$

where $\alpha' = \log(\alpha)$.

The model (4.2.9) has been used by the Australian Bureau of Statistics and by Statistics Canada.

This is also applied by the GENESEES software for qualitative variables; hence it is used in all ISTAT surveys on households such as the Italian Labour Force Survey and the Italian Consumer Expenditure Survey.

Use of log transformation is generally recommended as it tends to reduce the impact of extreme values on the model, thus making residuals more symmetric and homoscedastic (Johnson and King, 1987).

Johnson and King (1987) also examined the effect of adding further variables to the variance model, particularly information about the 'sign' of the design effect (lower than 1/greater than 1). They showed that inclusion of this new variable significantly improved the model. They also suggest measuring the goodness of fit of a model by comparing the actual and the predicted standard errors: they recommend evaluating the goodness of fit by the percentage of variances which are underestimated by more than 20 %. As underestimation is considered more serious than overestimation, they also recommend that underestimated observations be over-weighted.

The GVF procedure has to be flexible enough to fit most of the commonly used sample strategies. Though this has interesting implications, there is no scientific evidence for this claim. For instance, weight calibration (Deville and Särndal, 1992) should not affect the model, although the procedure might result in extreme observations in terms of variance of the estimators. *In order to fit a 'good' variance model, outliers should be deleted. An alternative is to weight the observations so as to reduce the influence of the most extreme ones, or to use log transformation (see model 4.2.9).*

Some feasibility studies of GVFs in sample surveys can be found in Johnson and King (1987), Bieler and Williams (1990), Finamore (1999), Swanepoel and Stoker (2000), Cho *et al* (2002).

Success of the GVF technique critically depends on the grouping of survey statistics, i.e. on whether all statistics within a group behave according to the same mathematical model (Wolter, 2007). According to equation (4.2.3), this means that all statistics within a group have to share the same *Deff* and the same planned sample size. For instance, the Current Population Survey (CPS) statistics have been divided into six groups, with GVF fitted independently to each of them (Wolter, 2007):

- agricultural employment,
- total or non-agricultural employment,
- males only in total or non-agricultural employment,
- females only in total or non-agricultural employment,
- unemployment,
- unemployment for black and other races.

In general, the GVF procedure will only work if Deff is the same for all statistics within a group. Each group should therefore include statistics coming from the same sample survey, regardless of the survey wave (unless sampling design has been modified from one wave to the next), and dealing with the same domain of interest (e.g. region, unemployed people). Another important requirement is that all statistics within a group be of the same type (total, means, ratios, etc.): the GVF method is 'indicator-specific'. In fact, in most cases, GVFs deal with estimators of population and sub-population totals. Finally, it has to be borne in mind that country-specific variance models are needed for cross-national surveys, as sample selection procedures generally vary between countries.

To justify the use of GVFs, Valliant (1987) introduced a class of super-population models for which relative variance has the same form as in (4.2.2). He also discussed the properties of Weighted Least Squares as a fitting methodology for the variance model. A condition for relative variance to be the same as in (4.2.2) occurs when the model-based variance can be expressed as a linear combination of the model-based mean and its square. This condition is satisfied in the 'binary case' (target variable is a dummy variable) and in the 'Poisson' case (model-based variance is equal to the mean).

If we assume stratified two-stage sampling, then for a set V of survey statistics that follow the same model, we need further requirements from the predictive perspective: sub-populations should be evenly spread among clusters in the same stratum, and intra-cluster correlation coefficients should be approximately constant from one cluster to another within the same stratum. The latter two requirements may be difficult to satisfy with rare sub-populations and some care is therefore required with model (4.2.2) when applied to rare sub-populations.

Valliant also proved the consistency of the GVF estimator when all statistics in group V provide the same values for the model parameters α and β and if consistent point estimators of the relative variance are used in estimating such parameters. Finally, an empirical study supports the choice of model (4.2.2), with model parameters estimated by Weighted Least Squares (WLS) when specific conditions hold (see Valliant, 1987).

When it comes to quantitative variables, there is no analogous theory that supports the use of Generalised Variance Functions. However, the Italian Statistical Office (ISTAT) has used the

following relationship. Let $\hat{Y}$ be the estimator of the total Y of a quantitative variable. Then we get:

$$V\left(\hat{Y}\right)=\alpha+\beta Y+\gamma Y^2. \quad (4.2.10)$$

Note that variance of $\hat{Y}$ can be written as:

$$V\left(\hat{Y}\right)=Deff\times\left(N^2\frac{1-\frac{n}{N}}{n}S_y^2\right), \quad (4.2.11)$$

where $S_y^2=\frac{1}{N-1}\sum_{i=1}^{N}\left(y_i-\frac{Y}{N}\right)^2=\frac{1}{N-1}\left(\sum_{i=1}^{N}y_i^2-\frac{Y^2}{N}\right)$. (4.2.12)

By using the relation: $\left(\sum_{i=1}^{N}y_i\right)^2=\sum_{i=1}^{N}y_i^2+2\sum_{i=1}^{N}\sum_{i'>i}^{N}y_iy_{i'}$, (4.2.13) we get the following:

$$V\left(\hat{Y}\right)=Deff\times\left[\frac{(N-1)}{N}AY^2-2A\left(\sum_{i=1}^{N}\sum_{i'>i}^{N}y_iy_{i'}\right)\right], \quad (4.2.14)$$

where $A=N^2\cdot\frac{1-\frac{n}{N}}{n}\cdot\frac{1}{N-1}$. (4.2.15)

Thus, under the assumption that *Deff* is approximately constant for all estimators in V, we can formulate the following model:

$$\mathrm{Re}\,l\,\mathrm{var}\left(\hat{Y}\right)=\alpha+\frac{\beta}{Y^2}+u, \quad (4.2.16)$$

where u designates an error term. Furthermore, in order to take into account $\sum_{i=1}^{N}\sum_{i'>i}^{N}y_iy_{i'}$ it is possible to introduce a heteroscedasticity component in the model:

$$E\left(u_\omega^2\right)=\sigma_\omega^2=\sigma^2 f\left(\sum_{i=1}^{N}\sum_{i'>i}^{N}y_{\omega,i}y_{\omega,i'}\right) \quad (\omega=1,...,\Omega). \quad (4.2.17)$$

Hence Weighted Least Squares (WLS) has to be used when using this component σ_ω^2 $(\omega=1,...,\Omega)$, or an estimate has to be provided, thereby increasing the computational burden of the method.

It may therefore be preferable to adopt an empirical relationship that shows good fit, as in the model introduced above:

$$V\left(\hat{Y}\right)=\alpha+\beta\hat{Y}+\gamma\hat{Y}^2+u. \quad (4.2.18)$$

An in-depth analysis of the fit is required in all cases. For qualitative variables we should check the Deff associated with all variables to see if they can fit the model. And of course in the quantitative case, data analysis can suggest an alternative formulation for the appropriate GVF model, so a customary analysis of the model should be applied.

The variance estimation software GENESEES, developed by ISTAT, contains a module for analysing models. This helps in selecting proper relationship between errors and estimates that need to be provided to data users together with the planned disseminated statistics (Pagliuca, 1995). For all household surveys, ISTAT produces annexes containing the estimated coefficients for models (differentiated by areas) and R^2 (coefficients of determination) together with examples of target amounts (corresponding to specified proportions) and their respective coefficients of variation (percent).

An example of the regression model as used to provide measures of associated variances can be found in the Labour Force Volume (in Italian) http://www.istat.it/dati/catalogo/20100217_00/forze_di_lavoro_media_2008.pdf.

Replication methods

Replication methods have already been described in the previous Section 3.3.

Under the integrated approach, the advantage of using replicate weights is that a single formula can be used to calculate the standard error of many types of estimates. Contrary to the approach based on generalised variance functions, the use of replicate weights provides a direct way of computing standard errors. This specifically means that there is no longer the need for grouping statistics. Standard errors are estimated statistic-by-statistic from replicate weights. Direct variance estimates are often expected to be more accurate than 'indirect' ones based on variance functions, although they may be more inconvenient for some users to calculate since they require some extra programming and, more generally, technical assistance.[40]

Many statistical bodies release files with replicate weights along with public use microdata files, thus enabling data users to perform variance calculations. For instance, the U.S. Census Bureau releases each fall a public use data file for the Current Population Survey (CPS) and a public use replicate weight file (U.S. Census Bureau, 2006). Similarly, several Statistics Canada surveys, like the Survey of Labour and Income Dynamics (SLID) and the National Population Health Survey (NPHS), provide bootstrap weights, or variants thereof, with their microdata for the purpose of variance estimation.

Whatever the replication method used, calculating the variance of an estimator is done in a somewhat similar fashion, with only minor changes depending on the exact method used.

Let $\hat{\theta}$ be an estimator for a parameter θ. The latter can have a linear or a more complex form. Broadly speaking, by using a given set of replicate weights, an estimate $\hat{V}\left(\hat{\theta}\right)$ of the variance of the estimator $\hat{\theta}$ is given by (Eurostat, 2002; European Central Bank, 2010; Asparouhov and Muthén, 2010):

$$\hat{V}(\hat{\theta}) = C\sum_{a=1}^{A} c_i \left(\hat{\theta}_{(a)} - \hat{\theta}_{(\cdot)}\right)^2, \quad (4.2.19)$$

[40] In addition to replicate weights, syntax files (under SAS and SPSS) may be provided to data users to help them with this issue. A quick user guide may also be desirable in this respect.

where

$\hat{\theta}_{(a)}$ is the weighted estimate of θ based on the same formula as $\hat{\theta}$, obtained in replicate sample A,

$\hat{\theta}_{(\cdot)}$ is the mean of the *A* values $\hat{\theta}_{(a)}$,

c_i and C are method-specific constants, whose values are given in the following table,

n_h is the number of sampled units in stratum *h*, and *n* is the sample size.

Table 4.2.1 Replication methods

Method	C	c_i
JK1	$(n-1)/n\,;(A-1)/A$	1
JKn	1	$(n_h-1)/n_h$
Bootstrap	$1/(A-1)$	1
BRR	$1/A$	1
RG	$1/A(A-1)$	1

Confidentiality issues may arise from the release of replicate weights with public use microdata files (Yung, 1997). A possible solution consists of averaging the bootstrap weights over a fixed number D of bootstrap samples. Statistics Canada used this approach for many of its surveys (Phillips, 2004). For example, in the General Social Survey (GSS), Statistics Canada produced *A*=5000 bootstrap weight variables. These were then averaged in groups of size *D*=25 in order to obtain 200 mean bootstrap weights that accompany the microdata. Similarly, Statistics Canada's Workplace and Employee Survey (WES) provides 100 mean bootstrap weights, each of which is the mean of *D*=50 bootstrap weights.

Modifying the bootstrap variance estimator presented in (4.2.19), and taking A instead of $A-1$ as the denominator, we obtain the 'mean bootstrap variance estimator' as follows:

$$\hat{V}_{\bar{BS}}(\hat{\theta}) = \frac{D}{A}\sum_{a}\left(\hat{\theta}_{(a)} - \hat{\theta}_{(\cdot)}\right)^2, \quad (4.2.20)$$

where each a^{th} mean bootstrap sample set of weights is equal to the means of *D* bootstrap weights. In this specification, the term $\hat{\theta}_{(a)}$ is obtained using the a^{th} mean bootstrap weight variable as opposed to the standard bootstrap weight variable used in equation (4.2.19).

An adjustment is made by inserting the integer *D* into the numerator of the variance estimator, which re-introduces the variability that had been removed by using an average weight. See Chowhan and Buckley (2005).

Finally, the choice of method for the integrated approach should be made after the possible methods have been analysed under the criteria of applicability, accuracy and administrative considerations (Section 3.3).

Summary

A generalised variance function (GVF) is a mathematical model that describes the relationship between a statistic (such as a population total) and its corresponding variance.

Data producers can provide the parameters for a predefined set of GVFs (for a group of survey statistics) so that data users are then able to estimate the variance for any national and European statistic (from that group) they may be interested in. To determine parameters, data producers apply direct computations of variance (with analytical, linearisation, replication methods). Then, the parameters and the GVFs can be carried over from one data collection to another with similar features (by assuming that sampling design remains unchanged).

Success of the GVF technique critically depends on the grouping of survey statistics, i.e. on whether or not all statistics within a group behave according to the same mathematical model. This means that design effect should be the same for all statistics within a group. All statistics should refer to the same domain of interest. The statistics within a group should be of the same type because the GVF method is indicator-specific.

GVFs may be very useful and provide quick results when dealing with repeated surveys that produce short-term statistics (e.g. LFS quarterly surveys) since direct computations may provide variance estimates for the first wave of data, while GVFs based on first-wave results can be valuable instruments for estimating variances from the second wave onwards.

A distinguishing feature of variance estimation based on GVFs is that the approach is mainly empirical, especially when it comes to quantitative variables.

The GVF procedure should be flexible enough to fit most of the commonly used sample strategies; there is however no scientific evidence for this claim.

Contrary to the approach based on generalised variance functions, the use of replicate weights in replication methods provides a direct way of computing standard errors.

Direct variance estimates are often expected to be more accurate than 'indirect' ones based on variance functions, although they require some extra programming and technical assistance.

The advantage of using replicate weights is that a single formula is used to calculate the standard error of different complex sampling designs and many types of indicators. Replication methods can deal with complex statistics, unlike the approach based on variance functions, which often deals with linear estimators.

Confidentiality issues may arise from the release of replicate weights with public use microdata files. A possible solution consists of averaging the bootstrap weights over a fixed number of samples.

The choice of method for the integrated approach should be made after the possible methods have been analysed under the criteria of applicability, accuracy and administrative considerations (Section 3.3).

5. Possible ways of assessing compliance with precision requirements

Denisa Camelia Florescu and Jean-Marc Museux (Eurostat)

This chapter sets out and explains the different approaches to assessing compliance with precision requirements from a Eurostat and NSI perspective. It underlines the need for tolerance and regular monitoring. It also introduces the metadata template (in *Appendix 7.3*) to support one of the approaches.

The objective of compliance monitoring is to detect major failures of target objectives that are defined in the legal acts.

Non-compliance can result from the following features:

- *a badly designed statistical instrument;*
- *badly conducted survey operations (non-response, failure in editing system, low quality of micro records).*

Compliance monitoring should target these features, especially when their nature is liable to hamper ESS statistics quality by producing lower precision at reporting domains (breakdowns) or a lack of comparability of indicators across countries.

In addition, compliance monitoring requires a minimum degree of harmonisation of, for instance, methods for computing the actual standard errors.

The essential features of compliance monitoring are:

- transparency of procedure and predictability;
- detection of major defects;
- warning system that triggers correction measures;
- assurance of overall output quality.

A monitoring system should allow smooth handling of conjunctural (non-structural) defects. It should not lead to a continuous redesign of the system but should be part of a rolling review strategy. It should not be sensitive to changes in the estimated phenomena.

Compliance monitoring is not an easy task because:

- best/standard practices are not widespread and accepted with respect to variance estimation;
- the way requirements are phrased in legal texts is not always clear-cut;
- Eurostat does not have the resources to monitor compliance by taking on board all survey design specificities.

The following three strategies for assessing compliance are possible:

1. The first strategy consists of estimating output quality on the basis of closed and ad-hoc formulae whose parameters are frozen for a period of time.

   For instance, precision can be estimated on the basis of formulae under simple random sampling, by including a design effect that is specific to a country and to a class of

estimators. These parameters are evaluated at the beginning of the period by Eurostat and Member States according to best practices. *The validity of the approximation should be tested on training data sets.*

The design effect can be:

- estimated by NSIs and then reported to Eurostat;
- estimated by Eurostat on the basis of the actual standard errors transmitted by countries for a limited list of indicators.

However, estimating the design effect is a not straightforward task. Please see more details on design effect in *Appendix 7.2.*

2. *With regard to the second strategy, under the assumption that the standard errors are produced regularly and timely, for instance through quality reports, only systematic deviations should be traced.* The actual precision of survey results can be used for assessment, i.e. by considering a design effect which changes in time.

Compared to the first strategy, this second strategy for compliance assessment seems to be more feasible in that it addresses the actual design effect and not the expected one.

NSIs use different methods and tools to compute precision estimates (decentralised approach, see Section 4.1), whereas Eurostat relies only on information provided by countries in their quality reports.

Compliance monitoring plays a key role in ensuring that the precision reported by countries accords with the proper methodology (process). In relation to this, a metadata template (checklist) is recommended in Appendix 7.3. This requires NSIs to report on their national sampling designs, survey processes (e.g. non-response adjustment, imputation, calibration) and the variance estimation methods and tools used. The aim of the metadata template is to assess the soundness and appropriateness of the variance estimation method and tool, in relation to the sampling design used and the type of indicators (some related guidance being provided by the matrix in Appendix 7.4). The template also requires information on whether the effects of different procedures used in the survey process, e.g. non-response adjustment, imputation, calibration, have been accounted for in precision estimation.

The metadata template was conceived to be as comprehensive as possible, so as to be relevant for several statistical domains. *In order to use it as an element in a compliance assessment strategy for a specific statistical domain (survey), it should be adapted to the specific features of that domain.*

The following remarks concern the relationship between the metadata template and the standard for quality reports:

- The standard for quality reports provides recommendations for preparation of comprehensive quality reports, for a full range of statistical processes and their outputs. There are six statistical processes, two of which are sample survey and census (Eurostat, 2009a). The metadata template is relevant to sample surveys and can be adapted for censuses.
- The standard for quality reports requires information on relevance, accuracy, timeliness and punctuality, accessibility and clarity, coherence and comparability, user needs and perceptions, performance, cost and respondent burden etc. The

metadata template has a more restricted scope, i.e. the precision part of accuracy (the metadata template does not cover bias).

- The metadata template collects information in a more structured, clear and detailed manner on issues like sampling design and variance estimation methods and tools, to feed into the previously specified objective. It is not always easy to understand the features of national sampling designs from quality reports, especially when the information provided is not structured and detailed. The metadata template helps to standardise the reporting of specific information.

Eurostat coordinators for specific statistical domains/surveys may decide to detail the relevant part of the quality report according to the metadata template to ensure that NSIs do not report the same amount of information twice.

3. National precision estimates can be computed centrally by Eurostat, through replication methods or generalised variance functions (see chapter 4 for more details). Using a common method to compute national precision estimates greatly facilitates compliance assessment.

However, this strategy holds true only when a *fully centralised or integrated approach* to the estimation of standard errors has been set up (see Section 4.1).

It is recommended that the compliance assessment strategy be based on the principles of transparency and tolerance.

Transparency is needed on how an assessment is carried out, so that everybody can verify whether agreed procedures have been observed. *Transparency concerns both NSIs and Eurostat. They should accurately provide each other with all relevant information. An explanation of all elements in the precision requirements should be made available, together with an explanation for why those choices were made.*

*In addition, some **tolerance** should be accepted when comparing the estimated standard errors with benchmarks indicated in the precision requirements*. Tolerance should be accepted for the following reasons:

- the results produced by different, though suitable, methods may not perfectly match under a decentralised approach where NSIs use different variance estimation methods (which may use different degrees of approximation, and may account for different sources of variability);
- what can be computed is not the 'true' standard error for a given estimate but only an estimated standard error, which in turn has its own variance. Thus, the estimated standard error can be higher than the true one (which is unknown);
- the regulation may set up a threshold for estimating standard error, but the estimate depends on the value of the estimated percentage. *Therefore, to be on the safe side and to ensure that NSIs avoid continuous adjustment of the survey design, some tolerance should be used to assess design efficiency in cases where the upper threshold is surpassed but the estimated percentage is, say, exceptionally high within the range 0%-50% (for instance when the estimated percentage of unemployed people in the working age population is exceptionally high).*

The requisite tolerance may be reflected either directly in the requirements or in the compliance assessment strategy. The LFS Group of Experts discussed the pros and cons of both approaches and agreed on the latter. The Group preferred to have a stricter rule in the

requirement. In fact, the former would in practice be perceived as a relaxation of the requirements, which may lead to requests to reduce the sample size of the LFS (Eurostat, 2010d).

*A distinction should be made between **occasional deviations** and **systematic deviations**. The TF strongly recommends taking into account only systematic deviations and not occasional deviations when ruling on non-compliance. Whenever precision thresholds are repeatedly or systematically not met, countries should provide measures (e.g. for a better survey design) to improve the degree of compliance in/on subsequent survey waves/occasions. However, compliance assessment should focus on the design features of national surveys. Non-compliance should arise from insufficient sample size, high non-response, ineffective stratification, systematic imbalances in the actual samples, etc. An increase in standard error arising only from the change in the value of the estimated percentage should not be considered as non-compliance. Likewise, a higher value of the estimated variance that arises because of the variability of the variance estimator should not be considered as non-compliance.* For example, let us consider a rotating panel survey with two-stage sampling design, where the primary sampling units are localities and the secondary sampling units are dwellings. Let us also consider that between successive waves, the part of the sample which is replaced consists of dwellings selected from the same localities (only the sample of dwellings is rotated). The sampling variance of an estimator from such a sample has two sources — the first-stage and the second-stage variance. Given that the sample of localities is stable over a long time, the contribution of the first-stage variance to the overall sampling variance is stable as well. *If the sample of localities, because of an unlucky draw, causes the estimated variance of the first-stage to be higher than the true variance, even if the deviation is systematic, this deviation should not be considered as non-compliance.*

In the case of the LFS, the Group of Experts agreed that the compliance assessment strategy should be spelled out, jointly by Eurostat and NSIs, in a gentlemen's agreement, as a means of ensuring that assessment is based on agreed rules and common understanding. Elements to be included in the gentlemen's agreement are (Eurostat, 2010d):

- reference to methods and tools for estimating variance for the different sampling designs. This is an element in ensuring that standard errors are correctly estimated. Some reference can be provided by the matrix on methods for variance estimation prepared by the DIME TF (*Appendix 7.4*);
- information on the way standard errors are estimated. This is another element in ensuring that standard errors are correctly estimated. The metadata template *(Appendix 7.3)* prepared by the DIME TF (which should be adapted to the specific needs of the LFS) is to make it possible to assess whether standard errors are correctly estimated for a specific sampling design;
- definition of the principle of tolerance and how it should be applied. This is crucial to determining when a country should be considered as non-compliant with the requirements.

Summary

The chapter reviews three strategies for assessing compliance:

- fixed normative rules (closed and ad-hoc formulae whose parameters — e.g. design effect — are estimated once for a period of time) which are agreed in advance between NSIs and Eurostat. The shortcomings of this approach are that it may differ significantly from the

actual precision (depending on progress over time and the quality of estimates) and that the design effect is not easy to calculate;

- systematic deviations can be traced on the basis of information from quality reports. *Use of the metadata template in Appendix 7.3 of the handbook is recommended when assessing whether the variance estimation method and tool used are appropriate in relation to sampling design and type of indicators. The purpose of the template is also to assess whether the effects of the different procedures used in the survey process, e.g. non-response adjustment, imputation, calibration, have been accounted for in the estimation of precision;*
- national precision estimates are computed centrally by Eurostat, through replication methods or generalised variance functions (see chapter 4).

It is recommended that the compliance assessment strategy be based on the principles of transparency and tolerance. Tolerance may be granted either directly in the requirements or in the compliance assessment strategy. The second approach may be preferred, as the former would in practice be perceived as a relaxation of the requirements, with a spurious effect such as an artificial reduction of sample size.

The TF strongly recommends taking into account only systematic deviations and not occasional deviations when ruling on non-compliance. Non-compliance is characterised by precision thresholds that are repeatedly or systematically surpassed and should arise from insufficient sample size, high non-response, ineffective stratification, systematic unbalances in the actual samples, etc. An increase in standard error arising only from the change in the value of the estimated percentage should not be considered as non-compliance. Likewise, a higher value of the estimated variance that arises because of the variability of the variance estimator should not be considered as non-compliance. In the event of non-compliance, preference should be given to investigating the source of the increased variability and to taking measures to improve the degree of compliance in/on subsequent survey waves/occasions.

6. References

Andersson, C. and Nordberg, L. (1994). *A method for variance estimation of non-linear functions of totals in surveys — Theory and software implementation*. Journal of Official Statistics, Vol. 10, No 4, pp. 395-405.

Ardilly, P. (2006). *Les techniques de sondage*. Second Edition, Paris: Technip.

Ardilly, P. and Osier, G. (2007). *Cross-sectional variance estimation for the French 'Labour Force Survey'*. Survey Research Methods, Vol. 1, No 2, pp. 75-83.

Ardilly, P. and Tillé, Y. (2005). *Sampling methods: exercises and solutions*. New York: Springer.

Asparouhov, T. and Muthén, B. (2010). *Resampling Methods in Mplus for Complex Survey Data*. Mplus Technical Report, Los Angeles, CA.

Bean, J. A. (1975). *Distribution and Properties of Variance Estimators for Complex Multistate Probability Sample*. Vital and Health Statistics, Series 2, No 65, National Centre for Health Statistics, Public Health Service, Washington, DC.

Beaumont, J. F. and Mitchell, C. (2002). *The system for estimation of variance due to nonresponse and imputation (SEVANI)*. Proceedings of Statistics Canada Symposium, 2002.

Bellhouse, D. R. (1985). *Computing Methods for Variance Estimation in Complex Surveys*. Journal of Official Statistics. Vol. 1, No 3, pp. 323–329.

Berger, Y. G. *Variance due to imputation*. Southampton Statistical Sciences Research Institute, University of Southampton, UK.

Berger, Y. G. (2004). *Variance estimation for measures of change in probability sampling*. Canadian Journal of Statistics, Vol. 32.

Berger, Y. G. (2005). *Variance estimation for systematic sampling from deliberately ordered populations*. Communications in Statistics — Theory and Methods, Vol. 34, No 7, 1533-1541.

Berger Y. G. (2007). *A jackknife variance estimator for unistage stratified samples with unequal probabilities*. Biometrika, Vol. 94, No 4, 953-964.

Berger, Y. G. (2008). *A note on asymptotic equivalence of jackknife and linearisation variance estimation for the Gini Coefficient*. Journal of Official Statistics, Vol. 24, No 4, 541-555.

Berger, Y. G and Priam, R. (2010). *Estimation of correlations between cross-sectional estimates from repeated surveys — an application to the variance of change*. Proceedings of the 2010 statistics Canada Symposium.

Berger Y. G. and Rao J. N. K. (2006). *Adjusted jackknife for imputation under unequal probability sampling without replacement*. Journal of the Royal Statistical Society: Series B (Statistical Methodology), Vol. 68, No 3, 531-547.

Berger, Y. G. and Skinner C. J. (2005). *A jackknife variance estimator for unequal probability sampling*. Journal of the Royal Statistical Society: Series B (Statistical Methodology), Vol. 67, No 1, 79-89.

Berger, Y. G. and Tillé, Y. (2009). *Sampling with unequal probabilities*. In: D. Pfeffermann and C.R. Rao. (editors). Elsevier.

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

Bethlehem, J. and Schouten, B. (2004): *Nonresponse adjustment in household surveys*. Discussion paper 04007, Statistics Netherlands.

Bieler, G. S. and Williams, R. L. (1990). *Generalised standard error models for proportions in complex design surveys*. Proceedings of Section on Survey Research Methods of the American Statistical Association, pp. 272-277.

Binder, D. A. (1983). *On the variances of asymptotically normal estimators from complex surveys*. International Statistical Review, 51(3), pp. 279–292.

Binder, D. A. (1998*). Longitudinal surveys: why are these surveys different from all other surveys*? Survey Methodology, 24, pp. 101-108.

Bjørnstad, J. F. (2007). *Non-Bayesian Multiple Imputation* (*with Discussion*). Journal of Official Statistics, Vol. 23, No 4, 433-491.

Booth, J., Butler, R., and Hall, P. (1994). *Bootstrap methods for finite populations*. Journal of the American Statistical Association, Vol. 89, pp. 1282–1289.

Burke, J. and Rust, K. (1995). *On the performance of jackknife variance estimation for systematic samples with small numbers of primary sampling units*. Proceedings of the American Statistical Association Section on Survey Research Methods, pp. 321-326.

Canty, A. J. and Davison, A. C. (1999). *Resampling-based variance estimation for Labour Force Surveys*. The Statistician, 1999, 48, Part 3, pp. 379 – 391.

Caron, N. (1998). *Le logiciel poulpe: aspects méthodologiques*. Proceedings of the Journées de Méthodologie Statistique.

Chauvet, G. (2007). *Méthodes de Bootstrap en population finie*. PhD Thesis.

Cho, M. J., Eltinge, J. L., Gershunskaya, J. and Huff, L. (2002). *Evaluation of generalised variance function estimators for the U.S. Current Employment Survey*. U.S. Bureau of Labor Statistics.

Chowhan, J. and Buckley, N. J. (2005). *Using Mean Bootstrap Weights in Stata: A BSWREG Revision*. Statistics Canada Research Data Centres: Information and Technical Bulletin, 2 (1), 23-38.

Cochran, W. G. (1977). *Sampling techniques*. New York: John Wiley and Sons.

Davison, A. C. and Sardy, S. (2004). *Resampling methods for variance estimation*. Technical report, DACSEIS deliverable D5.1.

Davison, A. C. and Sardy, S. (2007). *Resampling variance estimation in surveys with missing data*. Journal of Official Statistics, Vol. 23, No 3, pp. 371–386.

Deville, J. C. (1987). *Réplications d'échantillons, demi-échantillons, jackknife, bootstrap.* In 'Les Sondages', Paris. Economica.

Deville, J. C. (1999). *Variance estimation for complex statistics and estimators: Linearisation and residual techniques.* Survey Methodology, December 1999, Vol. 25, No 2, pp. 193-203.

Deville, J. C. and Särndal, C. E. (1992). *Calibration estimators in survey sampling*. Journal of the American Statistical Association, No 87, pp. 376-382.

Deville, J. C. and Särndal, C. E. (1994). *Variance estimation for the regression imputed Horvitz-Thompson estimator.* Journal of Official Statistics, Vol.10, No 4, pp. 381–394.

Deville, J. C., Särndal, C. E. and Sautory, O. (1993). *Generalised Raking Procedures in Survey Sampling*. JASA, Vol. 88, No 423, pp.1013-1020.

Deville, J. C. and Tillé, Y. (2004). *Efficient balanced sampling: the cube method*. Biometrika Vol. 91, pp. 893–912.

Deville, J. C. and Tillé, Y. (2005). *Variance approximation under balanced sampling*. Journal of Statistical Planning and Inference, Vol. 128, pp.569-591.

Di Consiglio, L. and Falorsi, S. (2010). *Design effect*. Paper prepared for the DIME Task Force on Accuracy (Precision Requirements and Variance Estimation).

Dippo, C. S. and Wolter, K. M. (1984). *A comparison of variance estimators using the Taylor Series Approximation*. Proceedings of the American Statistical Association Section on Survey Research Methods, pp. 113-121.

Duncan, G. J. and Kalton, G. (1987). *Issues of design and analysis of surveys across time.* International Statistical Review, Vol. 55, pp. 97-117.

Efron, B. and Tibshirani, R. (1986). *Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy*. Statistical Science 1, 54-75.

Escobar, E. L. and Berger, Y. G. (2010). *A novel jackknife variance estimator for two-stage without replacement sampling designs.* In 'Abstracts of Communications of the 10th International Vilnius Conference on Probability Theory and Mathematical Statistics', pp. 144. TEV, Vilnius. ISBN 978-609-433-009-4.

European Central Bank (2010). *Household Finance and Consumption Network. Variance estimation for the HFCS*.

Eurostat (2001). *Meeting of the Working Group on Statistics on Income and Living Conditions (EU-SILC).* European Community Household Panel (ECHP), Luxembourg, 26-27 April 2001.

Eurostat (2002). *Variance estimation methods in the European Union*. Monographs of official statistics. Luxembourg: Office for Official Publications of the European Communities.

Eurostat (2005). *The SAS macros for linearising EU-SILC complex income indicators*. Available on CIRCA.

Eurostat (2009a). *ESS handbook for quality reports*. Eurostat Methodologies and Working papers. Luxembourg: Office for Official Publications of the European Communities.

Eurostat (2009b). *European Health Interview survey. Task Force III report on sampling issues.* November 2009.

Eurostat (2010a). *Income and living conditions in Europe*. Luxembourg: Office for Official Publications of the European Communities.

Eurostat (2010b). *Methodological manual for statistics on the information society*.

Eurostat (2010c). *Minutes of the Group of Experts on Precision Requirements for the Labour Force Survey*. Luxembourg, 26-27 April 2010.

Eurostat (2010d). *Minutes of the Group of Experts on Precision Requirements for the Labour Force Survey*. Luxembourg, 4-5 October 2010.

Eurostat (2010e). *ESS Guidelines for the Implementation of the ESS Quality and Performance Indicators*. Luxembourg: Office for Official Publications of the European Communities.

Eurostat (2011). *Working Group meeting 'Statistics on Living Conditions'. Processing of design variables for variance estimation.* Luxembourg, 11-13 May 2011.

Falorsi, S., Moretti, D., Righi, P. and Rinaldelli, C. (2008). *Experiences of resampling approach in the household sample surveys.* Paper presented at the Q2008 European Conference on Quality in Official Statistics, Rome, 8-11 July 2008.

Falorsi S. and Russo P. (2001). *Il disegno di rilevazione per le indagini Panel sulle famiglie*. Rivista di statistica ufficiale, Vol. 3, pp. 55-90.

Fay, B. E. (1991). *A design-based perspective on missing data variance*. Proceeding of the 1191 Annual Research Conference, U.S. Bureau of the Census. pp. 429-440.

Finamore, J. M. (1999). *Generalised variance parameters for the 1997 National Survey of College Graduates*. U. S. Bureau of the Census, Demographic Statistical Methods Division, Health Surveys and Supplements Branch.

Frankel, M. R. (1971). *Inference from survey samples*. Institute for Social Research, Ann Arbor, Michigan.

Gabler, S., Häder, S. and Lahiri, P. (1999). *A model based justification of Kish formula for design effects for weighting and clustering*. Survey Methodology, Vol. 25, No 1, pp. 105-106.

Gabler, S., Häder, S. and Lynn, P. (2003). *Refining the concept and measurement of design effects.* Paper for the 54th session of the International Statistical Institute, Berlin, August 2003.

Ganninger, M. (2006). *Estimation of design effects for ESS Round II. ESS Round 2: European Social Survey (2008): ESS-2 2004 Documentation Report*. Edition 3.1, Bergen, European Social Survey Data Archive, Norwegian Social Science Data Services.

Ganninger, M. (2009). *Design Effects: Model-based versus Design-based Approach*. GESIS — Leibniz-Institut für Sozialwissenschaften.

Ghangurde, P. D. (1981). *Models for estimation of sampling errors*. Proceedings of Section on Survey Research Methods of the American Statistical Association, pp. 209-212.

Ghellini G., Verma V., Betti G. (2009). *Clarification and guidelines on the sample structure variables required for sampling error computations*. Università degli Studi di Siena, report prepared for Eurostat.

Gross, S. (1980). *Median estimation in sample surveys*. Proceedings of the Survey Research Section of the American Statistical Association, pp. 181–184.

Hájek, J. (1964). *Asymptotic theory of rejective sampling with varying probabilities from a finite population*. Ann. Math. Statist. 35, 1491–1523.

Hansen, M. H., Hurwitz, W. N., & Madow, W. G. (1953). *Sample survey methods and theory* (Vol. I). New-York: Wiley.

Hartley, H. O., Rao, J. N. K., (1962). *Sampling with unequal probabilities and without replacement.* Annals of Mathematical Statistics 33, 350-374.

International Statistical Institute. *Glossary of statistical terms.* http://isi.cbs.nl/glossary/index.htm.

Johnson, E. G. and King, B. F. (1987). *Generalised variance functions for a complex sample survey*. Journal of Official Statistics, Vol. 3, No 3, pp. 235-250.

Kalton, G. (1979). *Ultimate cluster sampling*. Journal of the Royal Statistical Society, Series A, 142, pp. 210-222.

Kalton, G., Brick, J. M. and Lê, T. (2005). *Estimating components of design effects for use in sample design*. UNStat.

Kim, J. K. and Fuller, W. (2004). *Fractional hot deck imputation.* Biometrika (2004), 91, 3, pp. 559–578.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Kish, L. (1987). *Statistical design for research*. New York: John Wiley and Sons.

Kish, L. (1990). *Rolling samples and censuses*. Survey Methodology, Vol. 16, No 1, pp. 63-71.

Kish, L. (1994). *Multi-population survey designs*. International Statistical Review, Vol. 62, pp. 167-186.

Kish, L. (1995). *Methods for design effects*. Journal of Official Statistics, 11, pp. 55-77.

Kish, L. (1998). *Space / time variations and rolling samples*. Journal of Official Statistics, Vol. 14, pp. 31-46.

Kish, L. and Frankel, M. R. (1974). *Inference from complex samples*. Journal of the Royal Statistical Society, Series B, 36.

Kott, P. (2001). *The Delete-A-Group Jackknife*. Journal of Official Statistics, 17, pp. 521-526.

Kovacevic, M. S. and Binder, D. A. (1997). *Variance estimation for measures of income inequality and polarisation — The Estimating Equations Approach.* Journal of Official Statistics, Vol. 13, No 1, pp. 41-58.

Kovar, J. G. and Whitridge, P. J. (1995). *Imputation of business survey data*. Business Survey Methods, eds Cox, B. G., Binder, D. A., Chinnappa, B. N., Christianson, A., Colledge, M. J. and Kott, P. S., pp 403–423. New York. Wiley.

Krewski, D. and Rao, J. N. K. (1981). *Inference from stratified samples: Properties of the Linearisation, Jackknife and Balance Repeated Replication Methods*. The Annals of Statistics, Vol. 9, pp. 1010-1019.

Lavallée, P. (2005). *Estimation et calcul de précision pour des échantillons rotatifs non chevauchants.* Proceedings of the Journées de Méthodologie Statistique.

Lavallée, P. (2007). *Indirect sampling*. New York: Springer Series in Statistics.

Lavallée, P. and Caron, P. (2001). *Estimation Using the Generalised Weight Share Method: The use of Record Linkage.* Survey Methodology 27, 155–169.

Lê, T. and Verma, V. (1997). *An analysis of sample designs and sampling errors for the demographic and health surveys*. DHS Analytical Reports, No 3. Calverton, Maryland: Macro International, Inc.

Lehtonen, R. and Pahkinen, E. J. (1996). *Practical methods for design and analysis of complex surveys*. New York: John Wiley and Sons.

Liberts, M. (2012). *Precision Requirements in European Safety Survey (SASU).* 20 January 2012.

Lohr, S. L. (1999). *Sampling: Design and analysis*. Pacific Grove: Brooks/Cole publishing company.

Lynn, P. (2009). *Methodology of longitudinal surveys*. Wiley Series in Survey Methodology.

Mac Carthy, P. and Snowden, C. (1985). *The bootstrap and finite population sampling*. Technical report.

Mak, T. K. (2004). *Estimating variances for all sample sizes by the bootstrap*. Computational Statistics and Data Analysis, Vol. 46, Number 3, June 2004, 459-467.

Miller, Jr. R. G. (1974). *The Jackknife — A Review*. Biometrika, Vol. 61, pp. 1-15.

Moretti, D. and Rinaldelli, C. (2005). *EU-SILC complex indicators: the implementation of variance estimation*. ISTAT report.

Mulry, M. H. and Wolter, K. M. (1981). *The effect of Fisher's Z-transformation on confidence intervals for the correlation coefficient.* Proceedings of the Survey Research Section, American Statistical Association, 601-606.

Münnich, R., Zins, S. and Bruch, C. (2011a). *Variance estimation for complex surveys.* Technical report, AMELI deliverable D3.1. http://ameli.surveystatistics.net.

Münnich, R., Zins, S. and Bruch, C. (2011b). *Variance estimation for Laeken Indicators.* Technical report, AMELI deliverable D3.2. http://ameli.surveystatistics.net.

Nikolaidis, I. (2008). *The effects of weighting, clustering and stratification on sampling errors as process quality indicators for the Greek Labour Force.* Paper for the Workshop on LFS Quality Assurance, Athens, 23-24 October 2008.

Nikolaidis, I. (2010a). *The effect of implicit stratification in multistage sample surveys*. Paper prepared for the DIME Task Force on Accuracy (Precision Requirements and Variance Estimation), Luxembourg.

Nikolaidis, I. (2010b). *Traditional method for the variance estimation*. Paper prepared for the DIME Task Force on Accuracy (Precision Requirements and Variance Estimation), Luxembourg, 22 March 2010.

Nordberg, L. (2000). *On variance estimation for measures of change when samples are coordinated by the use of permanent random numbers*. Journal of Official Statistics, Vol. 16, No 4, pp. 363-378.

OECD. *Glossary of statistical terms*. http://stats.oecd.org/glossary/index.htm.

OECD (2010). *Programme for the International Assessment of Adult Competencies (PIAAC).* Technical Standards and Guidelines, December 2010.

Osier, G. (2009). *Variance estimation for complex indicators of poverty and inequality using linearisation techniques*. Survey Research Methods, Vol. 3, No 3, pp. 167-195.

Osier, G. (2010). *Variance requirements and estimation for ESS surveys.* Paper commissioned by Eurostat for the DIME Task Force on Accuracy (Precision Requirements and Variance Estimation), under a contract with Sogeti.

Osier, G. (2011). *Sampling one/all persons per household: pros and cons.* Paper prepared for the meeting of the EU Task Force on Victimisation, Luxembourg, 27-28 October 2011.

Osier, G. (2012). *The linearisation approach implemented by Eurostat for the first wave of EU-SILC: what could be done from second wave onwards?* Paper prepared for the workshop on standard error estimation and other related sampling issues in EU-SILC organised in the context of the EU-funded 'Net-SILC2' project, Luxembourg, 29-30 March 2012.

Owen, A.B. (2001). *Empirical Likelihood*. Chapman & Hall/CRC.

Pagliuca, D. (1995). *GENESEES V. 3.0 Funzione Stime ed Errori — Manuale utente e aspetti metodologici*. Tecniche e Strumenti Istat.

Park, I. and Lee, H. (2004). *Design effects for the weighted mean and total estimators under complex survey sampling*. Survey methodology, Vol. 30, No 2, pp. 183-193.

Park, I., Winglee, M., Clark, J., Rust, K., Sedlak, A. and Morganstein, D. (2003). *Design effects and survey planning.* Proceedings of Section on Survey Research Methods of the American Statistical Association, pp. 3179-3186.

Phillips, O. (2004). *Using Bootstrap Weights with WesVar and SUDAAN*. The Research Data Centres Information and Technical Bulletin, (Fall) 1(2):1-10, Statistics Canada Catalogue no. 12-002-XIE.

Place, D. (2008). *Longitudinal variance estimation in the French Labour Force Survey*. Paper presented at the Q2008 European Conference on Quality in Official Statistics, Rome, 8-11 July 2008.

Preston, J. (2009). *Rescaled bootstrap for stratified multistage sampling*. Survey Methodology, 35 (2), pp. 227 – 234.

Raj, D. (1968). *Sampling theory*. McGrawHill Inc., New York, USA.

Rao, J. N. K. (1975). *Unbiased variance estimation for multistage designs*. Sankhya, C, 37, 133 – 139.

Rao, J. N. K. (1988). *Variance estimation in sample surveys*. In Handbook of Statistics, Vol. 6, (Eds. P.R. Krishnaiah and C.R. Rao), Amsterdam: Elsevier Science, pp. 427-447.

Rao, J. N. K. (1990). *Variance estimation under imputation for missing data*. Technical Report, Statistics Canada, Ottawa.

Rao, J. N. K. and Shao, J. (1992). *Jackknife variance estimation with survey data under Hot Deck Imputation*. Biometrika, Vol. 79, No 4 (Dec., 1992), pp. 811-822.

Rao, J. N. K. and Wu, C. F. J. (1984). *Bootstrap inference for sample surveys*. Proceedings of Section on Survey Research Methods of the American Statistical Association, pp. 106-112.

Rao, J. N. K. and Wu, C. F. J. (1985). *Inference from stratified samples: second-order analysis of three methods for non-linear statistics*. Journal of the American Statistical Association 80, 620-630.

Rao, J. N. K. and Wu, C. F. J. (1988). *Resampling inference with complex survey data*. Journal of the American Statistical Association, Vol. 83, pp. 231–241.

Rao, J. N. K., Wu, C. F. J. and Yue, K. (1992). *Some recent work on resampling methods for complex surveys*. Survey Methodology, 18 (2), pp. 209 – 217.

Rinaldelli, C. (2006). *Experiences of variance estimation for relative poverty measures and inequality indicators.* COMPSTAT Proceedings in Computational Statistics, pp. 1465-1472.

Roberts, G. and Kovacevic, M. (1999). *Comparison of cross-sectional estimates from two waves of a longitudinal survey*. Proceeding of the survey methods section, RH Coats Bldg., Ottawa, Ontario K1A 0T6.

Rosen, B. (1991). *Variance for systematic pps-sampling*. Technical report, Report 1991:15, Statistics Sweden.

Rubin, D. B. (1976). *Inference and missing data (with discussion).* Biometrika, 63, pp. 581–592.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons, Inc.

Rust, K. (1985). *Variance Estimation for Complex Estimators in Sample Surveys.* Journal of Official Statistics, 1(4):381-397, 1985.

Rust, K. and Kalton, G. (1987). *Strategies for Collapsing Strata for Variance Estimation.* Journal of Official Statistics, Vol. 3, No 1, pp. 69–81.

Saigo, H., Shao, J. and Sitter, R.R. (2001). *A repeated half-sample bootstrap and balanced repeated replications for randomly imputed data*. Survey Methodology, 27, 189-196.

Salonen, R. (2008). *Some remarks on calculation of CVs for estimates of annual averages and estimates of changes between two successive quarters in the case of the Finnish LFS*. Workshop on LFS Quality Assurance, 23-24 October 2008, Athens.

Särndal, C. E. (1990). *Methods for estimating the precision of survey estimates when imputation has been used*. Proc. Symp. Measurement and Improvement of Data Quality, pp. 337–347, Ottawa: Statistics Canada.

Särndal, C. E. (2007). *The calibration approach in survey theory and practice*. Survey Methodology, Statistics Canada, Vol. 33, No 2 (Dec., 2007), pp. 99-119.

Särndal, C. E. and Lundström, S. (2005). *Estimation in surveys with nonresponse*. London: Wiley.

Särndal, C. E., Swensson, B. and Wretman, J. (1989). *The weighted residual technique for estimating the variance of the general regression estimator of the finite population total.* Biometrika, Vol. 76.

Särndal, C. E., Swensson, B. and Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer-Verlag.

Shao, J., Chen Y. and Chen Y. (1998). *Balanced repeated replication for stratified multistage survey data under imputation.* Journal of American Statistical Association, Vol. 93, No 442, pp. 819-831.

Shao, J. and Sitter, R. R. (1996). *Bootstrap for imputed survey data*. Journal of the American Statistical Association, Vol. 91, No 435 (Sep., 1996), pp. 1278-1288.

Shao, J. and Steel, P. (1999). *Variance estimation for survey data with composite imputation and non-negligible sampling fractions.* Journal of the American Statistical Association, Vol. 95, No 445. pp. 254-265.

Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap.* Springer.

Shao, J. and Wu, C. F. J. (1989). *A general theory for jackknife variance estimation.* The Annals of Statistics, 17 (3), pp. 1176 – 1197.

Sitter, R. (1992a). *Comparing three bootstrap methods for survey data*. Canadian Journal of Statistics, Vol. 20, pp. 135–154.

Sitter, R. (1992b). *A resampling procedure for complex survey data*. Journal of the American Statistical Association, Vol. 87, pp. 755–765.

Statistics Canada (2010). *Guide to the Labour Force Survey*, pp. 30-31.

Statistics Finland (2007). *Quality Guidelines for Official Statistics*. 2nd Revised Edition.

Steel, D. and McLaren, C. (2009). *Design and analysis of surveys repeated over time*. Handbook of Survey Statistics 29B, Amsterdam: Elsevier, pp. 289 – 313.

Swanepoel, R. and Stoker, D. J. (2000). *The estimation and presentation of standard errors in a survey report*. Statistics South Africa.

Tam, S. M. (1984). *On covariances from overlapping samples*. The American Statistician, Vol. 34, No 4, pp. 288-289.

Tepping, B. J. (1968). *Variance estimation in complex surveys*. Proceedings of the American Statistical Association, Social Statistics Section, pp. 11-18.

U.S. Census Bureau (1993). *Variance Computation by Users of SIPP Micro-Data Files*.

U.S. Census Bureau (2006). *Current Population Survey*: *Design and Methodology*. Technical Paper 66, October 2006.

Valliant, R. (1987). *Generalised variance functions in stratified two-stage sampling*. Journal of the American Statistical Association, Vol. 82, No 398, pp. 499-508.

Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York: John Wiley & Sons.

Verma V. (1982). *Estimation and presentation of sampling errors*. WFS Technical Bulletins, No 11. The Hague: International Statistical Institute.

Verma, V. and Betti, G. (2007). *Cross-sectional and longitudinal measures of poverty and inequality: variance estimation using Jackknife Repeated Replication*. Conference 2007 'Statistics under one umbrella', Bielefeld University.

Verma, V. and Betti, G. (2011): *Taylor linearisation sampling errors and design effects for poverty measures and other complex statistics*. Journal of Applied Statistics, 38, 8, 1549-1576.

Wolter, K. M. (2007). *Introduction to variance estimation*. Second Edition, New York: Springer-Verlag.

Woodruff, R. S. (1952). *Confidence Intervals for Medians and Other Position Measures.* Journal of the American Statistical Association, Vol. 47, No 260, pp. 635–646.

Woodruff, R. S. (1971). *A simple method for approximating the variance of a complicated estimate.* Journal of the American Statistical Association, 66, pp. 411-414.

Yung, W. (1997). *Variance estimation for public use files under confidentiality constraints*. Proceedings of Section on Survey Research Methods of the American Statistical Association, pp. 434-439.

7. Appendix

7.1 Glossary of statistical terms

This glossary proposes definitions and related notations for technical terms that are closely linked with precision requirements and estimation, with a view to ensuring a common and clear understanding of the concepts.

Population, survey variables, parameters and sample

U	Population (or universe) of the survey, for which estimates are wanted. The population U is supposed to be finite and composed of N elements $u_1, u_2 \cdots u_N$ which are clearly defined, identifiable and observable. For the sake of simplicity, we generally represent the element u_i with its numerical label i: $U = \{u_1, u_2 \cdots u_N\} = \{1,2 \cdots N\}$.
U_h	In case of stratified sampling, subset of the population elements in stratum h $(h = 1,2 \cdots H)$.
U_d	Subset of the population elements in domain d $(d = 1,2 \cdots D)$.
y_i	Value taken by a variable of interest y on population unit i $(i = 1,2 \cdots N)$. Broadly speaking, y can be one-dimensional (scalar) or multi-dimensional (vector).
θ	A population parameter, that is, a quantitative measure of a population. θ is a function f of values $y_1, y_2, \ldots, y_N$: $\theta = f(y_1, y_2, \ldots, y_N)$.
s	Sample, of size n, that is, a partial list of population units: $s = \{i_1, i_2 \cdots i_n\}$ where $i_1, i_2 \cdots i_n$ are the labels of the n sample units. For the sake of simplicity, we may identify the sample unit i_j with the numerical label j: $s = \{i_1, i_2 \cdots i_n\} = \{1,2 \cdots n\}$. A sample is said to be with replacement when population units may appear more than once in the sample, while a sample is said to be without replacement when population units may appear only once. A sample is selected from a population using either a so-called 'probability' sampling scheme (all population units have a known, fixed in advance, probability of being selected) or not.

S_0	The set of all possible samples of the population.
$p(s)$	Probability for a sample s to be selected. The probability distribution $p = \{p(s), s \in S_0\}$ provides a probability-based description of the sampling design: this is a random mechanism which assigns the selection probability $p(s)$ to a specific sample s in S_0. For a given sampling design $p(.)$, we can regard any sample s as the outcome of a set-valued random variable $\tilde{S}$, whose probability distribution is specified by the function $p(.)$. Then, we have: $\Pr(\tilde{S} = s) = p(s)$, for all s in S_0.
$\hat{\theta}$	Estimator of the population parameter $\hat{\theta}$, that is, a function of the random set $\tilde{S}$: $\hat{\theta} = \hat{\theta}(\tilde{S})$. The term 'statistic' may also be used instead of estimator. An estimator is a stochastic variable in that it takes different values from one realisation of $\tilde{S}$ to another.
$\hat{\theta}_s$	Value taken by the estimator $\hat{\theta}$ on a realised sample s: this value provides an estimate (and not an estimator) of the parameter θ. *A clear distinction should be made between an estimator, that is, a function of the random set $\tilde{S}$ and an estimate, which is a particular outcome of the estimator: an estimator is a stochastic variable, while an estimate is a numerical value.* However, for simplicity's sake, an estimator is generally noted as a function of the sample s (see previous). In fact, such a notation is not rigorously correct as s is a particular realisation of the sample selection random mechanism, and not the mechanism itself.
$E(\hat{\theta})$	Expected value of the estimator $\hat{\theta}$, that is, the average value over all possible samples: $E(\hat{\theta}) = \sum_{s \in S_0} p(s) \cdot \hat{\theta}_s$.
$y_1, y_2 \cdots y_n$	Values taken by the variable of interest y on sample units i $(i = 1,2 \cdots n)$.
π_i, π_{ij}	π_i: First-order inclusion probability of i. This is the a priori probability for a population unit i $(i = 1,2 \cdots N)$ to be selected in the sample. This is the sum of the selection probabilities of all samples which contain i: $\pi_i = \sum_{\substack{s \in S_0 \\ s \ni i}} p(s)$. π_{ij}: Second-order inclusion probability for selection of i and j. This is the

	a priori probability for population units i and j to be selected together in a sample. This is the sum of the selection probabilities of all samples which contain both i and j: $$\pi_{ij} = \sum_{\substack{s \in S_0 \\ s \ni i \& j}} p(s).$$
d_i	Design weight of i, that is, the inverse probability of selection π_i.
w_i	Sampling weight of i. In general, sampling weights arise from adjusting the design weights for total non-response and possibly calibrating them to external data sources. Sampling weights are used to draw inference from sample s to the target population U. For example, a standard estimator of the population total Y of variable y is given by: $$\hat{\theta}(s) = \sum_{i=1}^{n} w_i \cdot y_i \ .$$
$\hat{\theta}_{HT}$	The Horvitz-Thompson estimator, defined as: $$\hat{\theta}_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i}.$$ Assuming full response to the survey, $\hat{\theta}_{HT}$ provides a design-unbiased estimator of the population total of y, in the sense that: $$E\left(\hat{\theta}_{HT}\right) = \sum_{i=1}^{N} y_i \ .$$
$\hat{\theta}_w$	A linear estimator, that is, a weighted sum of the sample values: $$\hat{\theta}_w = \sum_{i=1}^{n} w_i \cdot y_i,$$ where w_i is the sampling weight of i. $\hat{\theta}_w$ provides an estimator of the population total of y.
S_y^2	Dispersion of the variable of interest y over the population U: $$S^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(y_i - \bar{Y}\right)^2 ,$$ where $\bar{Y}$ designates the population mean of y: $\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} y_i$.
s_y^2	Dispersion of the variable of interest y over the sample s, of size n:

	$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2,$$ where $\bar{y}$ designates the sample mean of y: $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. s_y^2 provides an unbiased estimate of the population dispersion S_y^2 under simple random sampling.
ρ^*	Intra-cluster correlation coefficient. It measures the degree to which members of a cluster resemble each other. Members of a cluster resemble each other more than elements of the population in general. When the intra-cluster correlation coefficient for a variable in a population is large, it may be necessary to select a much larger sample of elements and clusters. Mathematically, let the population consist of N elements, grouped into M clusters, with cluster i, $i = 1,2,\ldots,M$, consisting of N_i elements. Moreover, let $\bar{Y}$ be the population mean of the variable of interest and $\bar{Y}_i$ be the corresponding mean of the elements of cluster i. ρ^* is given by: $$\rho^* = 1 - \frac{S_W^2}{S_y^2},$$ where $S_W^2 = \frac{1}{N-M}\sum_{i=1}^{M}\sum_{j=1}^{N_i}(y_{ij} - \bar{Y}_i)^2$ denotes the variance within clusters and $S_y^2 = \frac{1}{N-1}\sum_{i=1}^{M}\sum_{j=1}^{N_i}(y_{ij} - \bar{Y})^2$ the total population variance.

Sampling designs

Simple random sampling without replacement of size n	Every sample of size n receives the same probability of being selected: $$p(s) = \begin{cases} 1\Big/\binom{N}{n} & \text{if } s \text{ is of size } n \\ 0 & \text{otherwise} \end{cases}$$
Simple random sampling with replacement of size n	Ordered design that gives the same selection probability $1/N^n$ to every ordered sample of size n: $$p(s) = p(\{i_1, i_2 \ldots i_m\}) = \begin{cases} 1/N^n & \text{if } m = n \\ 0 & \text{otherwise} \end{cases}$$
Stratified simple random sampling	The population U is divided into H non-overlapping parts $U_1, U_2 \ldots U_H$ called strata, of sizes $N_1, N_2 \ldots N_H$ $\left(\sum_{h=1}^{H} N_h = N\right)$. Then a simple random

	sample s_h without replacement of size n_h is selected in each stratum h. The selections are assumed to be independent: $$p(s_h) = \begin{cases} 1 \Big/ \binom{N_h}{n_h} & if\ s_h\ is\ of\ size\ n_h \\ 0 & otherwise \end{cases}$$
Proportional-to-size (πps) sampling of size n	In πps sampling, a sample of size n is selected without replacement so that the first-order inclusion probability π_i of a unit i is proportional to an auxiliary variable x_i: $$\pi_i = n x_i \Big/ \sum_U x_i \ .$$ πps sampling generally leads to more accurate estimators if the auxiliary variable x_i is correlated with what the survey intends to measure.
Systematic sampling of size n with equal probabilities	The units in the population are numbered in some order. To select a sample of n units, we select a unit at random from the first k units and every k^{th} unit thereafter. For instance, if k is 15 and if the first unit drawn is 13, the subsequent units are numbers 28, 43, 58 and so on. In fact, the selection of the first unit determines the whole sample.
Systematic sampling with unequal probabilities	Generalisation of the above to accommodate unequal probabilities of selection.
Single-stage cluster sampling	The population U is divided into M non-overlapping parts $U_1, U_2 \ldots U_M$, called clusters. A random sample of m clusters is selected and then all eligible units in the selected clusters are interviewed. If the elements from a cluster tend to be similar to each other with respect to the main survey target characteristics (more homogeneity), cluster sampling will be less accurate (i.e. result in higher sampling errors) than simple random sampling of the same size.
Indirect cluster sampling	A sample of clusters is obtained from a sample of other units. For example, a sample of individuals is selected from a population register and then a sample of households is obtained by taking all the households having at least one of their current members in the original sample of individuals.
Multi-stage element sampling	Multi-stage sampling refers to sampling designs in which the population units are hierarchically arranged and the sample is selected in stages corresponding to the levels of the hierarchy. The sampling units for the different stages are different. Sampling units are selected in various stages but only the last sample of units is studied.
Two-stage element sampling	This is a particular case of a multi-stage element sampling, when there are two sampling stages. The population is first grouped into disjoint sub-populations, called Primary

	Sampling Units (PSUs). A random sample of PSUs is drawn (first-stage sampling). In the second stage, a random sample of Secondary Sampling Units (SSUs) is drawn from each PSU in the first-stage sample.
Multi-phase sampling	Multi-phase sampling refers to sampling designs in which the same type of sampling unit (e.g. individuals) is sampled multiple times. In the first phase, a sample of units is selected and every unit is measured on some variable. Then, in a subsequent phase, a subsample of units of the same type is selected only from those units selected in the first phase and not from the entire population. The sample of units selected in each phase is adequately studied before another sample is drawn from it. Sampling information may be collected at the subsequent phase at a later time, and in this event, information obtained on all sampled units of the previous phase may be used, if this appears advantageous. Multi-stage sampling is a particular case of multi-phase sampling arising by imposing the requirements for invariance and independence of the second phase designs. See Särndal *et al* (1992), Section 4.3.1.
Two-phase sampling	This is a particular case of a multi-phase element sampling, when there are two sampling phases. Two-phase sampling is sometimes called ‘double sampling’.

Survey errors

Sampling errors	Sampling error is a measure of the variability between estimates from different samples, which disregards any variable errors and biases that result from the measurement and sample implementation process. Sampling error occurs only in surveys based on samples. Of course, sampling error represents only one component of the total survey error. For estimates based on small samples, this component may be the dominant one. In other situations, non-sampling errors may be much more important.
Non-sampling errors	Non-sampling errors are errors in the estimates which cannot be attributed to sampling fluctuations. Non-sampling errors may be categorised as: - coverage errors; - non-response errors; - measurement errors; - processing errors; - model assumption errors. Non-sampling errors occur in sample surveys and also in censuses. Non-sampling errors can be random and systematic. *Random non-sampling errors are a source of variability in estimates and should be taken into account when estimating total variance.* Systematic non-sampling errors are

	a source of bias in estimates.
Non-response errors	Non-response refers to the failure to obtain a measurement on one or more study variables for one or more sample units. When a whole unit is missed, we have unit non-response. When a unit is included but information on some items for it is missed, we have item non-response. Non-response causes an increase in variance due to decreased effective sample size and/or due to weighting and imputation introduced to control its impact. More importantly, it causes bias in so far as non-respondents are selective with respect to the characteristic being measured.
Coverage (frame) errors	Coverage errors arise from discrepancies between target and frame populations, and also from errors in selecting the sample from the frame. The condition of 'probability sampling' is violated if: (a) the survey population is not fully and correctly represented in the sampling frame; (b) the selection of units from the frame into the sample is not random with known non-zero probabilities for all units; or (c) not all units selected into the sample are successfully enumerated.
Measurement errors	These arise from the fact that what is measured about the units included in the survey can depart from the actual (true) values for those units. These errors concern accuracy of measurement at the level of individual units enumerated in the survey, and centre on the substantive content of the survey: definition of survey objectives and questions, ability and willingness of the respondent to provide information sought and quality of data collection, recording and processing. This group of errors can be studied in relation to various stages of the survey operation.
Processing errors	Processing errors are of the same nature as measurement errors. Possible sources of processing errors are data entry, data editing (checks and corrections) or coding.
Model assumption errors	Errors that occur with use of methods such as model-based (model-dependent) estimation, benchmarking, seasonal adjustment, forecasting and other methods that rely on assumptions that the model holds.

Precision measures, their estimators and specific estimates

The delineation between these concepts relies on contributions made by Martin Axelson, Kari Djerf, Mārtiņš Liberts, Ioannis Nikolaidis and Guillaume Osier.

Precision

	The amount of random error in the estimation. It is measured by variance and other precision measures derived from variance.

Variance

$V(\hat{\theta})$	The variance of the estimator $\hat{\theta}$, that measures the expected variability of $\hat{\theta}$ over all possible samples s, is given by: $$V(\hat{\theta}) = \sum_{s \in S_0} p(s) \cdot \left[\hat{\theta}_s - E(\hat{\theta})\right]^2 = E(\hat{\theta}^2) - E(\hat{\theta})^2 .$$ It incorporates sampling variability (sampling errors) and (random) variability from non-sampling errors.
$\hat{V}(\hat{\theta})$	Estimator of the variance $V(\hat{\theta})$, is a function of the random set $\tilde{S}$.
$\hat{V}_s(\hat{\theta})$	Value taken by estimator $\hat{V}(\hat{\theta})$ on a specific sample s: this value provides an estimate (and is not an estimator) of the variance $V(\hat{\theta})$: $$\hat{V}_s(\hat{\theta}) = \hat{V}(\hat{\theta})(s).$$ $\hat{V}_s(\hat{\theta})$ is a numerical value (having no variance), while $\hat{V}(\hat{\theta})$ (the estimator) is a function of the random set $\tilde{S}$.

Standard error

$SE(\hat{\theta})$	Standard error of the estimator $\hat{\theta}$, defined as the square root of the variance: $$SE(\hat{\theta}) = \sqrt{V(\hat{\theta})}.$$ It incorporates sampling variability (sampling errors) and (random) variability from non-sampling errors.
$\hat{SE}(\hat{\theta})$	Estimator of standard error $SE(\hat{\theta})$, is a function of the random set $\tilde{S}$.
$\hat{SE}_s(\hat{\theta})$	Value taken by estimator $\hat{SE}(\hat{\theta})$ on a specific sample s: this value provides an estimate (and is not an estimator) of standard error $SE(\hat{\theta})$. $\hat{SE}_s(\hat{\theta})$ *is one single numerical value which does not vary; as the variance of* $\hat{SE}_s(\hat{\theta})$ *is zero, one should avoid formulations like: 'the variance of* $\hat{SE}_s(\hat{\theta})$ *' or 'the variance of the estimate'.* $$\hat{SE}_s(\hat{\theta}) = \sqrt{\hat{V}_s(\hat{\theta})}.$$ Unlike coefficient of variation (see below), the estimate of standard error is often not quite useful for comparing precision between estimates with different measurement units or widely different means, without additional specification of the estimate.

Coefficient of variation

$CV(\hat{\theta})$	Coefficient of variation of the estimator $\hat{\theta}$. Measure of the expected variability of $\hat{\theta}$, over all possible outcomes of $\hat{\theta}$ based on all possible samples. A parameter in itself. The ratio of the standard error of the estimator $\hat{\theta}$ to its expected value. A traditional quality indicator.
$\hat{CV}(\hat{\theta})$	Estimator for $CV(\hat{\theta})$.
$\hat{CV}_s(\hat{\theta})$	Value taken by the estimator $\hat{CV}(\hat{\theta})$ on a specific sample s: this value provides an estimate (and is not an estimator) of the coefficient of variation $CV(\hat{\theta})$. *$\hat{CV}_s(\hat{\theta})$ is one single numerical value which does not vary; as the variance of $\hat{CV}_s(\hat{\theta})$ is zero, one should avoid formulations like: 'the variance of $\hat{CV}_s(\hat{\theta})$' or 'the variance of the estimate'.* $$\hat{CV}_s(\hat{\theta}) = \frac{\hat{SE}_s(\hat{\theta})}{\hat{\theta}_s} \cdot 100 = \frac{\sqrt{\hat{V}_s(\hat{\theta})}}{\hat{\theta}_s} \cdot 100.$$ It is a dimensionless number and allows comparison of precision between estimates with different measurement units or widely different means. It also allows drawing conclusions on precision of estimates, without necessarily looking at the estimates. However, *one should be careful when the estimate is close to zero or binomial or multinomial. In the latter cases, the coefficient of variation depends on the value of the estimate, and hence is high for low values of the estimate (i.e. close to 0 in binomial (0,1) case) and low for high values of the estimate (i.e. close to 1, respectively).*

Using the data set of a specific sample s, specific estimates for the following precision measures can be derived:

Confidence interval

CI	The random interval, which is likely to contain the unknown true value of a population parameter. The wider the confidence interval, the lower the precision (under a fixed confidence level). If a large number of independent samples tending to infinity are taken repeatedly from the same population, and a confidence interval is calculated from each sample, then a certain percentage (close to the chosen confidence level) of the intervals will contain the unknown true value of a population parameter. Confidence

	intervals are usually calculated so that this percentage is close to 95 %, but we can produce confidence intervals with 90 %, 99 %, 99.9 % (or other) confidence levels for the unknown parameter: $\hat{CI}_s(\hat{\theta}) = \left(\hat{\theta}_s - z_{1-\frac{a}{2}} \hat{SE}_s(\hat{\theta}), \hat{\theta}_s + z_{1-\frac{a}{2}} \hat{SE}_s(\hat{\theta}) \right)$, where:
α	the confidence level;
$z_{1-\frac{a}{2}}$	the quantile value at $1-\frac{\alpha}{2}$ of the standard normal distribution; could be replaced by $t_{n-1,1-\frac{\alpha}{2}}$ (the quantile of t distribution at $1-\frac{\alpha}{2}$ with $n-1$ degrees of freedom) for small sample sizes to improve the coverage rate of confidence interval (Särndal *et al*, 1992, p. 281).

Absolute margin of error

d	The 'radius' (or half of the width) of the confidence interval: $d = z_{1-\frac{a}{2}} \hat{SE}_s(\hat{\theta}).$ Like confidence intervals, the absolute margin of error can be defined for any desired confidence level, but usually a level of 90 %, 95 %, 99 % or 99.9 % is chosen (typically 95 %).

Relative margin of error

$d\%$	The absolute margin of error as a percent of the estimate: $d\% = \frac{z_{1-\frac{a}{2}} \hat{SE}_s(\hat{\theta})}{\hat{\theta}_s} \cdot 100.$ For example, if the estimate is equal to 50 percentage points, and the statistic has a confidence interval 'radius' of 5 percentage points, then the absolute margin of error is 5 percentage points. The relative margin of error is 10 % (because 5 percentage points are ten percent of 50 percentage points). Often, however, the distinction is not explicitly made, yet usually is apparent from the context. Just like in the case of confidence intervals, the relative margin of error can be defined for any desired confidence level, but usually a level of 90 %, 95 %, 99 % or 99.9 % is chosen (typically 95 %).

Mathematically, all the above-mentioned precision measures hold the same information about precision. If we know the values of $\hat{\theta}_s$, $z_{1-\frac{a}{2}}$ and any of the precision measures, we can compute all other precision measures (except if $\hat{\theta}_s$ =0).

Other terms used in practice are **relative standard error** and **percentage standard error.**

Relative variance	This is the square of the coefficient of variation, according to the glossary of statistical terms of the International Statistical Institute.
Relative standard error	The glossary of statistical terms of the International Statistical Institute considers the term relative variance as synonym for square of the coefficient of variation (see previous). Thus, relative standard of error becomes an equivalent term for coefficient of variation.
Percentage standard error	The same glossary mentioned above specifies that the term percentage standard deviation is the coefficient of variation. Therefore, percentage standard error should be understood as coefficient of variation or relative standard error (see above).

7.2 Design effect

This section relies on contributions from Loredana Di Consiglio, Stefano Falorsi and Guillaume Osier.

The concept of design effect was first introduced by Kish (1965) in order to measure the gain or loss of sampling efficiency resulting from the use of a 'complex' design. Basically, a complex design is any design which significantly differs from simple random sampling (Ganninger, 2009). This happens when units are selected with unequal probabilities, or when sampling design includes several stages.

The effect of these complexities on sample accuracy is well known. For example, a cluster sample is generally less accurate (i.e. has more sampling errors) than a simple random sample of the same size. The reason is that units in a cluster generally tend to be similar to each other with regard to survey characteristics. As a result, collecting information from the same cluster would cause a loss of sample accuracy compared to collecting information from units selected independently from the whole population. Although the design efficiency of a multi-stage cluster sample is generally lower than for a simple random sample of the same size, multi-stage samples have other advantages in terms of economy and operational efficiency that make them the commonly used samples in practice.

The design effect $Deff$ is the ratio of the variance of an estimator $\hat{\theta}$ under the actual sampling design to the variance that would have been obtained from a hypothetical simple random sample without replacement of the same size:

$$Deff = \frac{V(\hat{\theta})}{V_{SRSWOR}(\hat{\theta}^*)} . \qquad (7.2.1)$$

$\hat{\theta}^*$ is an 'equivalent' estimator of θ under simple random sampling without replacement.

Deff compares two strategies. One consists of a sampling design, that is, a probability distribution $\{p(s), s \in S_0\}$ over all possible samples, and an estimator $\hat{\theta}$ of the population parameter θ. The formula (7.2.1) compares the strategy consisting of the actual sampling design and the estimator $\hat{\theta}$ with another consisting of simple random sampling without replacement of same size and the estimator $\hat{\theta}^*$. The latter raises an issue regarding the denominator of (7.2.1) as, in theory, there is more than one possible estimator $\hat{\theta}^*$ of θ under simple random sampling. However, according to Gabler *et al* (2003), *the estimator $\hat{\theta}^*$ should be chosen as 'equivalent' to the estimator used for the numerator $\hat{\theta}$, that is, it should have the same structure.* The estimator $\hat{\theta}^*$ at the denominator is different, for instance, for linear statistics and ratio type statistics.

When θ is a linear parameter, for instance, the population total of a variable y:

$$\theta = Y = \sum_{i \in U} y_i = \sum_{i=1}^{N} y_i , \qquad (7.2.2)$$

then a linear estimator of θ is given by the weighted sum of the sample values of the variable y. The 'equivalent' linear estimator that would be obtained under simple random sampling without replacement and of size n is given by:

$$\hat{\theta}^* = N\bar{y} , \qquad (7.2.3)$$

where N is the size (number of elements) of the population, and $\bar{y}$ is the sample mean of y.

Therefore, *Deff* can be written as:

$$Deff = \frac{V(\hat{\theta})}{V_{SRSWOR}(N\bar{y})} = \frac{V(\hat{\theta})}{N^2(1-f)S_y{}^2/n} , \qquad (7.2.4)$$

where $f = n/N$ is the sampling rate, $1 - f$ is called the finite population correction (Cochran, 1977).

For stratified simple random sampling with proportional allocation, the design effect is given by (Cochran, 1977):

$$Deff = \frac{\sum_{h=1}^{H} N_h S_h^2}{\sum_{h=1}^{H} N_h (\bar{Y}_h - \bar{Y})^2 + \sum_{h=1}^{H} N_h S_h^2} , \qquad (7.2.5)$$

where N_h is the size (number of elements) of the population in stratum h, $\bar{Y}_h$ and S_h^2 are the population mean and variance respectively of stratum h, and $\bar{Y}$ is the overall population mean. It is clear that *Deff* is approximately the share of within-stratum variance in total variance. So the more homogenous the strata, the more efficient stratified SRS is to SRSWOR.

In the case of simple random sampling with non-response, the conditional variance of the estimator of Y given the number n_R of respondents is (Lohr, 1999):

$$V\left(\hat{Y} \mid n_R\right) = N^2\left(1-\frac{n_R}{N}\right)\frac{S_y^2}{n_R}. \quad (7.2.6)$$

The unconditional variance, i.e. after averaging over n_R, contains additional terms of order $\frac{1}{n_R^2}$. Therefore the design effect is approximately:

$$Deff \approx \frac{n}{n_R}\cdot\frac{1-\frac{n_R}{N}}{1-\frac{n}{N}} \quad (7.2.7)$$

and if the sampling fraction is small, then $Deff \approx \frac{n}{n_R} = \left(\text{responserate}\right)^{-1}$.

Therefore the smaller the response rate, the larger the design effect.

Let us assume now that θ is a non-linear parameter — for instance, the ratio of the population totals of two variables y and x. The most common way to estimate θ is to estimate the totals of y and x individually using a linear formula, and then take the ratio of the two estimators as an estimator of θ (Särndal *et al*, 1992).

Under simple random sampling, this leads to an estimator $\hat{\theta}^*$ which can be expressed as the ratio of the sample means $\bar{y}$ and $\bar{x}$ of y and x, respectively:

$$\hat{\theta}^* = \frac{\bar{y}}{\bar{x}}. \quad (7.2.8)$$

The variance of $\hat{\theta}^*$ under simple random sampling without replacement of size n can be calculated by the Taylor linearisation technique (Wolter, 2007):

$$V_{SRSWOR}\left(\hat{\theta}^*\right) = V_{SRSWOR}\left(\frac{\bar{y}}{\bar{x}}\right) \approx N^2\left(1-f\right)S_U^2/n, \quad (7.2.9)$$

where

$S_U^2 = \frac{1}{N-1}\sum_{i=1}^{N}u_i^2$ and $u_i = \frac{1}{X}\left(y_i - \theta x_i\right)$.

According to Lehtonen and Pahkinen (1996), the reference design to which we compare the actual one can be simple random sampling with or without replacement. For simple random sampling with replacement, and assuming that the estimator $\hat{\theta}$ has a linear form, the design effect can be written as:

$$Deff = \frac{V\left(\hat{\theta}\right)}{V_{SRS}\left(\hat{\theta}^*\right)} = \frac{V\left(\hat{\theta}\right)}{N^2\sigma_y^2/n}, \quad (7.2.10)$$

where $\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)^2$.

σ_y^2 is an alternative measure of the dispersion of the variable of interest y over the reference population U. Assuming that $y_1, y_2, \ldots, y_N$ are independent realisations of random variables with the same mean and variance ('super-population' model), the use of N in the denominator (rather than $N-1$) makes σ^2 slightly biased as an estimator of variance.

When the population size N is large enough we have:

$$\frac{\sigma_y^2}{n} = \frac{\frac{N-1}{N}S_y^2}{n} = \frac{N-1}{N}\cdot\frac{1}{1-f}\cdot(1-f)\frac{S_y^2}{n} \approx (1-f)\frac{S_y^2}{n}. \quad (7.2.11)$$

Therefore, (7.2.10) and (7.2.4) lead to almost the same results. *In practice, however, we recommend using (7.2.4) as this Deff is not affected by population size.*

We usually refer to the square root of *Deff* as *Deft*. This is the ratio of standard errors between two strategies: an actual one and a hypothetical one that consists of simple random sampling without replacement and of the same size. Verma (1982) proposed that *Deff* be referred to as 'design effect' and *Deft* as 'design factor', though Lê and Verma (1997) noted that this proposal has not been widely adopted. Kish (1995) used 'design effect' as a generic term to encompass not only both *Deff* and *Deft*, but also more general concepts in relation to the effect that a design might have on accuracy. It is also possible to define *Deff* as a ratio of Mean Square Errors (MSEs) rather than variances.

Kish (1995) pointed out that design effects are only tools, rather than a theory or even a method. However, since its first introduction, *Deff* has been used extensively in the field of survey sampling, at both design and analysis stages. According to Gabler *et al* (2003), the primary use of these tools is to convey information about survey design. For this, we need somehow to estimate (predict) the likely design effects.

A model-based design effect may be used for this purpose. In general, for a multi-stage cluster sampling design, *Deff* can be predicted directly as follows:

$$\hat{Deff}_{PRIOR}(\hat{\theta}) = \left[n\sum_{l=1}^{L} n_l w_l^2 \Big/ \left(\sum_{l=1}^{L} n_l w_l\right)^2 \right] \cdot \left[1 + (\bar{b} - 1)\rho^*\right], \quad (7.2.12) \text{ (Gabler } et\ al\text{, 1999)}$$

where

l = weighting class $(l = 1 \ldots L)$,

n_l = number of sample observations in weighting class l,

ρ^* = intra-cluster correlation coefficient (see *Appendix 7.1*),

$\bar{b}$ = mean number of sample observations per cluster,

$\left[n\sum_{l=1}^{L} n_l w_l^2 \Big/ \left(\sum_{l=1}^{L} n_l w_l\right)^2 \right] = \hat{Deff}_w(\hat{\theta})$ is the design effect due to unequal weighting,

$\left[1 + (\bar{b} - 1)\rho^*\right] = \hat{Deff}_c(\hat{\theta})$ is the design effect due to clustering.

The above formula (7.2.12) expresses the overall design effect as a product of the design effect due to unequal weighting and the design effect due to clustering. Ganninger (2006)

presents the calculation of the design effect due to unequal weighting and of the design effect due to clustering and illustrates it by numerical examples. Calculation of the design effect due to unequal weighting requires the inclusion probabilities of each sampling unit to be known at every sampling stage. On the other hand, calculation of the design effect due to clustering is based on the above model-based estimator, which needs data on average cluster size (or, if cluster size shows great variation, then an alternative measure of size is used) and on the intra-cluster correlation coefficient, which requires the selection of specific core variables.

Basically, the design effect due to clustering depends on:

- the variable under study, i.e. the same sampling design may lead to different design effects that depend on the degree of autocorrelation or similarities of units within a cluster, as measured by the intra-cluster correlation coefficient ρ^*. We can even imagine some extreme cases where ρ^* is positive with respect to one variable, thus leading to a $Deff$ with a value higher than 1, and negative for another variable, so that $Deff$ would be less than 1. Values of design effect can differ greatly across variables and sub-populations within the same survey (Eurostat, 2010a);
- the mean number of sample observations $\bar{b}$ per cluster.

The formula (7.2.12) does not take into account the effect of stratification on accuracy. However, in most cases, stratification improves the accuracy, so not taking it into account would result in a more conservative estimator. This might turn out to be a problem, though, when we oversample certain small sub-populations, as the overall accuracy might deteriorate. However, loss of accuracy is generally limited.

If we consider the value of autocorrelation of the main variables of interest as transferrable information, then using formula (7.2.12) makes it easy to derive the design effect associated with a given cluster design structure, and consequently permits us to calculate the variance for the estimator for the new sampling strategy.

With a view to calculating the design effect for general multi-stage sampling designs with stratification of primary sampling units and selection of units at different stages with probability proportional to size without replacement (πpswor) mostly adopted in large scale surveys, it is not unusual to consider just the variability between primary sampling units (PSUs). This choice is based on the hypothesis of selecting PSUs with probability proportional to size and with replacement (πpswr). This means that simple estimation formulae for sampling variances that do not involve the second-order inclusion probabilities between PSUs can be applied in this simplified framework. Each stratified multi-stage sampling design can be approximated with a πpswr selection in each stratum, whereby PSUs are considered as ultimate clusters, i.e. the aggregate of all elementary units selected from the same PSU. Therefore, all second, third and successive stage units selected from the PSU are treated as a single unit (Särndal *et al*, 1992).

Then, in this simplified but general multi-stage sampling context in which larger PSUs are certainly selected, let us denote h as a generic stratum, with H_{sr} as the total number of *self-representing*[41] (*sr*) strata, H_{nsr} as the total number of *non self-representing* (*nrs*) strata and H as the overall number of strata. Index i will then denote a generic PSU, where for stratum h, N_h and n_h represent population and sample size of PSUs in the strata, being for *sr* strata $N_h = n_h = 1$, while index j denotes a generic elementary unit. For stratum h:

[41] Strata in which no sub-sampling takes place at the first stage.

$$M_{sr,h} = \sum_{i=1}^{N_h} M_{sr,hi} \quad \text{and} \quad m_{sr,h} = \sum_{i=1}^{n_h} m_{sr,hi}$$

$$M_{nsr,h} = \sum_{i=1}^{N_h} M_{nsr,hi} \quad \text{and} \quad m_{nsr,h} = \sum_{i=1}^{n_h} m_{nsr,hi} \ ,$$

where $M_{sr,hi}$, $m_{sr,hi}$, $M_{nsr,hi}$ and $m_{nsr,hi}$ are the population and sample sizes of elementary units clustered in the i-th PSU of stratum h for *sr* and *nsr* respectively.

If the same number of elementary units in each *PSU* is selected and under the hypothesis that $S^2_{sr,h} = S^2$ for $h = 1,\ldots,H_{sr}$ and $S^2_{nsr,h} = S^2$ for $h = 1,\ldots,H_{nsr}$, then the design effect depends on the intra-class correlation coefficients $\rho_{sr,h}$ and $\rho_{nsr,h}$ for *sr* and *nsr*. It is given by:

$$\hat{Deff}(\hat{Y}) = \frac{m}{M}\left[\sum_{h=1}^{H_{sr}} \frac{M^2_{sr,h}}{m_{sr,h}}\left[1+\rho_{sr,h}\left(\overline{m}_{sr,h}-1\right)\right]+\sum_{h=1}^{H_{nsr}} \frac{M^2_{nsr,h}}{m_{nsr,h}}\left[1+\rho_{nsr,h}\left(\overline{m}_{nsr,h}-1\right)\right]\right], \qquad (7.2.13)$$

where $\overline{m}_{sr,h}$ and $\overline{m}_{nsr,h}$ are the mean number of elementary units at PSU level for the h^{th} stratum in *sr* and in *nsr* domain. In large scale surveys, the design effect takes into account the impact of non-response. If we denote τ as the overall response rate and τ_h as the h^{th} stratum response rate, then the design effect formula is given by:

$$\hat{Deff}_{\tau}(\hat{Y}) = \frac{m\,\tau}{M}\left[\sum_{h=1}^{H_{sr}} \frac{M^2_{sr,h}}{m_{sr,h}\tau_h}\left[1+\rho_{sr,h}\left(\overline{m}_{sr,h}\,\tau_h-1\right)\right]+\sum_{h=1}^{H_{nsr}} \frac{M^2_{nsr,h}}{m_{nsr,h}\,\tau_h}\left[1+\rho_{nsr,h}\left(\overline{m}_{nsr,h}\,\tau_h-1\right)\right]\right]$$

(7.2.14).

The expression $\hat{Deff}(\hat{Y})$ depends on some known quantities that are related to the planned sampling design, i.e. m, $m_{sr,h}$, $M_{sr,h}$, $m_{nsr,h}$, $M_{nsr,h}$, $\overline{m}_{sr,h}$ and $\overline{m}_{nsr,h}$, and on some parameters that are unknown in the planning stage and which depend on the characteristics of the target population, i.e. $\rho_{sr,h}$ and $\rho_{nsr,h}$, and on the response rates, i.e. τ and τ_h, for $\hat{Deff}_{\tau}(\hat{Y})$. Under the hypothesis that the values of these unknown quantities are stable over time, these can be estimated using data from previous surveys. *For this reason, the difference between estimates and actual values should be evaluated once the survey has been carried out.*

Nikolaidis (2008) specifies that the design effect measures the impact of the sampling design (clustering, stratification) and unequal selection probabilities (random weighting) on the variance of estimates. Park and Lee (2004) mention that Kish (1965) considered cases where the unequal weights arise from 'haphazard' or 'random' sources such as frame problems or non-response adjustments. Kish (1987) proposes a decomposition model of the overall design effect as a product of two individual components — clustering and unequal weighting — while Park *et al* (2003) consider a three-factor decomposition model.

Such prediction has to be distinguished from estimating the design effect subsequent to a survey, on the basis of data collected. In that case, the survey data can be used to input

information into the model. Often, the model is entirely discarded and a design-based estimate of the design effect is calculated.

A design-based estimation of the design effect relies on separate calculations of the numerator and the denominator. Standard methods can provide a design-based estimator of the numerator, that is, the variance under the actual sampling design.

When it comes to estimating the variance that would be obtained under simple random sampling without replacement with the same sample size n, a common mistake is to apply the classical variance estimator under simple random sampling:

$$\hat{V}_{SRSWOR}\left(\hat{\theta}^*\right) = N^2\left(1-f\right)\frac{s^2}{n}, \quad (7.2.15)$$

where s^2 refers to the sample variance of the study variable[42] and $f = n/N$ is the sampling rate. The problem with (7.2.15) is that the formula provides a design-unbiased estimator of the variance under simple random sampling only when the sample is a simple random one. In line with Gabler *et al* (2003), if the sample units had unequal inclusion probabilities (e.g. with disproportionate stratification), then (7.2.15) could no longer be applied as it would lead to biased estimators of the variance under simple random sampling, that is, the denominator of *Deff* . Although the denominator of *Deff* accounts for the variance under simple random sampling, it generally has to be estimated from a sample, which may be far from a simple random one.

However, *when the sample is self-weighted, that is, all units in the population have the same probability of selection (e.g. stratified sampling with proportional allocation), then (7.2.15) still provides a consistent and asymptotically unbiased estimator of the variance* (Ardilly and Tillé, 2005). *When sample units have unequal selection probabilities, the following should be used* (Ardilly and Osier, 2007):

$$\hat{V}_{SRSWOR}\left(N\bar{y}\right) = \frac{1}{n}\left(1-\frac{n-1}{\hat{N}-1}\right)\cdot\hat{N}\cdot\sum_{i\in s} w_i\left(y_i-\bar{y}_s\right)^2, \quad (7.2.16)$$

where

$\hat{N} = \sum_{i\in s}\frac{1}{\pi_i}$ = estimated size of the population,

$\bar{y}_s = \frac{\sum_{i\in s} w_i y_i}{\sum_{i\in s} w_i}$ = weighted sample mean of the study variable.

When the size of population is large enough, (7.2.16) leads to an almost unbiased estimator of the variance under simple random sampling. (7.2.16) is used in the POULPE variance estimation software, developed by the French NSI (INSEE), in order to yield *Deff* estimates. The formula is valid even for multi-phase sampling. For more information, see Ardilly and Osier (2007). While this is easy to implement for means and totals, it is difficult to find the right formula for the denominator of the design effect for any other type of estimate, e.g. ratios, regression coefficients and difference of proportions.

[42] If the estimator is not linear, the study variable is replaced by the 'linearised' variable, that is, the variable which results from a linear approximation of the estimator (Osier, 2009).

In the design effect formula, the variance at the denominator should be calculated in the same units (individuals or households) as for the variance at the numerator in household surveys. This is done for those sampling units which are relevant for statistics (e.g. in the LFS: unemployment indicators for individuals and job-related indicators for households).

Di Consiglio and Falorsi (2010) introduced the concept of the **estimator effect**, which seeks to measure the efficiency of an estimator with respect to the Horvitz-Thompson estimator. For example, calibration estimators (Deville and Särndal, 1992) are often used in practice to deal with unit non-response, thereby increasing sample efficiency (reducing standard errors), or for coherence purposes with external sources.

Consider an estimator $\hat{\theta}$ of a parameter θ. Let us assume that $\hat{\theta}$ is a linear estimator, that is, a weighted sum of the sample values. The extension to non-linear estimators can be achieved by linearisation (Osier, 2009). Basically, the estimator effect of $\hat{\theta}$ is the ratio between its variance under the actual sampling design and the variance of the Horvitz-Thompson estimator $\hat{\theta}_{HT}$ under the same sampling design. It measures the increase (or decrease) in efficiency obtained with the proposed final estimator:

$$Eff\_est = \frac{V(\hat{\theta})}{V(\hat{\theta}_{HT})} \quad . \quad (7.2.17)$$

To obtain a better evaluation of the actual variance with the proposed sampling strategy, we also need to estimate the effect of using a different estimator. Based on (7.2.17), the overall design effect can be written as:

$$Deff = \frac{V(\hat{\theta})}{V_{SRSWOR}(\hat{\theta}^*)} = \frac{V(\hat{\theta})}{V(\hat{\theta}_{HT})} \cdot \frac{V(\hat{\theta}_{HT})}{V_{SRSWOR}(\hat{\theta}^*)} = Deff_1 \cdot Deff_2 . \quad (7.2.18)$$

$Deff_1 = \frac{V(\hat{\theta})}{V(\hat{\theta}_{HT})} = Eff\_est$ is the estimator effect of $\hat{\theta}$.

$Deff_2 = \frac{V(\hat{\theta}_{HT})}{V_{SRSWOR}(\hat{\theta}^*)}$ is the sampling design effect, that is, the effect of sampling features (stratification, unequal weighting, clustering) on accuracy.

If successive waves of the same survey use the same auxiliary variables for calibration and if between these waves the relationship between the auxiliary variables and the target variable remains constant, then we can also suppose the estimator effect to be constant over the successive waves, and so we are able to apply a previously estimated value. After the survey wave is carried out, we can check whether the estimated estimator effect based on the new sample data is close to the assumed (previous) value, and correct the assumed value if necessary.

Liberts (2012) disaggregated $Deff$ into three components, $Deff = Deff_1 \cdot Deff_2 \cdot Deff_3$ with $Deff_1$ and $Deff_2$ as before and $Deff_3 = \frac{1}{(1-\hat{OC}_r)(1-\hat{NR}_r)}$ to represent the **effect of non-sampling errors** (such as non-response and over-coverage). $Deff_3$ is always greater than 1, because non-sampling errors are unavoidable. In the last equation, $\hat{OC}_r$ is the expected

unweighted over-coverage rate (Eurostat, 2010e, p.5) and $\hat{NR}_r$ the expected unweighted non-response rate (Eurostat, 2010e, p.8).

Measurement and processing errors also affect *Deff* but it is difficult to quantify their impact. Moreover, processing errors are unlikely to be clustered, so they may reduce the *Deff*, whereas coverage errors may be clustered, so they may increase the *Deff*.

In the European Social Survey, *design effects are mainly used to calculate the minimum required sample size.* The European Social Survey requires a minimum effective sample size of 1500 in countries having more than two million inhabitants, 800 otherwise. Countries have to estimate *Deff* prior to the survey, by using (7.2.12) for instance, and then use that value as an inflator for the minimum effective sample size, in order to obtain the minimum required sample size (Ganninger, 2006).

The design effect can be determined under the model-based approach (7.2.12 or 7.2.13/7.2.14) or under the design-based approach. However, we recommend using the term 'design effect' with its coverage clearly defined in terms of clustering, unequal weighting, etc., and in terms of whether it considers full response — theoretical situation or a variance adjusted for non-response, etc. Using the term 'design effect' without a clear definition will lead to misunderstandings and very different interpretations. The precision requirements of EU-SILC (EP and Council Regulation No 1177/2003 of 16 June 2003 and accompanying technical documents) suffer from precisely this lack of a clear definition of the 'design effect'.

7.3 Metadata template for variance estimation

This metadata template has been prepared for surveys based on samples. In case of censuses, only a part of the questions of the metadata template are relevant and have to be answered.

The metadata template was conceived to be as comprehensive as possible, in order to be relevant for several statistical domains. As mentioned in chapter 5, *in order to use it as an element of compliance assessment strategy for a specific statistical domain (survey), it should be adapted to the specific features of that statistical domain.*

In case the fully centralised approach is implemented, the metadata template would be particularly useful and is recommended in order to collect clear and detailed information on the sampling designs.

A. Frame population
1. What is the frame population used? *Please mention the frame population and the units listed therein (e.g. districts, municipalities, addresses, households, persons, telephone numbers, etc.).* *If the sample is selected from a sample of another survey, from a micro-census or from a master sample (in the case of multi-phase sampling designs), then please mention the frame population used for the other survey/the micro-census/the master sample.* *If more than one sampling frame is used e.g. a sampling frame for each sampling stage or a sampling frame for each national region, then please mention all of them.*

Please describe if different frames are used to draw the sample and to gross up.

Please mention if RDD (Random Digit Dialling) is used.

2. Is the sample drawn from another survey sample?

☐ Yes

If yes, please name the survey

If yes, then the sampling stages used to select the other survey sample have to be further included in the description of sampling design.

If yes, then we have a case of multi-phase sampling (see Section 3.1 and Appendix 7.1 for more information).

☐ No

3. Is the sample drawn from the micro-census?

☐ Yes

If yes, then the sampling stages used to select the micro-census have to be further included in the description of sampling design.

If yes, then we have a case of multi-phase sampling (see Section 3.1 and Appendix 7.1 for more information).

☐ No

4. Is the sample drawn from the master sample?

☐ Yes

If yes, then the sampling stages used to select the master sample have to be further included in the description of sampling design.

If yes, then we have a case of multi-phase sampling (see Section 3.1 and Appendix 7.1 for more information).

☐ No

5. What are the main errors in the frame?

Frame errors refer to under-coverage, over-coverage and multiple listings. Over-coverage and multiple listings are sources of additional variability of the estimator (see Section 3.2 and Appendix 7.1 for more information).

If more than one sampling frame is used e.g. a sampling frame per sampling stage/phase or a sampling frame for each national region, then please mention the errors in all of them.

In the first column (Y/N), please mention whether or not the specified error (over-coverage/multiple listings) occurs. In the second column (%), please indicate the error rate (or an estimate thereof) as a percentage of the total number of records in the frame. If not possible, please provide a qualitative assessment in the third column.

	Y/N	%	Qualitative assessment
Over-coverage	☐		

Multiple listings ☐

B. Target indicators and reporting domains

6. What are the main target indicators for which precision estimates are required?

Please specify the type of each indicator e.g. total, mean, proportion, ratio, other smooth non-linear statistics e.g. regression coefficients, correlation coefficients, non-smooth statistics e.g. quantiles, poverty rates, etc. For proportions and ratios, please see those definitions given for the variance estimation context in Section 2.2.

Main target indicator	Type of indicator

(insert additional rows if needed)

7. Are the precision estimates required over domains?

In this case, please consider both planned and unplanned domains. See Section 2.3 for definitions of planned and unplanned domains.

☐ Yes

If yes, what are these domains? Please specify if domains are planned or unplanned.

Domain	Planned/unplanned

(insert additional rows if needed)

☐ No

C. Sampling design

8. Is the sampling design a probability sampling design?

A probability sampling design ensures known probabilities for units selected. In practice, non-response generally makes samples depart from the probability ones. However, the point here is to report on whether or not the gross sample (net sample plus non-respondents) has been selected in a probability way.

☐ Yes

☐ No

9. What is the number of sampling stages?

If the survey sample is selected from a sample of another survey, from the micro-census or from the master sample, then please include the number of sampling stages from all sampling

phases into the total number of sampling stages.

If there are differences in the same country with regard to the number of sampling stages for different population groups, e.g. one-stage sampling in urban areas and two-stage sampling in rural areas, then report the number of sampling stages for each of the population groups.

10. What is the sampling unit at stage 1 (the primary sampling unit PSU)?

Examples: census enumeration areas, sections, municipalities, communes, villages, settlements, households, individuals, etc.

If there are differences in the same country with regard to the type of PSUs, e.g. households as PSUs in urban area and villages as PSUs (and households as SSUs) in rural areas, then report the relevant sampling unit at stage 1 for each of the population groups. Please do this also for the sampling units at further stages at the next questions.

11. What is the sampling unit at stage 2 (the secondary sampling unit SSU)?

Examples: dwellings, households, individuals, etc.

(Please insert additional rows when needed for additional stages)

12. What is the sampling unit at the ultimate stage?

Examples: dwellings, households, individuals.

13. What are the interviewed units?

Interviewed units are units from which data are collected. The interviewed unit can be different from the ultimate sampling unit.

For instance, the sampling unit at an ultimate stage can be a household and the interviewed unit can be an individual (all eligible individuals in the household are interviewed — this is a cluster sampling).

Furthermore, the sampling unit at the ultimate stage can also be an individual and the interviewed unit can be all eligible individuals in the same household. This is an indirect cluster sampling. See Section 3.1 and Appendix 7.1 for more information.

14. Is there (explicit) stratification at stage 1?

If there are differences as regards stratification at stage 1 between population groups (e.g. rural/urban, etc.), then please answer separately for each case.

☐ Yes
If yes, what are the stratification variables at stage 1?
Examples:
-region/ province/ county/ district/ code of administrative territories;
-size/ population density/ degree of urbanisation;
-type of municipality/ settlement;
-type of residence: urban/ rural;
-age, gender, etc.
☐ No

15. What is the sampling method at stage 1?

The sampling method (for the sampling units) refers to the way the sample is selected. For example, the sampling method can be a simple random sampling, whereby all samples are given the same probability of selection. Other possible methods include systematic sampling with equal or unequal probabilities, other proportional-to-size sampling (πps), etc.

Please mention if the systematic sampling has stratification effect (gives rise to implicit stratification). See Section 3.4 for more information about implicit stratification. If there are differences as regards the sampling design at stage 1 between population groups (e.g. rural/urban, etc.), then please answer separately for each case.

☐ Exhaustive selection
☐ Simple random sampling
☐ Systematic sampling with equal probabilities
  ☐ With stratification effect, please mention the related auxiliary variable
  ☐ Without stratification effect
☐ Systematic sampling with probabilities proportional-to-size
  ☐ With stratification effect, please mention the related auxiliary variable
  ☐ Without stratification effect
☐ Other proportional-to-size (πps) sampling, please indicate
☐ Other, please indicate

For stage 1 it is important to know if there are self-representing primary sampling units (with probability of selection equal to 1). Please mention if this is the case:

(Please insert additional rows when needed for intermediate stages, for the questions on explicit stratification and sampling method.)

16. Is there (explicit) stratification at the ultimate stage?

If there are differences as regards stratification at the ultimate stage between population groups (e.g. rural/urban, etc.), then please answer separately for each case.

☐ Yes
If yes, what are the stratification variables at the ultimate stage?
☐ No

17. What is the sampling method at the ultimate stage?

If there are differences as regards the sampling method (for the sampling units) at the ultimate stage between population groups (e.g. rural/urban, etc.), then please answer separately for each case.

☐ Exhaustive selection
☐ Simple random sampling
☐ Systematic sampling with equal probabilities
  ☐ With stratification effect, please mention the related auxiliary variable
  ☐ Without stratification effect
☐ Systematic sampling with probabilities proportional-to-size
  ☐ With stratification effect, please mention the related auxiliary variable
  ☐ Without stratification effect
☐ Other proportional-to-size (πps) sampling, please indicate
☐ Other, please indicate:

18. What is the sample allocation among strata?

In case the sample has been stratified, then sample allocation among the strata can be either equal, proportional to the stratum sizes, or optimal (Neyman) in that it yields estimators with the lowest standard error for certain variables of interest. Compromise allocations can also be used, in order to ensure an acceptable level of accuracy both for national and regional estimates.

If there are differences as regards to allocation between population groups (e.g. rural/urban, etc.), then please answer separately for each case.

☐ Equal allocation, for stratification at stage(s)
☐ Proportional allocation, for stratification at stage(s)
☐ Optimal (Neyman) allocation, for stratification at stage(s)
☐ Compromise allocation, for stratification at stage(s)
☐ Other (please mention) , for stratification at stage(s)

19. Does the survey have a longitudinal component (i.e. the survey collects data from the same sample elements on multiple occasions over time)?

☐ Yes, please indicate:
If the survey sample is based on rotation groups, please specify:
How many rotation groups are considered?
Are the rotation groups of equal sizes?
☐ Yes
☐ No, please provide more information:
What is the frequency of rotation of groups?
☐ Monthly
☐ Quarterly
☐ Annual
☐ Other, please specify:
How are new rotation groups selected?
Please specify if the selection of a new rotation group is made from the same e.g. primary sampling units used in the previous wave or a selection of new e.g. primary sampling units is made for each wave. Please provide any additional relevant information.
If the survey sample is drawn from another survey sample/micro-census/master sample, then does the rotation take place at the level of the other survey sample/micro-census/master sample?

☐ Yes
☐ No

☐ No

20. Please provide any additional information on the sampling design.

D. Weights

21. How are design weights calculated?

Design weights are defined as the inverse of the units' selection probabilities. Under full response, design weights provide unbiased estimates for linear parameters. This point aims at collecting textual information on the way design weights are calculated. In case the approach departed from the usual one that consists of taking the inverse of the inclusion probabilities, then the latter should be explained.

22. Is balanced sampling used?

A sampling design is called balanced if it ensures that the Horvitz-Thompson estimators of some 'balancing' variables are equal to the known totals.

☐ Yes
If yes, what are the balancing variables?
Please describe the method:
☐ No

23. Are there very low numbers of respondents for specific strata which make it difficult to calculate precision measures?

☐ Yes
If yes, please give more details and specify the techniques used (e.g. collapse of strata, etc.)
☐ No

24. Has re-weighting for unit non-response been performed?

☐ Yes
Which method has been used?
If yes, the method used to determine the correction factors should be explained: re-weighted Horvitz-Thompson estimator, ratio estimation, regression estimation, etc. Please indicate if response homogeneity groups have been created.
☐ No

25. Has adjustment to external data sources been performed?

Generally, samples are adjusted to external data sources in order to make their accuracy better. For instance, the calibration technique aims at calculating new weights which provide error-free estimates for a certain number of characteristics. If the characteristics are strongly

correlated with the variables of interest, then the level of accuracy for most of the survey estimates is improved.

☐ Yes

If yes, please list the calibration variables used and their sources:

Examples of calibration variables:
-region/province/administrative territories, degree of urbanisation, settlement size/type, household size, household composition/structure, place of residence (urban/rural), etc. (for households);
-region, settlement size, urban density, place of residence (urban/rural), gender, type of settlement/family, age (band), marital status, level of education, employment status, professional activity, gross income, nationality, citizenship, etc. (for individuals).
Example of sources: population register, updated population register, etc.

☐ No

26. Has any other adjustment been performed?

Further adjustments might be done to correct coverage errors or measurement errors. Besides, in order to avoid extreme weights, the distribution might be trimmed or top-coded: trimming refers to the removal of observations that are greater than a certain threshold, while top-coding consists of recoding observations greater than a given maximum to this maximum value.

☐ Yes

If yes, please describe the adjustment:

☐ No

27. Is the sample self-weighted?

The elements of a self-weighted sample have the same probability of selection. For instance, simple random samples are self-weighted. In practice, however, non-response generally makes samples depart from self-weighting ones. The point here is to report on whether or not the sample is nearly self-weighted (subjective assessment). In this regard, summary measures on the weight distribution might be useful (e.g. the coefficient of variation of the weight distribution).

This question is relevant to know if the inclusion probabilities are equal or unequal. This makes a difference on the estimator of the variance under simple random sampling that should be used at the denominator of the design effect (see Appendix 7.2).

☐ Yes
☐ No

E. Substitution and imputation

28. Has substitution been used for the main target indicators for which precision estimates are required?

Substitution means replacement of a sampling unit by a new one.

☐ Yes

If yes, please mention the main target indicators (in case there are several):

☐ No

29. What was the substitution rate for the main target indicators?

Please give the proportion (%) of the sampling units that were replaced by substitutes.

Main target indicator	Substitution rate (%)

(insert additional rows if needed)

30. On which criterion has the selection of the substituted units been based?

The choice of the substituted units may rely on statistical considerations (a substituted unit should be similar to the original unit with respect to certain characteristics) but also on administrative considerations.

Please explain:

31. Are the main target indicators for which precision estimates are required affected by item-response?

☐ Yes, please indicate the main target indicators (in case there are several) and the item non-response rate, which gives an assessment of the influence of this on variability:

Have the main target variables been imputed?

☐ Yes

☐ No

☐ No

32. What imputation methods have been used?

(Multiple choices possible)

☐ **Deductive imputation**

An exact value can be derived as a known function of certain characteristics (e.g. the value received for a family allowance is a known function of certain characteristics like income class, age of children, etc. As soon as those characteristics are known, it becomes possible to calculate the value of a family allowance without error.)

Deterministic imputation

Deterministic imputation leads to estimators with no random component, that is, if the imputation were to be re-conducted, the outcome would be the same.

☐ Mean/Median

☐ Mean/Median by class

☐ Regression-based

☐ Donor
☐ Other (please specify):

Random imputation
Random imputation leads to estimators with a random component, that is, if the imputation were re-conducted, it would lead to a different result.
☐ Hot-deck
☐ Cold-deck
☐ Simulated residuals
☐ Other (please specify):

☐ **Multiple imputation**
Multiple imputation methods offer the possibility of deriving variance estimators by taking imputation into account. In multiple imputation each missing value is replaced (instead of a single value) with a set of plausible values that represent the uncertainty of the right value to impute. The incorporation of imputation variance can be easily achieved based on the variability of estimates among the multiply imputed data sets.

33. What was the overall imputation rate for each of the main target indicators (cf q. 6)?

For each of the target indicators that you listed in your response to question 6, please report the proportion of observations that are imputed values. Moreover, if applicable, please report the share of the estimate that is contributed by the imputed values.

Main target indicator	Imp. rate (% of observations)	Imp. rate (share of estimate — %)

(insert additional rows if needed)

F. Variance estimation methods and tools

This section is concerned with variance estimation methods and software tools usually used by countries.

34. What are the main variance estimation methods used?

Please see the definitions of methods in Section 3.3.

For each method, please specify the name of the main indicators the method was applied to (out of those specified for question 6) and the type of parameter. We can distinguish between the following types of parameters:

- *parameters at specific time points;*
- *parameters defined as averages over a period of time;*
- *net change between two time points, i.e. to provide estimates of the difference, the ratio, or some other measure of change, between parameters at different time points;*
- *gross change, i.e. aggregates of change at the element level between time points.*

See more details in Sections 3.3 and 3.7, and in Appendix 7.4.

☐ **Analytical method:**

Formula or reference in literature	**Name of indicator**	**Type of parameter**

Please add rows if needed.

☐ **Linearisation method:**

Linearisation method	**Name of indicator**	**Type of parameter**

☐ **Taylor linearisation**
☐ **Linearisation based on influence functions**
☐ Other, please specify:
Please add rows if needed.

☐ **Replication method:**

Replication method	**Name of indicator**	**Type of parameter**

☐ **Jackknife**
☐ **Bootstrap**
☐ **Balanced repeated replication/Balanced half-samples**
☐ **Random groups**
Please add rows if needed.

☐ **Other e.g. generalised variance functions, please specify:**

35. Please briefly describe the method(s):

36. What are the main variance estimation tools used?

☐ CLAN
☐ GENESEES
☐ SUDAAN
☐ POULPE
☐ CALJACK
☐ BASCULA
☐ ReGenesees
Other, please specify:

37. Do the methods/tools for variance estimation take into account the effect of:

☐ **unit non-response**?
The variance estimator $\hat{V}(\hat{\theta})$ has to be adjusted to take unit non-response into account. Different methods can be used: methods based on the assumption that respondents are missing at random or completely at random within e.g. strata or constructed response homogeneity groups, methods using the two-phase approach, etc. See Section 3.4 for more information.

If yes, please indicate:

☐ **imputation**?

Imputation variance can be estimated if multiple imputation is used.

Replication and analytical methods can be used to incorporate imputation into variance estimation.

Deville and Särndal (1994) proposed a method for the regression imputed Horvitz-Thompson estimator.

See Section 3.4 for more information.

If yes, please indicate:

☐ **coverage errors (over-coverage, multiple listings)?**

Methodology of domain estimation can be used. Target population has to be defined as a domain of the frame population.

The related loss of precision can be quantified.

See Section 3.4 for more information.

If yes, please indicate:

☐ **implicit stratification?**

One way to consider implicit stratification is to define explicit strata, from which each of an independent sample is supposed to have been selected.

Other methods using analytical formulae are available. See Section 3.4.

If yes, please indicate:

☐ **rotating samples?**

In case of rotating sample schemes, the overlap of samples between e.g. successive quarters reduces the precision of the average of estimates from e.g. quarterly samples and increases the precision for e.g. the quarter-to-quarter estimates of change.

See Section 3.7.

If yes, please indicate:

☐ **calibration?**

Methods to account for the effect of calibration on variance should be used. Deville and Särndal method (1992) is presented in Section 3.4.

If yes, please indicate:

38. Please provide the main references in literature for variance estimation methods and software tools used

G. Availability of specific information which can be used to estimate standard errors by Eurostat (and other data users)

This section aims at collecting information on the possibility of estimating standard errors by Eurostat (and other data users) in the future.

39. Are design effects (*Deffs*) systematically calculated along with variance estimates?

☐ Yes

Please briefly describe the method:

This point deals with the estimation method for Design Effect (Deff). It is essential to explain how the variance under simple random sampling (denominator of the Deff) is estimated. Please see Appendix 7.2 for more details.

☐ No

40. If a replication method is used in the NSIs to estimate standard errors, then can replicate weights be transmitted to Eurostat together with the microdata?

This question aims to contribute to the assessment of the feasibility of estimating standard errors under an integrated approach (see chapter 4).

Please comment:

41. If generalised variance functions are used in the NSIs to estimate standard errors, then can the parameters and the functions be reported to Eurostat and would they be sufficient for estimating standard errors for all indicators and breakdowns?

This question aims at contributing towards assessment of the feasibility in order to estimate standard errors under an integrated approach (see chapter 4).

Please comment:

42. Can microdata be transmitted to Eurostat together with the following specific variables at record level?
- the stratum to which the ultimate sampling unit belongs;
- the primary, the secondary, etc. sampling units to which the ultimate sampling unit belongs;
- in case systematic sampling is used at any sampling stage, the order of selection of the primary, the secondary, etc. sampling units;
- the final sampling weight of the units used in the estimation. The final sampling weight should ideally incorporate adjustment due to non-response, calibration, etc. so that variance estimates can reflect their effects.

This question aims at contributing towards the assessment of the feasibility in order to estimate standard errors under a fully centralised approach (see chapter 4).

Please comment:

7.4 Suitability of variance estimation methods for sampling designs and types of statistics

This matrix lists some recommended methods which are suited to different sampling designs and types of statistics. The list of suitable methods does not claim to be complete. *This matrix also lists methods which are not recommended (bad practices).*

✔ Suitable method (recommended for use)
! Unsuitable method (not recommended for use)

Sampling designs	**Type of statistics**			
	Linear statistics (e.g. totals, means, proportions*)	**Ratios** (other than proportions)**	**Smooth non-linear statistics (other than ratios e.g. regression coef., correlation coef.) *** **	**Non-smooth statistics (e.g. Gini coefficient, functions of quantiles)*** **
Simple random sampling	**Analytical:** ✔ Cochran (1977 p 23) **Jackknife:** ✔ Wolter (2007 p 162) **Bootstrap:** ✔ Rao *et al* (1984) ✔ Booth *et al* (1994) ✔ Wolter (2007 p 196) ✔ Bootstrap with replacement (MacCarthy and Snowden, 1985) ✔ Rescaled bootstrap (Rao and Wu, 1988) ✔ Mirror-matched bootstrap (Sitter, 1992b) ✔ Bootstrap without replacement (Gross, 1980) **Random Groups:** ✔ Wolter (2007) ✔ Särndal *et al* (1992 p 427)	**Analytical:** ✔ Cochran (1977 p 32) **Jackknife:** ✔ Wolter (2007 p 162) **Bootstrap:** ✔ Rao *et al* (1984) ✔ Booth *et al* (1994) ✔ Wolter (2007 p 196) ✔ Bootstrap with replacement (MacCarthy and Snowden, 1985) ✔ Rescaled bootstrap (Rao and Wu, 1988) ✔ Mirror-matched bootstrap (Sitter, 1992b) ✔ Bootstrap without replacement (Gross, 1980) **Linearisation of the statistic:** ✔ Taylor linearisation, Särndal *et al* (1992 p 178) **Random Groups:** ✔ Wolter (2007) ✔ Särndal *et al* (1992 p 427)	**Jackknife:** ✔ The Extended Delete-A-Group jackknife (Kott, 2001) ✔ Berger (2007, 2008) **Bootstrap:** ✔ Rao *et al* (1984) ✔ Booth *et al* (1994) ✔ Wolter (2007 p 196) ✔ Bootstrap with replacement (MacCarthy and Snowden, 1985) ✔ Rescaled bootstrap (Rao and Wu, 1988) ✔ Mirror-matched bootstrap (Sitter, 1992b) ✔ Bootstrap without replacement (Gross, 1980) **Linearisation of the statistic:** ✔ Taylor linearisation, Särndal *et al* (1992 p 178) ✔ Generalised linearisation method relying on the concept of influence function (Osier, 2009) **Random Groups:** ✔ Wolter (2007) ✔ Särndal *et al* (1992 p 427)	**Jackknife:** ✔ The Extended Delete-A-Group JK (Kott, 2001) ✔ Berger (2007, 2008) **!** Delete-one or groups jackknife is inconsistent for non-smooth statistics (except for the Gini coefficient) (Miller, 1974; Berger, 2008) **Bootstrap:** ✔ Rao *et al* (1984) ✔ Booth *et al* (1994) ✔ Wolter (2007 p 196) ✔ Bootstrap with replacement (MacCarthy and Snowden, 1985) ✔ Rescaled bootstrap (Rao and Wu, 1988) ✔ Mirror-matched bootstrap (Sitter, 1992b) ✔ Bootstrap without replacement (Gross, 1980) **Linearisation of the statistic:** ✔ Linearisation based on estimating equations (Binder, 1983, Kovacevic and Binder, 1997) ✔ Generalised linearisation method relying on the concept of influence function (Osier, 2009) **!** Taylor linearisation **Random Groups:** ✔ Wolter (2007) ✔ Särndal *et al* (1992 p 427)

Sampling designs	Type of statistics			
	Linear statistics (e.g. totals, means, proportions*)	**Ratios** (other than proportions)**	**Smooth non-linear statistics (other than ratios e.g. regression coef., correlation coef.) *** **	**Non-smooth statistics (e.g. Gini coefficient, functions of quantiles)*** **
Stratified random sampling	**Analytical:** ✔ Cochran (1977 p 95) **Jackknife:** ✔ Wolter (2007 p 172) ✔ Berger (2007, 2008) **!** Delete-one jackknife should not be used in stratified designs. See e.g. Wolter (2007) p. 172-173 **Bootstrap:** ✔ Rao *et al* (1984) ✔ Booth *et al* (1994) ✔ Wolter (2007 p 207) ✔ Bootstrap with replacement (MacCarthy and Snowden, 1985) ✔ Rescaled bootstrap (Rao and Wu, 1988) ✔ Mirror-matched bootstrap (Sitter, 1992b) ✔ Bootstrap without replacement (Gross, 1980) **Random Groups:** ✔ Wolter (2007) ✔ Särndal *et al* (1992 p 428) **Balanced repeated replication ****:** ✔ Wolter (2007) ✔ Särndal *et al* (1992 p 432)	**Analytical:** ✔ Cochran (1977 p 164) **Jackknife:** ✔ Wolter (2007 p 172) ✔ Berger (2007, 2008) **!** Delete-one jackknife should not be used in stratified designs. See e.g. Wolter (2007) p. 172-173 **Bootstrap:** ✔ Rao *et al* (1984) ✔ Booth *et al* (1994) ✔ Wolter (2007 p 207) ✔ Bootstrap with replacement (MacCarthy and Snowden, 1985) ✔ Rescaled bootstrap (Rao and Wu, 1988) ✔ Mirror-matched bootstrap (Sitter, 1992b) ✔ Bootstrap without replacement (Gross, 1980) **Linearisation of the statistic:** ✔ Taylor linearisation, Särndal *et al* (1992 p 178) **Random Groups:** ✔ Wolter (2007) ✔ Särndal *et al* (1992 p 428) **Balanced repeated replication ****:** ✔ Wolter (2007) ✔ Särndal *et al* (1992 p 432)	**Jackknife:** ✔ The Extended Delete-A-Group jackknife (Kott, 2001) ✔ Berger (2007, 2008) **!** Delete-one jackknife should not be used in stratified designs. See e.g. Wolter (2007) p. 172-173 **Bootstrap:** ✔ Rao *et al* (1984) ✔ Booth *et al* (1994) ✔ Wolter (2007 p 207) ✔ Bootstrap with replacement (MacCarthy and Snowden, 1985) ✔ Rescaled bootstrap (Rao and Wu, 1988) ✔ Mirror-matched bootstrap (Sitter, 1992b) ✔ Bootstrap without replacement (Gross, 1980) **Linearisation of the statistic:** ✔ Taylor linearisation, Särndal *et al* (1992 p 178) ✔ Linearisation based on estimating equations (Binder, 1983, Kovacevic and Binder, 1997) **Random Groups:** ✔ Wolter (2007) ✔ Särndal *et al* (1992 p 428) **Balanced repeated replication ****:** ✔ Wolter (2007) ✔ Särndal *et al* (1992 p 432)	**Jackknife** ✔ The Extended Delete-A-Group jackknife (Kott, 2001) ✔ Berger (2007, 2008) **!** Delete-one or groups jackknife is inconsistent for non-smooth statistics (except for the Gini coefficient) (Miller, 1974; Berger, 2008) **!** Delete-one jackknife should not be used in stratified designs. See e.g. Wolter (2007) p. 172-173 **Bootstrap:** ✔ Rao *et al* (1984) ✔ Booth *et al* (1994) ✔ Wolter (2007 p 207) ✔ Bootstrap with replacement (MacCarthy and Snowden, 1985) ✔ Rescaled bootstrap (Rao and Wu, 1988) ✔ Mirror-matched bootstrap (Sitter, 1992b) ✔ Bootstrap without replacement (Gross, 1980) **Linearisation of the statistic:** ✔ Linearisation based on estimating equations (Binder, 1983, Kovacevic and Binder, 1997) ✔ Generalised linearisation method relying on the concept of influence function (Osier, 2009) **!** Taylor linearisation **Random Groups:** ✔ Wolter (2007) ✔ Särndal *et al* (1992 p 428) **Balanced repeated replication** ****: ✔ Wolter (2007) ✔ Särndal *et al* (1992 p 432)

Sampling designs	Type of statistics			
	Linear statistics (e.g. totals, means, proportions*)	**Ratios** (other than proportions)**	**Smooth non-linear statistics (other than ratios e.g. regression coef., correlation coef.) *** **	**Non-smooth statistics (e.g. Gini coefficient, functions of quantiles)*** **
Single-stage cluster sampling	**Analytical:** ✔ Cochran (1977 p 261) **Jackknife:** ✔ Wolter (2007 p 182) ✔ Berger (2007, 2008) **Bootstrap:** ✔ Rao *et al* (1984) ✔ Deville (1987) ✔ Booth *et al* (1994) ✔ Wolter (2007 p 210) ✔ Chauvet (2007) ✔ Preston (2009) **Random Groups:** ✔ Wolter (2007) ✔ Särndal *et al* (1992 p 427)	**Analytical:** ✔ Cochran (1977 p 271) **Jackknife:** ✔ Wolter (2007 p 182) ✔ Berger (2007, 2008) **Bootstrap:** ✔ Rao *et al* (1984) ✔ Deville (1987) ✔ Booth *et al* (1994) ✔ Wolter (2007 p 210) ✔ Chauvet (2007) ✔ Preston (2009) **Linearisation of the statistic:** ✔ Taylor linearisation, Särndal (1992 p 178) **Random Groups:** ✔ Wolter (2007) ✔ Särndal *et al* (1992 p 427)	**Jackknife:** ✔ The Extended Delete-A-Group jackknife (Kott, 2001) ✔ Berger (2007, 2008) **Bootstrap:** ✔ Rao *et al* (1984) ✔ Deville (1987) ✔ Booth *et al* (1994) ✔ Wolter (2007 p 210) ✔ Chauvet (2007) ✔ Preston (2009) **Linearisation of the statistic:** ✔ Taylor linearisation, Särndal *et al* (1992 p 178) ✔ Linearisation based on estimating equations (Binder, 1983, Kovacevic and Binder, 1997) **Random Groups:** ✔ Wolter (2007) ✔ Särndal *et al* (1992 p 427)	**Jackknife:** ✔ The Extended Delete-A-Group jackknife (Kott, 2001) ✔ Berger (2007, 2008) **!** Delete-one or groups jackknife is inconsistent for non-smooth statistics (except for the Gini coefficient) (Miller, 1974; Berger, 2008) **Bootstrap:** ✔ Rao *et al* (1984) ✔ Deville (1987) ✔ Booth *et al* (1994) ✔ Wolter (2007 p 210) ✔ Chauvet (2007) ✔ Preston (2009) **Linearisation of the statistic:** ✔ Linearisation based on estimating equations (Binder, 1983, Kovacevic and Binder, 1997) ✔ Generalised linearisation method relying on the concept of influence function (Osier, 2009) **!** Taylor linearisation **Random Groups:** ✔ Wolter (2007) ✔ Särndal *et al* (1992 p 427)

Sampling designs	Type of statistics			
	Linear statistics (e.g. totals, means, proportions*)	**Ratios** (other than proportions)**	**Smooth non-linear statistics (other than ratios e.g. regression coef., correlation coef.) *** **	**Non-smooth statistics (e.g. Gini coefficient, functions of quantiles)*** **
Stratified single-stage cluster sampling	**Analytical:** ✔ Cochran (1977 p 271) **Jackknife:** ✔ The Extended Delete-A-Group jackknife, Kott (2001) ✔ Berger (2007, 2008) **!** Delete-one jackknife should not be used in stratified designs. See e.g. Wolter (2007) p. 172-173 **Bootstrap:** ✔ Rao *et al* (1984) ✔ Deville (1987) ✔ Booth *et al* (1994) ✔ Wolter (2007 p 210) ✔ Chauvet (2007) ✔ Preston (2009) **Random Groups:** ✔ Wolter (2007) ✔ Särndal *et al* (1992 p 428) **Balanced repeated replication ****:** ✔ Wolter (2007) ✔ Särndal *et al* (1992 p 432)	**Analytical:** ✔ Cochran (1977 p 271) **Jackknife:** ✔ The Extended Delete-A-Group jackknife, Kott (2001) ✔ Berger (2007, 2008) **!** Delete-one jackknife should not be used in stratified designs. See e.g. Wolter (2007) p. 172-173 **Bootstrap:** ✔ Rao *et al* (1984) ✔ Deville (1987) ✔ Booth *et al* (1994) ✔ Wolter (2007 p 210) ✔ Chauvet (2007) ✔ Preston (2009) **Linearisation of the statistic:** ✔ Taylor linearisation, Särndal *et al* (1992 p 178) **Random Groups:** ✔ Wolter (2007) ✔ Särndal *et al* (1992 p 428) **Balanced repeated replication ****:** ✔ Wolter (2007) ✔ Särndal *et al* (1992 p 432)	**Jackknife:** ✔ The Extended Delete-A-Group jackknife (Kott, 2001) ✔ Berger (2007, 2008) **!** Delete-one jackknife should not be used in stratified designs. See e.g. Wolter (2007) p. 172-173 **Bootstrap:** ✔ Rao *et al* (1984) ✔ Deville (1987) ✔ Booth *et al* (1994) ✔ Wolter (2007 p 210) ✔ Chauvet (2007) ✔ Preston (2009) **Linearisation of the statistic:** ✔ Taylor linearisation, Särndal *et al* (1992 p 178) ✔ Linearisation based on estimating equations (Binder, 1983, Kovacevic and Binder, 1997) **Random Groups:** ✔ Wolter (2007) ✔ Särndal *et al* (1992 p 428) **Balanced repeated replication ****:** ✔ Wolter (2007) ✔ Särndal *et al* (1992 p 432)	**Jackknife:** ✔ The Extended Delete-A-Group jackknife (Kott, 2001) ✔ Berger (2007, 2008) **!** Delete-one or groups jackknife is inconsistent for non-smooth statistics (except for the Gini coefficient) (Miller, 1974; Berger, 2008) **!** Delete-one jackknife should not be used in stratified designs. See e.g. Wolter (2007) p. 172-173 **Bootstrap:** ✔ Rao *et al* (1984) ✔ Deville (1987) ✔ Booth *et al* (1994) ✔ Wolter (2007 p 210) ✔ Chauvet (2007) ✔ Preston (2009) **Linearisation of the statistic:** ✔ Linearisation based on estimating equations (Binder, 1983, Kovacevic and Binder, 1997) ✔ Generalised linearisation method relying on the concept of influence function (Osier, 2009) **!** Taylor linearisation **Random Groups:** ✔ Wolter (2007) ✔ Särndal *et al* (1992 p 428) **Balanced repeated replication ****:** ✔ Wolter (2007) ✔ Särndal *et al* (1992 p 432)

Sampling designs	Type of statistics			
	Linear statistics (e.g. totals, means, proportions*)	**Ratios** (other than proportions)**	**Smooth non-linear statistics (other than ratios e.g. regression coef., correlation coef.) *** **	**Non-smooth statistics (e.g. Gini coefficient, functions of quantiles)*** **
Multi-stage (cluster) sampling	**Analytical:** ✔ Särndal *et al* (1992 p 135) **Jackknife:** ✔ The Extended Delete-A-Group jackknife, Kott (2001) ✔ Berger (2007, 2008) **Bootstrap:** ✔ Rao *et al* (1984) ✔ Deville (1987) ✔ Booth *et al* (1994) ✔ Wolter (2007 p 210) ✔ Chauvet (2007) ✔ Preston (2009) **Random Groups:** ✔ Wolter (2007) ✔ Särndal *et al* (1992 p 429) **Balanced repeated replication ****:** ✔ Wolter (2007) ✔ Särndal *et al* (1992 p 435)	**Analytical:** ✔ Särndal *et al* (1992 p 135) **Jackknife:** ✔ The Extended Delete-A-Group jackknife, Kott (2001) ✔ Berger (2007,2008) **Bootstrap:** ✔ Rao *et al* (1984) ✔ Deville (1987) ✔ Booth *et al* (1994) ✔ Wolter (2007 p 210) ✔ Chauvet (2007) ✔ Preston (2009) **Linearisation of the statistic:** ✔ Taylor linearisation, Särndal *et al* (1992 p 178) **Random Groups:** ✔ Wolter (2007) ✔ Särndal *et al* (1992 p 429) **Balanced repeated replication ****:** ✔ Wolter (2007) ✔ Särndal *et al* (1992 p 435)	**Jackknife:** ✔ The Extended Delete-A-Group jackknife (Kott, 2001) ✔ Berger (2007, 2008) **Bootstrap:** ✔ Rao *et al* (1984) ✔ Deville (1987) ✔ Booth *et al* (1994) ✔ Wolter (2007 p 210) ✔ Chauvet (2007) ✔ Preston (2009) **Linearisation of the statistic:** ✔ Taylor linearisation, Särndal *et al* (1992 p 178) ✔ Linearisation based on estimating equations (Binder, 1983, Kovacevic and Binder, 1997) **Random Groups:** ✔ Wolter (2007) ✔ Särndal *et al* (1992 p 429) **Balanced repeated replication ****:** ✔ Wolter (2007) ✔ Särndal *et al* (1992 p 435)	**Jackknife:** ✔ The Extended Delete-A-Group jackknife (Kott, 2001) ✔ Berger (2007, 2008) **!** Delete-one or groups jackknife is inconsistent for non-smooth statistics (except for the Gini coefficient) (Miller, 1974; Berger, 2008) **Bootstrap:** ✔ Rao *et al* (1984) ✔ Deville (1987) ✔ Booth *et al* (1994) ✔ Wolter (2007 p 210) ✔ Chauvet (2007) ✔ Preston (2009) **Linearisation of the statistic:** ✔ Linearisation based on estimating equations (Binder, 1983, Kovacevic and Binder, 1997) ✔ Generalised linearisation method relying on the concept of influence function (Deville, 1999, Osier, 2009) **!** Taylor linearisation **Random Groups:** ✔ Wolter (2007) ✔ Särndal *et al* (1992 p 429) **Balanced repeated replication ****:** ✔ Wolter (2007) ✔ Särndal *et al* (1992 p 435)

* In Section 2.2 proportions and ratios are defined in the general context: a ratio is a ratio of two totals or means, while a proportion is a special case of ratio, where the numerator and denominator are counts of elements in domain A and domain B respectively, where domain A is a subset of domain B. In this general context, both ratios and proportions can have constant denominator or variable denominator. However, for the purpose of variance estimation and for simplification purposes, **proportion** is used to designate a linear statistic where the variance of the estimator in the denominator is zero. (The denominator is a constant, e.g. known from external sources.)

** In Section 2.2 proportions and ratios are defined in the general context: a ratio is a ratio of two totals or means, while a proportion is a special case of ratio, where the numerator and denominator are counts of elements in domain A and domain B respectively, where domain A is a subset of domain B. In this general context, both ratios and proportions can have constant denominator or variable denominator. However, for the purpose of variance estimation and for simplification, **ratio** is used to designate a ratio of two estimators with a denominator having a non-zero variance. (The denominator is not a constant, but a random variable being estimated from a survey). This can occur for example in the case of domain estimates.

*** **Smooth non-linear statistics** are differentiable non-linear statistics for which Taylor series expansions can be used. They can be expressed as differentiable functions of linear statistics. Other examples besides ratios of two linear statistics are the estimators for regression coefficients or for correlation coefficients. **Non-smooth non-linear statistics** are non-linear statistics for which Taylor series expansions can no longer be used; influence functions can be used for linearisation of non-smooth statistics (income quintile, Gini coefficient).

**** Other names for the ***balanced repeated replication (BRR)*** technique are ***balanced half-samples (BHS)*** and ***pseudo-replication***. Originally the balanced half-samples technique was used for the case with a large number of strata and a sample composed of only two elements (or two PSUs) per stratum. Modification of the technique has been suggested for cases where the sample sizes n_h exceed two. For a review, see Wolter (2007).

7.5 Suitability of software tools for sampling designs and related issues on variance estimation

	POULPE	CLAN	SAS/SPSS	R*	GENESEES	REGENESEES	SUDAAN	BASCULA
Simple random sampling	Yes. Let $\bar{y}$ be the sample mean of a variable y. Then we have: $\hat{V}(\bar{y}) = (1-f)\frac{s^2}{n}$, where n is the sample size, f the finite population correction and s^2 is the dispersion of variable y over the sample.	Yes. Same formula as POULPE.	Yes. Same formula as POULPE.	Yes.	Yes.	Yes.	Yes.	Yes.
Stratified random sampling	Yes. Let $\bar{y}_h$ be the sample mean of a variable y over stratum h. Let w_h be the relative weight of stratum h in the population. We have: $\hat{\bar{Y}} = \sum_h w_h \bar{y}_h$ and $\hat{V}\left(\hat{\bar{Y}}\right) = \sum_h (1-f_h)\frac{s_h^2}{n_h}$, where n_h is the sample size in stratum h, f_h the finite population correction in stratum h and s_h^2 is the dispersion of variable y over the sample in stratum h.	Yes. Same formula as POULPE.	Yes. Same formula as POULPE.	Yes.	A special feature available in GENESEES is the merging of strata (collapsing) for strata with small response frequencies (the user can choose the frequency level). This problem is especially notable when there is only one PSU in one or more strata.	Yes. Collapse strata technique for handling single PSUs (Rust and Kalton (1987)).	Yes.	Yes.
Single-stage cluster sampling and multi-stage	Yes. Every multi-stage sampling can be split into 'elementary' samplings. Variances in each 'elementary' sampling are	Yes. Ultimate cluster approximation: the second	Yes. Ultimate cluster approximation.	Yes. Ultimate cluster approximation.	Yes. Ultimate cluster approximation.	Yes, both via the ultimate cluster approximation	Yes. Like POULPE, SUDAAN uses an exact	BASCULA can handle the following

	POULPE	CLAN	SAS/SPSS	R*	GENESEES	REGENESEES	SUDAAN	BASCULA
(cluster) sampling	estimated and then combined so as to form an estimate for the overall variance. The underlying formula is due to Raj (1968). See Section 3.3 for this formula. The formula is the cornerstone of the software POULPE. If PSUs are assumed to be selected with replacement, then the variance of a multi-stage sampling can be estimated by the variance of the estimated totals of the PSUs. This approximation, called 'ultimate cluster approximation', is implemented in the software too. Its interest lies in its simplicity.	(third, etc.) stage's variance is omitted. However, one can introduce the second-stage variation into the calculations with additional computational efforts.				or by means of an actual multi-stage computation (Bellhouse (1985) recursive algorithm).	variance decomposition.	designs: i) Stratified two-stage sampling where both PSUs and secondary sampling units (SSUs) are selected by simple random sampling. ii) Stratified multi-stage sampling where PSUs are selected (possibly with unequal probabilities) with replacement.
Multi-phase (cluster) sampling	Variance estimation under multi-phase sampling is a long-established theory (Särndal *et al* 1992). It can be regarded as an extension of variance estimation under multi-stage sampling: the variance of a multi-phase design can be expressed as a sum of variance terms representing the contribution of each sampling phase. POULPE tackles multi-phase sampling designs by estimating the variances contributed by each phase of selection and then	Same approach as POULPE. The idea is to decompose for the variance of first-phase estimator and the variance for second-phase.	No formula. Can be programmed by the user.	Yes.	No.	No.	No.	No.

	POULPE	CLAN	SAS/SPSS	R*	GENESEES	REGENESEES	SUDAAN	BASCULA
	combining them in order to obtain an estimate for the overall variance.							
Can software tools take into account in the overall variance estimation the effects of:								
Implicit stratification?	Systematic sampling with equal probabilities can be treated using the following approximate estimator. Let $\hat{Y}$ be an estimator for the total Y of a variable y. A variance estimator is given by: $\hat{V}_i(\hat{Y}) = N^2\left(\frac{1}{n}-\frac{1}{N}\right)\sum_{j=2}^{n}\frac{(y_{i,j}-y_{i,j-1})^2}{2(n-1)}$ where y_{ij} denotes the value of y for the j th individual of the i th systematic sample.	Yes.	No formula. It has to be programmed.	No formula. It has to be programmed.	No.	No.	No.	No.
(Unit) non-response?	Non-response is viewed as an additional phase of selection. POULPE offers two possibilities as to the kind of selection carried out at second-phase: either Poisson selection or post-stratified one.	Yes.	No formula. It has to be programmed.	Use formulae for multi-phase sampling.	No.	Yes, with the calibration approach of Särndal and Lundström (2005). The calibration step intended to fight non-response bias and the calibration step intended to improve estimators efficiency can be performed simultaneously or subsequently.	Use replication methods (adjust the non-response weight in each replication).	Use replication methods (adjust the non-response weight in each replication).

	POULPE	CLAN	SAS/SPSS	R*	GENESEES	REGENESEES	SUDAAN	BASCULA
Imputation (for item non-response)?	No special treatment. Imputed values are treated as if they were the exact ones.	No special treatment. Imputed values are treated as if they were the exact ones.	No formula. Has to be programmed.	No formula. Has to be programmed.	No.	No.	Multiple imputation.	No.
Substitution?	Substituted units are treated as if they were the original ones.	Substituted units are treated as if they were the original ones.	Substituted units are treated as if they were the original ones.	Substituted units are treated as if they were the original ones.	Substituted units are treated as if they were the original ones.	No.	Substituted units are treated as if they were the original ones.	Substituted units are treated as if they were the original ones.
Calibration?	The variance of a calibration linear estimator is asymptotically equal to that of the estimator based on initial weights and using as variable of interest the regression residuals of the target variable over the calibration variables (Deville and Särndal, 1992).	Contrary to POULPE, CLAN has both a structure for calibrated weight creation and for variance estimation taking the calibration into account. The results are based on the properties of the GREG estimator (technique of residuals).	No formula. Has to be programmed (compute the regression residuals).	Two main functions for calibration: Post-stratify (post-stratification); Calibrate (regression calibration).	Yes (GREG estimators).	Yes, by exploiting the asymptotic equivalence of all calibration estimators to the GREG and by linearising the GREG for variance estimation (as in POULPE).	Post-stratification only.	Yes (GREG estimator).
Rotating schemes?	Yes. But simplifying assumptions must be made regarding the rotation groups (the sub-samples are assumed to be independent).	Yes.	Yes. But simplifying assumptions must be made.	Yes. But simplifying assumptions must be made.	No.	No.	No.	No.

	POULPE	CLAN	SAS/SPSS	R*	GENESEES	REGENESEES	SUDAAN	BASCULA
Joint inclusion probabilities?	An approximation formula is available for handling this kind of selection. Let n be the sample size and π_i the probability of selection of *i.* Let $\hat{Y} = \sum_{i \in s} \frac{y_i}{\pi_i}$ be the Horvitz-Thompson estimator for the total Y of the variable of interest y. A variance estimator for the estimated total $\hat{Y}$ is given by: $\hat{V}(\hat{Y}) = \frac{n}{n-1} \sum_{i=1}^{n} (1-\pi_i) \left(\frac{y_i}{\pi_i} - B \right)^2$ (Hájek, 1964), where $B = \frac{\sum_{i=1}^{n} \frac{y_i}{\pi_i}(1-\pi_i)}{\sum_{i=1}^{n}(1-\pi_i)}$.	Yes.	Ultimate cluster approximation.	Ultimate cluster approximation.	Yes. Ultimate cluster approximation.	No approximate expressions available for joint inclusion probabilities. Selection under πpswor treated as if it were under πpswr.	SUDAAN has an option denoted by 'UNEQWOR', available for the first stage only. It uses the Yates-Grundy-Sen variance estimator which is given by $\hat{V}(\hat{T}) = -\frac{1}{2} \times \sum_i \sum_{i>j} \Delta_{ij} \cdot \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2$ where $\Delta_{ij} = \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}}$ $\hat{Y} = \sum_i \frac{y_i}{\pi_i}$ and π_i and π_{ij} designate the simple inclusion probability of unit *i* and the joint inclusion probability of *i* and *j*, respectively.	Yes

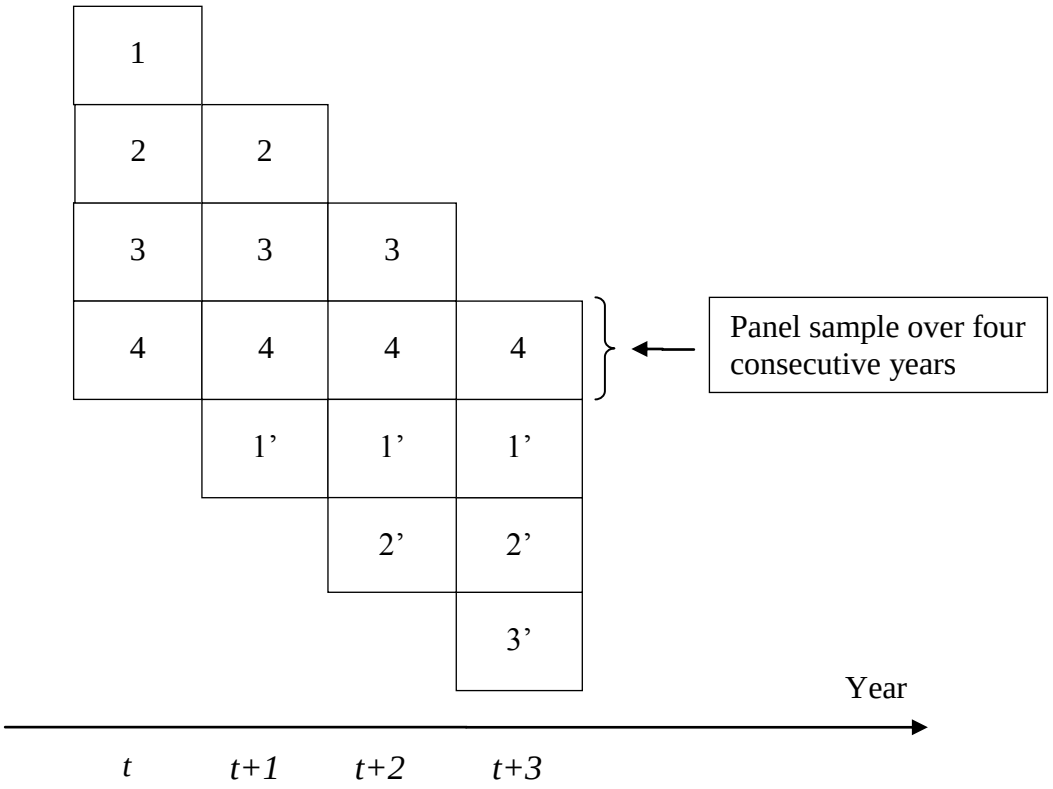
	POULPE	CLAN	SAS/SPSS	R*	GENESEES	REGENESEES	SUDAAN	BASCULA
Non-linear statistics?	The software handles linear statistics. In order to deal with non-linear statistics, the solution is to linearise them and to run POULPE on the linearised variable.	The software handles linear statistics. In order to deal with non-linear statistics, the solution is to linearise them and to run the software on the linearised variable.	SAS can deal with ratios using Taylor approximation. Otherwise, linearisation formulae must be programmed (Osier, 2009).	Use of Taylor linearisation. Otherwise, statistics have to be linearised prior to any calculations.	Use of Taylor linearisation. Otherwise, statistics have to be linearised prior to any calculations.	Yes, provided they can be expressed as closed-form, differentiable functions of Horvitz-Thompson or calibration estimators of totals and means. They can be even freely defined by the user. ReGenesees linearises such estimators automatically, on-the-fly.	Non-linear statistics have to be linearised first.	Non-linear statistics have to be linearised first. Another solution is to use replication methods.

* Reference is made to the packages *sampling* and *survey* presented in Section 3.5.

7.6 Minimum effective sample size for longitudinal estimates

As significant efforts are required to obtain panel data, it is only fair to require a certain level of precision for gross changes. Precision requirements for gross changes can be expressed as the minimum effective sample sizes that need to be achieved between any pair of consecutive waves. As in EU-SILC, let us consider a simple rotating design based on four rotation groups (once the system is fully established). The sample at a given year consists of four rotation groups which have been in the survey for 1-4 years. Any particular rotation group remains in the survey for four years; each year one of the four groups from the previous year is dropped and a new one is added. Between year *t* and *t+1* the sample overlap is 75%; the overlap between year *t* and year *t+2* is 50%; it is reduced to 25% from year *t* to year *t+3*, and to zero for longer intervals.

Figure 7.6.1: The EU-SILC rotating scheme



When a panel sample is nested in a larger structure, as in the above figure, the precision of the panel sample must be linked to that of the larger structure. More precisely, in the above figure, the precision of the panel (2,3,4) between *t* and *t+1* depends on that of the cross-sectional samples (1,2,3,4) at year *t* and (2.3.4.1') at year *t+1*. For instance, the EU-SILC Framework Regulation (No 1177/2003 of 16 June 2003) specifies minimum sample sizes for both the cross-sectional and the longitudinal dimension. For any pair of consecutive years, sample size for the longitudinal component refers to the number of households successfully interviewed in the first year in which all, or at least a majority, of the household members aged 16 or over are successfully interviewed in both years.

In practice, there are often precision requirements for the cross-sectional samples (1,2,3,4) and (2,3,4,1') too (as in EU-SILC). Let $n_{\min}^{(cross)}$ be the minimum sample size required at cross-sectional level. If a proportion α of the sample $(0<\alpha<1)$ rotates out at each wave, then the minimum sample size between two consecutive waves is given by:

$$n_{\min}^{(long)} = (1-\alpha)\cdot n_{\min}^{(cross)}. \quad (7.6.1)$$

Please note that in the above formula we assume that $n_{\min}^{(cross)}$ is fixed and we derive the $n_{\min}^{(long)}$. In the EU-SILC rotating scheme (Figure 7.6.1), we have $\alpha = 0.25$. In the case of a 'pure' panel (that is, no rotation of the sample), we have $\alpha = 0$.

(7.6.1) assumes full response from one wave to another, something which is, obviously, unrealistic. *Therefore, when we set up precision requirements for gross changes, we should take into account the loss of accuracy caused by non-response between two consecutive waves*. As a matter of fact, at least in the first years of the survey, many EU-SILC countries departed from Eurostat's recommendation of having a four-year rotating design by rotating out less than 25% of their sample at each wave. Thus, the longitudinal sample size was higher and the loss of sampling efficiency due to higher non-response was under control. This solution is interesting and could be generalised to all surveys which are based on rotating samples. Let us assume that a proportion α of the sample $(0<\alpha<1)$ rotates out at each wave. However, to accommodate a certain level of non-response between two consecutive waves, only a proportion β $(0<\beta<\alpha<1)$ of the sample is actually replaced. The value of β can then be derived from α and the (predicted) response probability r between two consecutive waves is given by:

$$\frac{1-\alpha}{1-\beta} = r. \quad (7.6.2)$$

In panel surveys, r is generally high, so β would be close to α. Thus, by rotating out fewer sample units than initially planned, we can allow for a certain degree of non-response and then ensure that we achieve the minimum required longitudinal sample size that would have been achieved under full response, when a proportion α of the sample rotates out at each wave.

An alternative solution is to issue minimum sample sizes for the longitudinal dimension on the basis of the minimum sample sizes required for the cross-sectional dimension, and by taking into account both the rotation rate and the probability of response between two waves. Thus, let $n_{\min}^{(cross)}$ be the minimum sample size required at cross-sectional level. Let us assume that a proportion α of the sample $(0<\alpha<1)$ rotates out at each wave and that there is a probability r for a unit to respond at t+1 given that the unit has responded at t. Thus, the minimum longitudinal sample size between t and $t+1$ is given by:

$$n_{\min}^{(long)} = r\cdot(1-\alpha)\cdot n_{\min}^{(cross)}. \quad (7.6.3).$$

This solution has the disadvantage of using a pre-determined value r for the response probability, which does not take into account the differences in response behaviour from one country to another. On the other hand, the former option (although we assume full response) leaves the responsibility of adjusting rotating schemes to the countries in order to meet the precision requirements.

Index

European Commission

Handbook on precision requirements and variance estimation for ESS household surveys

Luxembourg: Publications Office of the European Union

2013 — 166 pp. — 21 x 29.7 cm

Theme: General and regional statistics
Collection: Methodologies & Working papers

ISBN 978-92-79-31197-0
ISSN 1977-0375
doi:10.2785/13579
Cat. No: KS-RA-13-029-EN-N

HOW TO OBTAIN EU PUBLICATIONS

Free publications:

- one copy:
  via EU Bookshop (http://bookshop.europa.eu);
- more than one copy or posters/maps:
  from the European Union's representations (http://ec.europa.eu/represent_en.htm);
  from the delegations in non-EU countries (http://eeas.europa.eu/delegations/index_en.htm);
  by contacting the Europe Direct service (http://europa.eu/europedirect/index_en.htm) or calling 00 800 6 7 8 9 10 11 (freephone number from anywhere in the EU) (*).

  (*) The information given is free, as are most calls (though some operators, phone boxes or hotels may charge you).

Priced publications:

- via EU Bookshop (http://bookshop.europa.eu).

Priced subscriptions:

- via one of the sales agents of the Publications Office of the European Union (http://publications.europa.eu/others/agents/index_en.htm).

KS-RA-13-029-EN-N

ISBN 978-92-79-31197-0

Publications Office