

11-619E

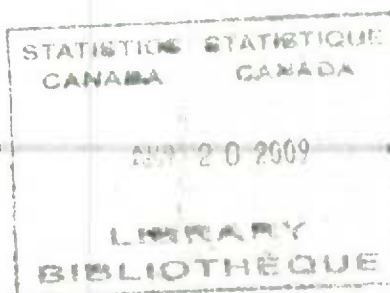


Statistics
Canada

Statistique
Canada

no.2009-002

c.3



Methodology Branch

Direction de la méthodologie

Household Survey
Methods Division

Division des méthodes
d'enquêtes des ménages

Canada

WORKING PAPER
METHODOLOGY BRANCH

**RETHINKING THE NLSCY WEIGHTING METHODOLOGY:
THE LESSONS LEARNED AND THE INNOVATIONS INTRODUCED**

HSMD-2009-002E

Editors: Claude Girard, Michel Ferland, Sarah Franklin and Marcelle Tremblay

Contributors:

Pierre Caron
Yves Lafortune
Bruno Lapierre
Scott Meyer
Michelle Simard
Mike Tam

Household Survey Methods Division
Statistics Canada
March 2009

The work presented in this paper is the responsibility of the editors and does not necessarily represent the views or policies of Statistics Canada.

Abstract

The National Longitudinal Survey of Children and Youth (NLSCY) is a longitudinal survey of Canadian children conducted by Statistics Canada. Introduced in the mid-nineties, it is one of the first longitudinal surveys conducted by the Agency. Over the last two years, a team of methodologists have undertaken a major revision of NLSCY's weighting strategy after evidence of bias surfaced and deficiencies in its implementation were noted. Along the way, many lessons were learned on how to conduct the weighting of a longitudinal survey. Moreover, innovations were introduced to address some of the issues raised. This paper describes these at length.

Résumé

L'Enquête longitudinale nationale sur les enfants et les jeunes (ELNEJ) est une enquête longitudinale canadienne menée par Statistique Canada. Introduite au milieu des années quatre-vingt-dix, l'ELNEJ est l'une des premières enquêtes longitudinales menée par l'Agence. Au cours des deux dernières années, une équipe de méthodologistes a entrepris une révision en profondeur de la méthodologie de pondération alors employée par L'ELNEJ après que des indications de biais dans les estimations et des failles dans sa mise en œuvre aient été notées. Au cours de l'étude, plusieurs leçons ont été apprises sur la façon de procéder pour la pondération d'une enquête longitudinale. De plus, cette étude a donné lieu à quelques innovations qui ont été introduites afin de répondre aux enjeux soulevés. Ce document décrit en détail ces enjeux, les leçons apprises et les innovations introduites.

Table of contents

Table of contents	5
Important note to the reader	7
1.0 Introduction.....	8
2.0 Overview of the NLSCY.....	9
2.1 An overview of the NLSCY sample design.....	9
2.2 An overview of the weighting and variance estimation processes of the NLSCY.....	11
3.0 General features of the NLSCY design weights.....	13
3.1 The LFS subweight	13
3.2 The eligible kid adjustment.....	15
3.3 The rotation group adjustment.....	15
3.4 The birth month coverage adjustment.....	15
3.5 The <i>histoire</i> file	16
4.0 Nonresponse in the NLSCY	18
4.1 Total nonresponse	20
4.2 Partial nonresponse	20
4.2.1 The current situation regarding imputation	21
4.2.2 Future plans for treating item nonresponse.....	22
5.0 Addressing total nonresponse	24
5.1 Weighting.....	24
5.1.1 Modelling longitudinal nonresponse: <i>saut-de-mouton</i> versus <i>saut-de-l'ange</i>	29
5.1.1.1 Model building: considerations when applying the <i>saut-de-mouton</i>	29
5.1.1.2 Model building: considerations when applying the <i>saut-de-l'ange</i>	34
5.1.2 Variance estimation considerations about weighting	35
5.2 Calibration.....	36
5.2.1 Calibration in practice: the weighting steps.....	37
5.2.2 Calibration in practice: the variance estimation process.....	38
6.0 The nonresponse weighting methodologies of Cycles 1 through 7	39
6.1 Cycles 1 and 2	40
6.2 Cycle 3	41
6.3 Cycles 4 and 5.....	42
6.3.1 Handling converted units in a <i>saut-de-mouton</i> framework.....	44
6.4 Cycles 6 and 7	46
6.4.1 Segmentation modelling versus logistic modelling.....	46
6.4.2 Using response propensity scores at Cycles 6 and 7	46
6.4.3 The cooperation variables	48
6.5 Comparing the nonresponse weighting methodologies of Cycles 1 to 6 using longitudinal consistency	50
7.0 Post-stratification and other issues.....	53
7.1 Post-stratification	53
7.2 Weighting sub-components of interest of the NLSCY.....	55
8.0 Conclusion	57
References.....	58
Perspective A: On the challenges of producing multi-purpose weights.....	61
Perspective B: An overview of the bootstrap as used in the NLSCY.....	63
Perspective C: Dissemination procedure including the RDC network and the ATT.....	87

Perspective D: Unequal weights in a survey like the NLSCY.....	91
Perspective E: On the mathematical activity of modelling	97
Perspective F: On nonresponse bias detection	105
Perspective G: On the estimation of response propensities	109
Perspective H: On the role of the weights in the creation of the RHGs.....	113
Perspective I: On our use of models and their role in the inferential framework.....	115
Perspective J: On the choice of a calibration distance	121
Perspective K: On post-stratification.....	125

Important note to the reader

The structure of this paper requires a few words of explanation. This paper documents the extensive review that was performed on many of the aspects of weighting within the NLSY, the issues we faced and the solutions we identified, along with the lessons learned along the way.

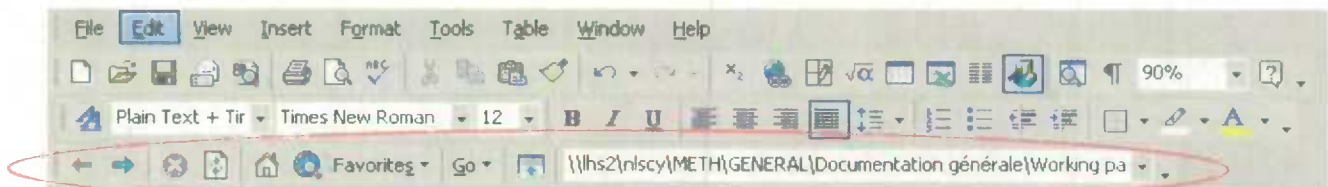
The main body is written in a more classical style and presents the reader with a detailed exposé of the issues we address here. At a few places, though, we felt a digression (referred to as a *Perspective*) was in order to provide the reader with more in-depth information. Such occurrences appear in the form of a box in the main text which looks like this:

Further exploration

This is a dummy example of a *Perspective*. In a real *Perspective*, the word would be underlined to indicate a hyperlink which would take the reader directly to the appropriate *Perspective* located in the second half of this document.

The style adopted in the *Perspectives* is purposely verbose and more familiar, as it suits more the conversational aspect of the discussion undertaken. At some point, the reader will want to return to the main text at the same spot he or she left before the digression. The painless way to navigate efficiently about the document is to activate the “Web” toolbar. This is very simple to do: from the main toolbar, select View->Toolbars->Web.

At this point the Word environment will show a series of buttons looking like those in the ellipse below:



Pay special attention to the left- and right-arrows this toolbar contains (they are located in the left-hand side of the ellipse in this snapshot); use them to navigate about the document, as if it were a Web environment. So, for example, after accessing *Perspective A* from Section 2, and having read it, the reader can return to Section 2 by clicking on the left arrow in the Web toolbar.

In the main text, cross-references and references are provided in the form of hyperlinks to help the reader navigate throughout the document (e.g., in a statement such as “[...] in Section X we discussed [...]”, “Section X” would be a hyperlink. Also, the equations, tables and figures contained in the document are assigned a sequential number within subsections. For example, the first table of subsection 3.1 is referenced as Table 3.1-1.

1.0 Introduction

The National Longitudinal Survey of Children and Youth (NLSCY) is a long-term study of Canadian children, and one of six longitudinal surveys launched by Statistics Canada in the 1990s. The other surveys are: the National Population Health Survey (NPHS), the Survey of Labour and Income Dynamics (SLID), the Youth in Transition Survey (YITS), the Longitudinal Survey of Immigrants to Canada (LSIC) and the National Graduates Survey (NGS) together with the Follow-up on Graduates Survey (FOG). Compared to other countries such as Great Britain and the United States, Canada is a relative newcomer to this field and consequently learned much during its first decade of designing and conducting longitudinal surveys.

As Statistics Canada's expertise and understanding grew about the challenges of longitudinal surveys, including their growing cost and complexity, the agency began to examine the future and direction of these large-scale surveys (Picot and Webber (2005)). This movement spurred a more detailed review of the first ten years of the NLSCY, with a particular focus on the methodology (Statistics Canada (2007)). One goal of this review (hereafter referred to as the 10-year review) was to determine if nonresponse bias was present in the survey estimates. An investigation concluded that bias was likely present, at least for some estimates, and that the methodology used to adjust for nonresponse at Cycles 1 to 5 had some deficiencies. It further revealed that changes to the existing methodology were required, to take into account both potential nonresponse bias and recent methodological developments (for example, in variance estimation). Indeed, our knowledge of (longitudinal) methodology has greatly increased since the methodology of the NLSCY was devised and it was thus important to revisit it in this new light.

The investigation into nonresponse bias became so comprehensive that it encompassed revisiting the derivation of the survey weights, from the design weight calculations, to the nonresponse adjustment and post-stratification, as well as the bootstrap variance methodology. In this paper we report our findings on all of the topics that broadly fall within the realm of weighting. We also share some of the thought processes that arose from the investigation of topics which have an indirect impact on the issues, such as modelling, the role of weights in the nonresponse methodology, *etc.* As mentioned in the note to the reader, these digressions are presented as *Perspectives* so as not to disrupt the natural flow of the main discussion of weighting issues. While the NLSCY hosts quite a few cohorts, to avoid repetition the focus in this paper will mainly be on the Original Cohort which has suffered the most extensive attrition and also provides a good panorama of the issues encountered with any one of the NLSCY cohorts.

The paper is aimed at methodologists with an interest in longitudinal surveys, hoping to provide some insight into how they should design and implement their methodologies, as well as analyze existing ones, but since many of the topics raised also apply to cross-sectional surveys, it should be of interest to methodologists working on these surveys as well.

2.0 Overview of the NLSCY

The NLSCY is designed to identify important factors that shape the development of Canadians from birth to adulthood. The content of the NLSCY is wide-ranging and includes information about a child's cognitive, emotional, and physical development. Questions touch on the child's health, learning, behaviour, and social environment (family, friends, schools, and communities). This information comes from a variety of sources: the selected child, the Person Most Knowledgeable about the selected child (or PMK, who is usually the mother), and on occasion principals and school teachers.

The NLSCY has been collecting information on children every two years since 1994 through various cohorts. Some of these are long-lived (*e.g.*, the Original Cohort introduced at the start of the survey) while others are designed to exist only for a few cycles. The analysts of NLSCY survey data are interested in a wide range of areas, and conduct many different types of analyses.

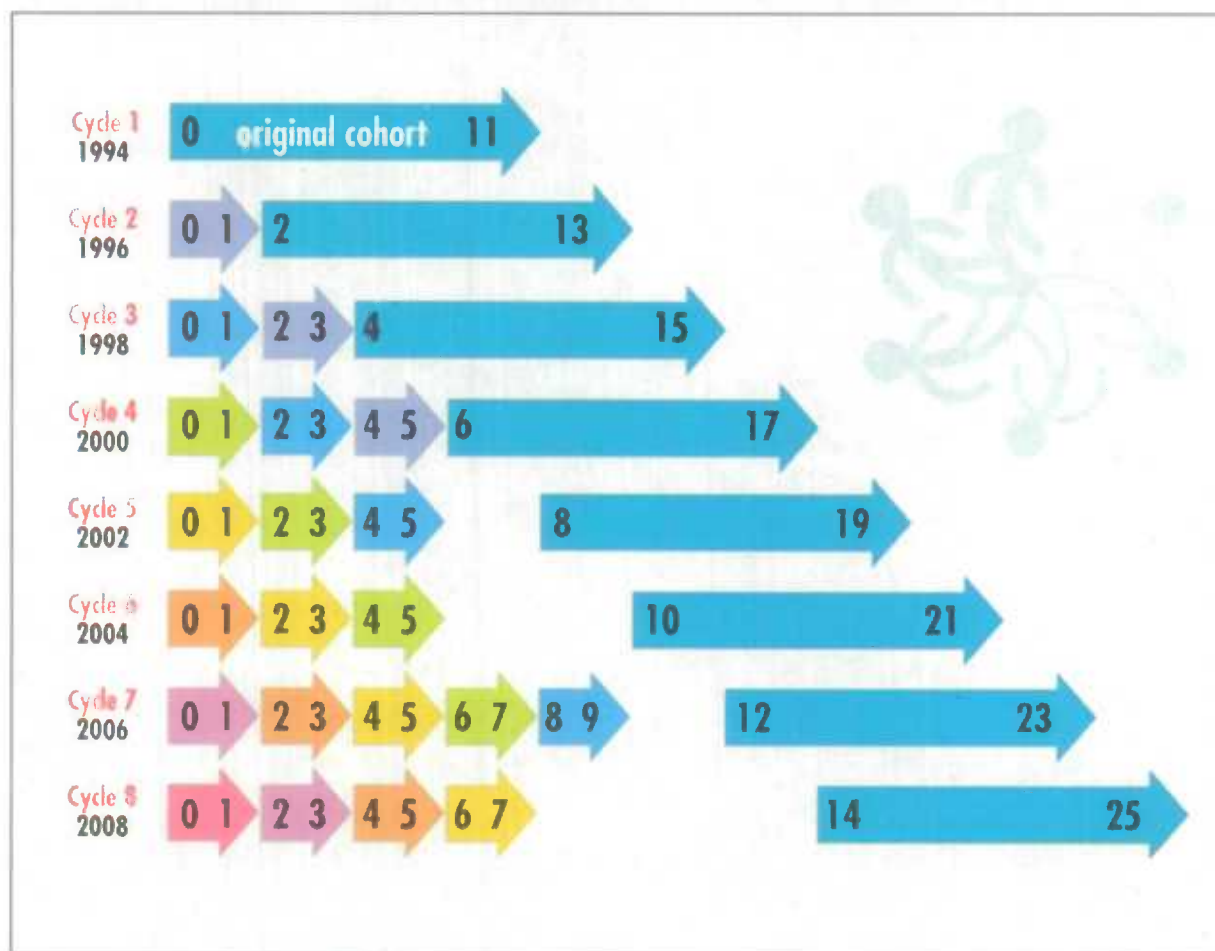
2.1 An overview of the NLSCY sample design

The first sample of children was initially surveyed in 1994-95 and again every two years since then. Known as the Original Cohort, this sample represents children who, as of December 31st, 1994, were 0-11 years old and resided in one of the 10 Canadian provinces during collection in 1994-95. No new members have been added to this cohort since the first sample selection. The Original Cohort was comprised of 16,903 responding children at Cycle 1.

The NLSCY also draws a new birth cohort of 0-1 year-olds at every cycle in order to follow these children a few cycles for the purpose of evaluating early childhood development. These cohorts are referred to as the ECD cohorts. Since the ECD cohorts serve both cross-sectional and longitudinal analytical needs, both cross-sectional and longitudinal weights are produced for these children, while only longitudinal weights are produced for the Original Cohort (Cycle 4 was the last cycle for which cross-sectional weights were produced for the Original Cohort).

Diagram 2.1-1 describes the various cohorts of the NLSCY as of Cycle 8; the long arrow represents the Original Cohort while the shorter arrows indicate ECD cohorts. Each row in Diagram 2.1-1 is a cycle and the numbers indicate the ages of the children.

Age of children at each cycle, original cohort versus the early childhood development cohorts



Source: Statistics Canada, National Longitudinal Survey of Children and Youth.

Diagram 2.1-1: The various cohorts of the NLSCY for Cycles 1 to 8.

With only a few exceptions, the NLSCY samples have been drawn over the years from the Canadian Labour Force Survey (LFS)¹. Therefore, in order to understand the NLSCY sample design, a brief description of the LFS is required. (For further information on the methodology of the Canadian Labour Force Survey, consult [Statistics Canada \(2008\)](#)).

The LFS is a monthly survey that provides estimates of employment and unemployment at the regional and national level. It is also a major source of information on the personal characteristics of the working-age population, including age, marital status, educational attainment and family characteristics.

¹ The exceptions are a *Santé Québec* sample of children who were part of the Cycle 1 sample that was dropped at Cycle 2 due to budget cuts, and samples at Cycles 3 and 4 that were drawn from provincial birth registry data. At Cycle 7, of these non-LFS samples, only the Cycle 3 birth registry sample was still actively surveyed. All children surveyed at Cycle 8 were drawn from the LFS.

The monthly target sample size for the LFS is about 54,000 households. The LFS uses a rotating panel sample design, in which households remain in the sample for six consecutive months. The sample is split into six representative sub-samples (or rotation groups) and each month one-sixth of the sample is replaced after it has completed its stay in the survey.

Because the LFS has a large sample and collects information on characteristics of Canadian households, it is often used at Statistics Canada as a vehicle for sampling other surveys. For example, at every cycle, the NLSCY selects a new cohort of babies by identifying and sampling responding LFS households with children who are 0-1 year old. Over the cycles, the NLSCY has modified the way it samples children from LFS households. For example, for the Original Cohort, up to two children per household were selected, whereas for the ECD samples, only one child per household is sampled.

There are several benefits to using the LFS as a vehicle for launching another survey, one of which is convenience. However, the downside is that the LFS is designed to estimate labour force characteristics of the population 15 years of age and above and not, as we would like for the NLSCY, characteristics of children. Consequently, the LFS sample design is not the most efficient design for the NLSCY. For example, in order to obtain an adequate sample of 0-1 year-olds, the NLSCY needs to sample a large number of LFS panels (as many as 19 in some provinces). The LFS also has a complex stratified, multi-stage cluster design, using an area frame, which leads to a large range in the size of the LFS design weights. Any survey sampling from the LFS inherits this complex design which results in highly variable survey weights (see [Section 3.1](#) for more on this).

For more details on the NLSCY sample designs for Cycles 1 to 8, see [Tremblay \(2009\)](#).

2.2 An overview of the weighting and variance estimation processes of the NLSCY

At the conclusion of each cycle of the NLSCY, the released products typically include a microdata file containing upwards of 1,000 variables describing the collected data, longitudinal and cross-sectional survey weights (depending on the cohort), and a file of bootstrap weights (which permit the calculation of design-based sampling variances). All survey weights provided in the release are corrected for total nonresponse and post-stratified to match demographic projections of children by age, gender and province. Cross-sectional analyses, using the corresponding weights, are possible only in the early cycles of a cohort, as the elapsed time results in immigration and other factors creating a significant discrepancy between the target population (the one of interest) and the survey population (the population actually surveyed). On the longitudinal front, two different sets of longitudinal weights are provided: one for those who respond at every cycle (so-called funnel weights), and those who do not (so-called longitudinal, non-funnel or Swiss-cheese weights). There are a whole series of different weights depending on different populations of interest (cross-sectional versus longitudinal), and different patterns of longitudinal nonresponse (funnel versus non-funnel). For a description of different patterns of nonresponse, see [Section 4.0](#).

In the NLSCY, for any given population of interest and pattern of nonresponse (if applicable), there is only one set of weights which can be used. Consequently, the weights are determined before anything about the analysis to be performed is known. This, in itself, explains many of the issues that arise in a multi-purpose survey like the NLSCY (compared to a typical business survey, say).

Further exploration

Producing multi-purpose weights for a versatile survey like the NLSCY poses challenges to the creation and delivery of survey weights. To read more on this topic access [*Perspective A*](#).

The NLSCY uses the bootstrap as its variance estimation method. A brief introduction to the bootstrap used by the NLSCY is provided in the *Perspective* below while a more comprehensive treatment can be obtained from [*Girard \(2007\)*](#). For each release file of the NLSCY there is an accompanying bootstrap file.

Further exploration

To read more about the bootstrap used in a survey like the NLSCY, access [*Perspective B*](#).

Access to the released data is an issue for most analysts who either have to physically visit one of the Research Data Centres (RDCs), submit a request to Statistics Canada detailing the analysis to be performed (on a cost-recovery basis) or send Statistics Canada a computer program previously tested on a synthetic data file provided by the Agency. The program is run by the Agency's staff who then examine the results to avoid any disclosure of confidential respondent information.

Further exploration

To read more about the data dissemination procedures in the context of the NLSCY such as the RDCs, access [*Perspective C*](#).

3.0 General features of the NLSCY design weights

As mentioned earlier, longitudinal weights are calculated for children in the Original Cohort and for each ECD birth cohort. There are two sets of longitudinal weights: funnel weights, for those who respond to every cycle, and non-funnel weights for those who do not. Cross-sectional weights are calculated for ECD children who are in-scope for a given cross-sectional population. Cross-sectional weights were produced for the Original Cohort for Cycles 1 to 4, after which it was decided that because the Original Cohort had not been topped-up to ensure proper cross-sectional representation (e.g., to include recent immigrants) it was unwise to continue to produce cross-sectional weights for this cohort².

The final NLSCY survey weight (longitudinal or cross-sectional) is a child-level weight whose general form for the i^{th} child at time t is:

$$w_{i,final}(t) = w_i(0) \times adj_{i,nonresponse}(t) \times adj_{i,post-stratification}(t), \quad t \geq 1 \quad (3.0-1)$$

where

- t indicates the cycle (i.e., $t=1$ is Cycle 1),
- $w_i(0)$ is the inverse of the probability of selection for the i^{th} child, based on the NLSCY sample design,
- $adj_{i,nonresponse}(t)$ is the adjustment for the i^{th} child to correct for nonresponse at time t ,
- $adj_{i,post-stratification}(t)$ is the adjustment at time t for the i^{th} child (mainly to correct for frame undercoverage) that ensures that the sum of the final weights match (either exactly or closely) known or projected census counts of children by age, gender and province.

The general formula for the NLSCY child-level design weight for the i^{th} child is:

$$w_i(0) = w_{i,LFS\ subweight} \times adj_{i,eligible\ kids} \times adj_{i,rotation\ group} \times adj_{i,birth\ month} \quad (3.0-2)$$

Each of these components is described in the subsections below. For more details on the NLSCY design weight methodology, see [Tremblay \(2008\)](#).

3.1 The LFS subweight

The LFS subweight, $w_{i,LFS\ subweight}$, is a household-level weight that has been adjusted for LFS household nonresponse, but has not been post-stratified to LFS control totals.

Like most survey frames, the LFS suffers from undercoverage, which is referred to as slippage and which the NLSCY inherits. The LFS slippage rates for 0-11 year-olds at Cycle 1 (i.e., the Original Cohort) are around 20%; the national slippage rates for 0-1 year-olds (i.e., the ECD birth cohorts) range from 9% to 12%. The NLSCY corrects for undercoverage by post-stratifying the final

² For ECD cohorts with a similar issue, a less restrictive approach was taken: cross-sectional weights are produced, but a warning was introduced in the User's Guide informing users of the issue.

NLSCY weights to projected census counts of children by age, gender and province (*i.e.*, the adjustment $adj_{i, post-stratification}(t)$ in Equation (3.0-1)). The LFS weights inherited by the NLSCY may differ considerably from one child to another. Table 3.1-1 below gives an idea of the distribution of the LFS weights for the children of the Original Cohort in the province of Ontario.

Quantile (in %)	LFS subweight
0 (=min)	4
1	56
5	69
10	79
25	114
50	145
75	239
90	435
95	789
99	1,007
100 (=max)	1,649

Table 3.1-1: Distribution of the LFS weights for Cycle 1 of the Original Cohort, for children in Ontario (rounded to the nearest integer).

In the context of a discussion on logistic analyses, Section 8 of [Scott \(2006\)](#) warns about the consequences of having a ratio of maximum to minimum weights larger than 10. Only by using the 5th and 95th percentiles instead of the min and max values, respectively, do the Ontario weights fall within this guideline. Unequal weights become an even more important issue when looking at the final Cycle 1 weights, since the various adjustments applied to the LFS weights only accentuate the trend. Table 3.1-2, below, illustrates this in the case of the Original Cohort in Ontario. For the final weights, the ratio of the 95th to the 5th percentile is now over 13.

Quantile (in %)	Final Cycle 1 weight
0 (=min)	45
1	65
5	96
10	120
25	172
50	281
75	440
90	763
95	1,259
99	2,416
100 (=max)	4,919

Table 3.1-2: Distribution of the final weights for the Original Cohort kids in Ontario for Cycle 1 (rounded to the nearest integer).

Further exploration

To read about the problems posed by unequal weights to the stability of variance estimation, see [*Perspective D*](#).

3.2 The eligible kid adjustment

The eligible kid adjustment, $adj_{i, \text{eligible kids}}$, is the inverse of the probability of selection of a child being sampled in a given household. Depending on how the LFS rotation group is used, there may be more in-scope children in the household than are sampled by the NLSCY. For example, in Ontario most rotation groups at Cycle 7 were used to sample 0-1 year-olds (one child per household), while in the Maritimes some rotation groups were used to sample 0-5 year-olds (*i.e.*, a top-up of 2-5 year-olds was not required in the Ontario Cycle 7 sample). Thus, the eligible child adjustment is:

$$adj_{i, \text{eligible kids}} = \frac{\text{number of in - scope kids for the NLSCY in the LFS household}}{\text{number of selected kids}} \quad (3.2 -1)$$

where the number of selected children may be as many as two for the Original Cohort, and is usually one for the ECD cohorts.

3.3 The rotation group adjustment

The rotation group adjustment reflects the fact that summing the weights of one LFS rotation group represents only 1/6th of the Canadian population of households. Consequently,

$$adj_{i, \text{rotation group}} = 6 \quad (3.3 -1)$$

3.4 The birth month coverage adjustment

This weight adjustment factors in the total number of rotation groups used to sample a particular age group. In the case of 0-year-olds this adjustment also corrects for any undercoverage by the LFS rotation group. Because of the way in which the NLSCY samples from the LFS, some rotation groups do not cover the entire reference birth year for 0-year-olds. For example, in Cycle 7, a 0-year-old was a child born in 2006, but the LFS rotation group used for sampling may have rotated out midway through the year, in which case it only covered half of the reference year, and so the weights need to be adjusted to reflect this (see Figure 3.4-1). At Cycle 6, a gross adjustment was performed to account for this undercoverage; at Cycle 7 the birth-month adjustment was refined.

	2006											
Panel	Jan.	Feb.	March	April	May	June	July	August	Sept.	Oct.	Nov.	Dec.
1	Recorded LFS births that can be selected for NLSCY sampling											
2												
3												
4												
5												
6	Births within LFS panels that are not captured											
7												
8												
9												
10												
11												
12												

Figure 3.4-1. A table representing births in 2006 that are not captured by the LFS files. The white space represents the missed births by month and panel to which the household belongs.

For sampled children who are 1 year old or older the LFS rotation group has complete coverage of the reference birth year in which case this adjustment simply corrects for the number of rotation groups used to sample that age group, and is thus:

$$adj_{t, \text{birth month}} = \frac{1}{M} \quad (3.4-1)$$

where M = total number of rotation groups used to sample the age group in a given province. (For children under the age of 1, see [Tremblay \(2008\)](#).)

Now that the design weight for the NLSCY has been described, we can move on to describe how nonresponse affects the weighting methodology.

3.5 The *histoire* file

The *histoire* file is the repository for basic information about any unit that has ever been part of the NLSCY sample, for all cycles that the unit has been involved in. It thus provides historical information related to sampling, response status, weighting, master file inclusion status, household membership, etc.

The *histoire* file was introduced early on in the NLSCY, but over the years it was not consistently maintained and so inconsistencies appeared. Recently, a refurbishment of the *histoire* file was undertaken and completed, and the extent of that undertaking showed just how difficult it is to keep track of relevant sampling and collection information for a sampled unit in a longitudinal survey. Since then, time and resources have been devoted to keeping it up-to-date. Actually, the *histoire* file has now become the cornerstone of the sampling and weighting processes. For example, it is used to determine the returning sample and to identify which units should be given a cross-sectional and/or a longitudinal weight on the master file. The *histoire* file is also the source for the

calculation of various response rates. Finally, the *histoire* file is often used to investigate special cases and to provide answers to *ad hoc* questions regarding the different cohorts of the NLSCY.

The necessity of having an *histoire* file in top shape is one of the important lessons learned with the NLSCY: the need to centralize all the longitudinal information.

4.0 Nonresponse in the NLSCY

All longitudinal social surveys are subject to various types of nonresponse, and the NLSCY is no exception. Nonresponse is of concern to methodologists because in addition to depriving the user of important analytical power due to the reduction of usable sample size, it may actually bias inferences if gone unaddressed. This problem is aggravated when the nonresponse mechanism is not missing at random (NMAR) or non-ignorable (for a discussion of these concepts, see [Section 5.1](#)). Moreover, proceeding to address the bias is unnerving because, if ill-advised, the methodology developed for that very purpose may end up doing more harm than good.

One type of nonresponse is total nonresponse (also called unit nonresponse), where all data items for an individual are missing at a given cycle. Another type is partial nonresponse (also called item nonresponse) where data are missing at a given cycle for one or more items, or even an entire block of items. The NLSCY weights only compensate for total nonresponse. Item nonresponse, as it affects a particular analysis, should be addressed by the analyst (*e.g.*, by performing a nonresponse weight adjustment that is tailored to the analysis - see [Section 4.2.2](#)).

With respect to total nonresponse, longitudinal surveys can have either a monotonic or non-monotonic design, also referred to as funnel and non-funnel, respectively. These terms refer to how at a given cycle we handle total nonresponse that has occurred at previous cycles. If we define converted units to be respondents to the current cycle who were nonrespondents to at least one earlier cycle, then we can say that a non-monotonic design allows for converted units, while a monotonic design does not. Put another way, with a monotonic design a nonresponding unit is never surveyed at subsequent cycles. The presence of converted units allows for higher response rates to be achieved at the current cycle which is why, at Cycle 3, the NLSCY switched to a non-monotonic design for the Original Cohort. The drawback of this approach is that the nonresponse weighting methodology required to account for converted units can be more complex than if a monotonic approach had been used.

It is common now to refer to a non-monotonic design as the Swiss cheese approach (because of the holes in the cumulative file any episode of nonresponse creates) while a monotonic design is often referred to as the funnel approach. These different types of nonresponse are illustrated in Figure 4.0-1.

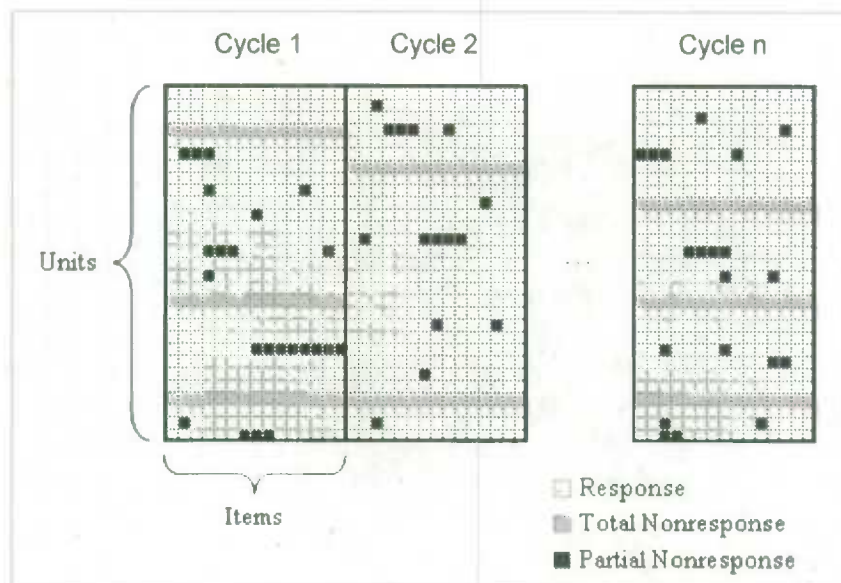


Figure 4.0-1. Illustration of the Swiss cheese approach (or the non-monotonic or non-funnel design)

In longitudinal surveys the problem of nonresponse is magnified by the fact that a unit may experience several episodes of cycle nonresponse over the life of the cohort. On the bright side, it also means that there is usually plenty of information to use to help reduce the impact of nonresponse at estimation.

Another challenge raised by nonresponse is to determine the proper variance estimation procedure that will account for nonresponse and the nonresponse treatment that is applied to the survey estimates. As we mention in *Perspective B*, the NLSCY and many other surveys at Statistics Canada use the bootstrap as their variance estimation procedure. This is because the weighting methodologies that these surveys use to handle total nonresponse can be well accommodated by the bootstrap, provided it is handled carefully (we will return to this topic in [Section 5.1.2](#)). However, proper variance estimation becomes more of a challenge when it comes to partial nonresponse and imputation. SEVANI ([Beaumont and Mitchell \(2002\)](#)) is a tool developed at Statistics Canada to provide estimates of variance in the presence of imputation, but it is not based on the bootstrap method.

Another issue associated with nonresponse is the loss of analytical power. Because of the missing items, the analysts end up with a reduced effective sample size for their analysis and the variance increases. In the context of longitudinal surveys and attrition, it can reach the point where it might not be possible to study some analytical domains at a certain point in the lifespan of the survey. Training and informing external data users regarding the treatment of nonresponse is also a challenge for a statistical agency like Statistics Canada, which operates in a framework where privacy and disclosure policies are crucial. It should be noted however that this training is crucial given that data users do not necessarily have a strong survey sampling background (or any statistical background) and they often tend to simply ignore the partial nonrespondents, discarding them without adjusting the weights to compensate for them, running the inherent risk of biased conclusions. In this context, whatever an analyst chooses to do to address the partial nonresponse that arises, he or she should make it explicit in the analysis and findings how nonresponse was dealt

with. It would also be important for the analyst to evaluate and report on the impact of the decisions taken to cope with the nonresponse.

4.1 Total nonresponse

As already mentioned, total nonresponse occurs when there is no useful survey information that was collected about a unit (a child in the case of the NLSCY). This is the type of nonresponse we seek to address through the final weights released to data users.

Although the total nonresponse rate at each cycle is relatively small, it accumulates quickly as the number of cycles grows as shown by Table 4.1-1 below about the Original Cohort. (Since the first cycle sampled households in the field, and then children, we can measure the nonresponse at the household level or the child level. Unless otherwise stated, total nonresponse refers to child-level nonresponse.) We can see that the Cycle 7 cumulative response rate is 57%, which means that we must account for the loss of 43% of initially sampled units.

<i>Cycle</i>	<i>Cumulative Response Rate (Household Level)</i>	<i>Cumulative Response Rate (Child Level)</i>	<i>Nonresponse Design</i>
Sampling Frame (LFS)	100%	N/A	N/A
1	87%	87%	Monotonic
2	79%	79%	Monotonic
3	77%	76%	Non-monotonic
4	68%	68%	Non-monotonic
5	65%	63%	Non-monotonic
6	61%	58%	Non-monotonic
7	61%	57%	Non-monotonic

Table 4.1-1 Longitudinal response rates for the NLSCY Original Cohort³.

This significant potential for bias must be addressed somehow, but how? What are the options? This important issue is tackled in detail in [Section 5.0](#).

4.2 Partial nonresponse

Partial nonresponse occurs whenever some items - but not all – that are asked of a unit are left unanswered. The preferred treatment option for partial nonresponse is usually imputation as it gives a complete (or square) file for the analysts and addresses the loss of analytical power ([Brick and Kalton \(2003\)](#)). However, variance estimation in the presence of imputation is a challenge. It can be done in theory with the bootstrap method by repeating the imputation process for each replicate ([Shao and Sitter \(1996\)](#)). This would require in practice the development of a new analytical tool as none currently exists providing such a feature. This approach would also require that the imputation process be simple enough to be automated in the analytical tool without affecting the quality of the imputations.

³ The NLSCY only samples LFS respondents. The NLSCY design weight uses an LFS weight that has been adjusted for LFS nonresponse (see [Section 3.1](#)). The LFS monthly response rate is over 90%.

An alternative approach advocated by some is multiple imputation ([Rubin \(1987\)](#)) which is basic imputation repeated multiple times. The analysis is then performed using, in turn, each of the proposed values, and the estimates obtained that way are combined into one final estimate and their variability enters the variance which is reported for the (aggregated) estimate. Multiple imputation has been integrated into SUDAAN[®] Version 9, which is one of the existing analytical tools able to properly handle variance estimation when using data coming from surveys with complex sample designs. Considering it is already a challenge for analysts to use the data file (in which they find just one occurrence of each variable – the reported data) that is currently disseminated by each of Statistics Canada's social surveys, making available the multiply imputed data could overwhelm some of them. This being said, more advanced data users who currently consider a multiple imputation approach to deal with partial nonresponse from our surveys would certainly greatly appreciate any efforts made in this direction by us, the data providers.

Setting aside the variance estimation matter, there are some other practical issues that have to be considered with imputation, such as the risk of creating inconsistencies at the micro record level. In the context of longitudinal surveys, both the cross-sectional covariance structure of the data and the covariances across cycles are of concern, and with social surveys composed of many variables (often categorical), taking into account everything is simply not possible. Furthermore, the preservation of the longitudinal covariance structure of the data may require the revision of previously released imputations when data for a new cycle become available, thus resulting in the re-release of previously disseminated data files. (This is not desirable since any re-release could be inconsistent with an earlier one and could then create more confusion than anything else.) Implementation issues such as the time required to carry out the actual imputations and verifications without delaying too much the dissemination of the data also have to be considered. For example, the released data file for the NLSCY already contains over 1,000 variables to which an imputation flag would have to be added for each imputed variable. In such a setting, it becomes essential to focus on the imputation of a subset of key variables that would be of most benefit to the end users. In addition, while imputing a numerical variable like income is customary nowadays, it is much trickier and controversial to impute social data, such as an individual's score on a depression scale.

4.2.1 The current situation regarding imputation

For the reasons mentioned above, not much imputation is currently being done in our longitudinal social surveys at Statistics Canada. All six surveys impute income-related questions as they mostly consist of continuous variables that can be fairly easily and reliably imputed. The imputation strategy employed in the Survey of Labour and Income Dynamics (SLID) is more sophisticated as the income variables are very important in the context of the survey and a great deal of information closely related to income is collected. In the first five cycles of the NLSCY, total household income was imputed using a cross-sectional hot-deck imputation approach. Although this imputation method preserves the cross-sectional distribution of income, it does not ensure the longitudinal coherence of the imputed data. In Cycle 6 of the NLSCY, the imputation strategy was changed to overcome this issue when a longitudinal hot-deck imputation method was introduced, providing a reasonable balance between preserving the cross-sectional and longitudinal distributions of income.

In the NLSCY, the Motor and Social Development (MSD) items are also imputed for records with one or two missing items out of 15 using a simple cross-sectional hot-deck approach. This allows the imputation of about 80% of the partial nonrespondents and the derivation of a MSD score for these records. This is a good example of where a rather simple imputation effort is translated into a non-negligible gain for the analysts using the data. In the Longitudinal Survey of Immigrants to Canada (LSIC), mass imputation is used in some sections of the questionnaire in order to treat partial nonresponse. All the variables of the questionnaire's section are imputed, even the reported items, using the same donor record. Compared with filling only the holes (*i.e.*, impute only the missing items), this approach has the advantage of preserving the internal micro-record coherence, and thus the cross-sectional covariance structure, as the data come from a complete and coherent record (the donor). It should be noted, however, that the micro-record level coherence across cycles is not automatically ensured with mass imputation, although the situation is less problematic than with other strategies aimed at imputing only the missing items. Another advantage of mass imputation is that it presents fewer implementation challenges since it is quicker to implement and fit into the production schedule. However, variance in the presence of mass imputation remains an issue to this day. The most obvious and serious drawback of not resorting to imputation in such a case is that reported data may end up being discarded by the analyst who usually does not have the means to deal with partial nonresponse.

4.2.2 Future plans for treating item nonresponse

Statistics Canada plans to do more active work regarding the treatment of partial nonresponse in its longitudinal social surveys. In the particular case of the NLSCY, the plan is to work on three fronts. First, improve the training of the end users regarding the partial nonresponse issue and how it can be dealt with instead of simply being ignored. This has already started with a workshop addressing the issue, where imputation and weighting are covered as potential treatments for nonresponse (see [Lafortune \(2008\)](#)).

Second, the development of a custom weighting tool that analysts could use to treat the partial nonresponse affecting their variables of interest is currently being considered. Similar to the weighting process done for total nonresponse (which is discussed at length in [Section 5.1](#)), partial nonresponse would be analysed using a logistic regression approach in order to group into Response Homogeneous Groups (RHGs) the respondents and partial nonrespondents who have a similar propensity to respond. In the end, the weights of the partial nonrespondents would be redistributed among the remaining respondents within each RHG. Although the records affected by item nonresponse in the subset of variables involved in the analysis would still be excluded and some reported data would end up being thrown away, the potential bias would at least be addressed. An interesting feature of this approach is that variance estimation would not be an issue as the tool would also provide the user with a set of adjusted bootstrap weights to account for the partial nonresponse weighting process. We believe that such a tool would offer a rather simple and appealing alternative to analysts for the treatment of partial nonresponse.

Third, the NLSCY is considering increasing the number of variables to be treated by imputation. As mentioned previously, considering the rather large number of disseminated variables in the NLSCY data file, the plan will likely never be to impute all variables, but rather to focus on some

key variables and build from there in a step-by-step approach. We still, however, need to carefully weigh the benefits and issues of going further into the realm of imputation.

5.0 Addressing total nonresponse

There are essentially three options available to a methodologist in a longitudinal survey setting to address total nonresponse: calculate a nonresponse weight adjustment for each Response Homogeneous Groups (RHG), calibrate or mass impute. Since only the first two have been considered seriously in the context of the NLSCY, the following discussion focuses solely on them.

As we are about to dig deeper into nonresponse issues, it is worthwhile for the practitioner to give some thought to the entire endeavour of mathematical modelling. Indeed, the issue of addressing nonresponse in survey sampling is one place where a heavy use is made of models.

Further exploration

To read more on the mathematical activity of modelling, see [Perspective E](#).

Any effort invested in nonresponse modelling is made towards reducing the risk of potential bias. While the spectre of bias is easily perceived, in practice it is a very difficult task to address it head-on. Indeed, the very concept of bias calls for a quantity that we never know, namely the true value, and there rarely is a suitable proxy for it in a household survey. One indirect measure of it that we have devised in the NLSCY is that of longitudinal consistency. Essentially, longitudinal consistency claims that the most one can do in practice to limit bias is to ensure that with the current set of final survey weights one can reproduce earlier cycles' estimates, which used a previous set of weights. So, with longitudinal consistency, the idea is to ensure that the present is consistent with the past.

Further exploration

To read more on how to measure potential bias, see [Perspective F](#).

In this section we present the various issues and options available to us and in the following section we describe in detail what this all means in the context of the NLSCY.

5.1 Weighting

Weighting within RHGs involves the construction of a nonresponse model which, loosely speaking, seeks to explain the nonresponse that occurred using available information at hand. With this approach, one tries to thwart the effect that nonresponse is likely to have on the estimates by understanding how it operates. (This is akin to understanding the cause of an illness in order to have any hope of relieving the patient's symptoms.) To get a grip on nonresponse this way, we model its action as if it were an additional sampling mechanism forced upon us. We integrate the sampling we did perform with this imposed nonresponse "sub-sampling" into one scheme to help cope with the weighting and variance issues that arise. This is why the mathematical framework used to integrate RHGs into the overall estimation scheme is borrowed from the two-phase sampling (or multi-phase if the sampling itself involves more than one phase). If the nonresponse mechanism can reasonably be described this way, then its effects become more predictable and can

be (to some extent) thwarted. Nonresponse weight adjustments revise the weights initially assigned to responding units to compensate for the loss of nonresponding units (and their characteristics and their weights).

In the NLSCY, as with most of the other longitudinal social surveys at Statistics Canada, the nonresponse that occurred at a given cycle is addressed through weighting within RHGs or weighting classes (see, for instance, [Lohr \(1999\)](#), p.266). With this technique, units that have a similar propensity to respond (which has to be assessed somehow – this is where the main challenge with the approach lies) are grouped together, where each group contains both respondents and nonrespondents. The RHG approach is used when there are indications that the response mechanism is missing at random (MAR) or ignorable, (*i.e.*, when nonresponse does not depend on the variable of interest but rather on some auxiliary information X_1, X_2, \dots, X_p available for all units in the sample, as explained by [Little and Rubin \(1987\)](#)). In other words, while it may not be sensible to sustain the claim that nonresponse is uniform for the whole sample in a given context, it may be a reasonable assumption to make within well-defined subgroups.

Strictly speaking, while it is desirable in practice to use information which is available for the whole sample, we can do quite well if it is missing for some units. Indeed, the misclassification of these units into some “catch-all” group may lead in principle to a bias, but it is assumed to be small compared to the bias introduced by not using the incomplete information at all. This is actually a good place to reiterate the general point made with [Perspective E](#) about models as it applies here: we do not need to identify the perfect set of explanatory variables (should one such set even exist) to create a useful model – any significant information incorporated into a model will prove to be beneficial to address bias. It is not a matter of all or nothing: we ought to build the model using all the information at hand, even if it proves to be limited.

The theoretical support for this approach assumes that the following probability statement (from [Särndal et al. \(1992\)](#) Equation (15.6.6)) holds:

A given sample s splits into disjoint groups, the RHGs, indexed by h , for which the following holds:

$$p(k \in R_h | s_h) = \theta_{s_h} \neq 0 \quad \text{for all } k \in s_h \text{ and independently among units.} \quad (5.1-1)$$

where s_h is the set of sampled units found in the h^{th} RHG and R_h is the set of respondents.

This one (theoretic) probability per group, θ_{s_h} , is then somehow estimated and its inverse is used in the nonresponse adjustment. More precisely, the inverse of the product of the selection probability with the response probability is the weight after nonresponse. But how should that probability be estimated? It may come as a surprise to the reader to learn that in [Särndal et al. \(1992\)](#), this probability is evaluated as an unweighted ratio of the number of respondents in the h^{th} RHG to the entire sample size of the h^{th} RHG:

$$\hat{\theta}_{s_h} = \frac{r_h}{n_h} \quad (5.1-2)$$

Thus, it is the observed response rate within the h^{th} RHG, not a weighted response rate (see 15.6.7a on page 579). On the other hand, [Lohr \(1999\)](#) advocates using the weighted response propensity (see Section 8.5.1):

$$\hat{\theta}_{s_h, wgt} = \frac{\sum_{k \in R_h} w_k}{\sum_{k \in S_h} w_k} \quad (5.1-3)$$

where w_k is the design weight.

Which one to use? In practice, the use of (5.1-2) is done in conjunction with some other weighted adjustment so that in the end it gives rise to an adjustment which matches (5.1-3), which is how many surveys at Statistics Canada calculate this adjustment.

Further exploration

To read more on how the response propensities are evaluated in practice, see [Perspective G](#).

The theory just outlined assumes these RHGs already exist (that they were provided to us somehow) but in practice we need a way to create them before we can use the framework that we introduced above. This is one major challenge in practice since the construction of RHGs, which is meant to identify major sources of potential bias, leads to issues of its own. Indeed, though it appears simple, the RHG method can be implemented in a surprising variety of ways, including segmentation modelling and the response propensity scoring method (described below). Furthermore, for any given method, there is a wide array of information in a longitudinal setting that can be used for modelling, and decisions have to be made about how best to integrate this information.

Segmentation modelling uses the Chi-square Automated Interaction Detection Algorithm (CHAID) as suggested by [Kass \(1980\)](#) to generate the RHGs as collections of units with the same values for a subset of nonresponse predictors. (It is mainly known, though, through the software product name KnowledgeSeeker[®] which is typically used to perform the CHAID.) The idea here is to perform sequential chi-square tests to find significant differences between response and nonresponse groups (and between sub-groups of each of these groups, and so on).

For example, if males and females exhibit strongly different response rates, two groups could be built: one containing the males, the other containing the females. Within gender, further sub-groups could be identified if they exhibit very different response rates, and so on. When this investigation is over, the RHG is obtained by crossing all of the groups identified. Consequently, in practice, many more RHGs end up being created because several variables are usually identified by KnowledgeSeeker[®], which leads to a graphical representation known as a KnowledgeSeeker[®] tree. Figure 5.1-1 below is an example of such a tree, taken from the Cycle 2 longitudinal weighting documentation of the Original Cohort. It shows which variables enter the RHGs in the province of Ontario. For instance, we see that within Ontario the most significant variable that explains nonresponse is whether the PMK had a spouse (“Présence du conjoint” means “Presence of a

spouse”). We therefore conclude that at Cycle 2 the methodology based on KnowledgeSeeker[®] led to the creation of 17 RHGs solely within Ontario. (This number is obtained by counting the number of “leaves” the tree has).

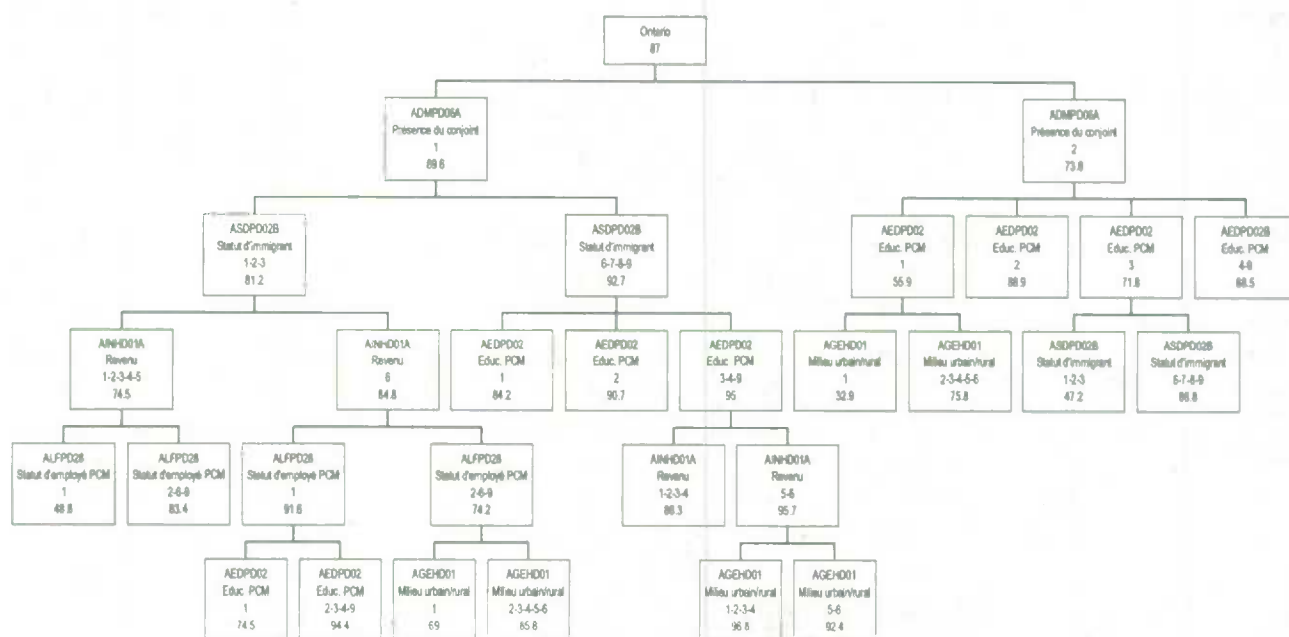


Figure 5.1-1: An example of a KnowledgeSeeker[®] tree, used for the NLSCY Cycle 2 longitudinal weights.

For the whole sample we ended up having 36 RHGs at Cycle 2, which is on the high side, but considerably fewer than the 400+ that were created at Cycle 1 for the Original Cohort. Upon reflection, though, it is to be expected that when RHGs are obtained by crossing variables, there will be a large number of them: very few are needed to create a large number of cross-categories. (This is not without consequence to weighting and variance estimation as we shall see in [Section 5.1.2](#)) The large number of RHGs created with this approach would be acceptable if there were evidence that so many weighting classes were necessary. But from what we observe with the RHGs at Cycle 2, ranking the distinct possible nonresponse adjustments from smallest to largest leads to successive adjustments differing only by 0.1% on average. This occurs because when KnowledgeSeeker[®] creates the homogeneous groups it does not monitor whether different groups have similar response rates.

Let us illustrate with an artificial example. Suppose at the first iteration we group records according to province because we observe that nonresponse varies geographically. At the second iteration, suppose that within each province we feel justified to further group records according to gender. Here is the trap: it is not because the first 10 groups of the first iteration had different response propensities that all 20 groups obtained at the end of the second iteration will. Indeed, it may very well be that two groups, say Gender *X* - Province *A* and Gender *Y* - Province *B*, have, in the end, about the same estimated response propensity. And if that is the case, then why should we maintain separation between these two (which would yield two RHGs)? Why not combine them instead to get just one RHG, one of a larger size? After all, based on estimated response propensities, there is nothing to justify keeping these groups separate.

With CHAID we group records according to their characteristics and nonresponse patterns but fail to track down the response propensities as further sub-groups are created. Consequently, we do not check whether the final groupings lead to estimated response propensities that are significantly different or not. And actually, knowing that over 400 response propensities are estimated, it is nearly impossible for all of these to co-exist and also be significantly different from one another.

Whether a tight watch on the ensuing estimated response propensities is kept or not, it remains that the segmentation approach does favour the proliferation of RHGs. An approach that was investigated for Cycle 6, and ultimately implemented, was to derive RHGs using the response propensity scoring method ([Little \(1986\)](#), [Eltinge and Yansaneh \(1997\)](#)). With this method, response propensities are estimated using a logistic regression model and then similar response propensities are grouped together. The idea with this approach is in some way the reverse of the segmentation approach. Indeed, while the segmentation approach groups together individuals based on some deemed-relevant characteristics (like, say, age and gender) and then assesses their response propensities, the logistic method first models the response propensities of individuals and then groups them based on similar predicted response propensities. As a result, individuals of a given RHG may only have in common a similar response propensity and not any obvious similar profile. In other words, while age and gender may have driven the segmentation approach, in a corresponding logistic-based RHG one may find individuals of all ages and both genders.

Forming the groups can be done by sorting the estimated probabilities and grouping them into deciles, or by using clustering algorithms to find more natural groupings of the estimated probabilities ([Haziza and Beaumont \(2006\)](#)). Consequently, the practitioner has more control over the number of RHGs to be created using the logistic approach than with segmentation. (A Monte Carlo study using data extracted from the Longitudinal Survey of Immigrants to Canada (LSIC) found the response propensity scoring method to be more robust than segmentation modelling for different types of nonresponse mechanisms; see [Alavi and Simard \(2006\)](#). [Haziza and Beaumont \(2006\)](#) also report several advantages of the response propensity scoring method over the segmentation modelling approach.)

One important issue is the role the weights get to play in the construction of the nonresponse model.

Further exploration

To read more on the use of weights in the construction of RHGs, see [Perspective H](#).

But even if at a given cycle only a few RHGs are created there remains the issue of how to combine several cycles' worth of nonresponse in a sensible way. Consequently, whatever approach we use for weighting, we first need to precisely determine what information we will use in the model itself at a given cycle. This is the focus of the next subsections.

5.1.1 Modelling longitudinal nonresponse: *saut-de-mouton* versus *saut-de-l'ange*

Whereas the RHG methodology described above can be readily implemented in any cross-sectional survey, more factors need to be considered in the case of a longitudinal survey. To develop a coherent methodology across all cycles, we need to consider how the strategy at one cycle is related to that of other cycles. Two general approaches were considered: the *saut-de-mouton* (literally, a sheep's jump) and the *saut-de-l'ange* (an angel's leap). The *saut-de-mouton* approach treats nonresponse one cycle at a time. Each cycle adds its own weight adjustment factor that corrects for the nonresponse at that cycle. To adjust for the cumulative nonresponse, the weight adjustment factors for all cycles are multiplied together. The *saut-de-mouton* approach is so called because it involves a step-by-step (or cycle-by-cycle) methodology. Conversely, the *saut-de-l'ange* approach addresses the cumulative nonresponse in just one step.

5.1.1.1 Model building: considerations when applying the *saut-de-mouton*

With the *saut-de-mouton* approach, nonresponse is modelled at each individual cycle and then the nonresponse adjustments obtained are combined into one nonresponse adjustment for use in the most current cycle through:

$$w_i(t) = w_i(0) \times \frac{1}{\hat{\phi}_i(1)} \times \cdots \times \frac{1}{\hat{\phi}_i(t-1)} \times \frac{1}{\hat{\phi}_i(t)}, \quad t \geq 1 \quad (5.1.1.1-1)$$

where $w_i(t)$ is the nonresponse adjusted weight at Cycle t , $w_i(0)$ is the design weight, and $\hat{\phi}_i(t)$ is the response probability at Cycle t , which for the NLSCY is computed as the weighted response rate of the RHG; that is,

$$\hat{\phi}_i(t) = \frac{\sum_{k \in R(t) \cap RHG_i(t)} w_k(t-1)}{\sum_{k \in RHG_i(t)} w_k(t-1)} \quad (5.1.1.1-2)$$

where $R(t)$ is the set of respondents at Cycle $t \geq 1$, and $RHG_i(t)$ is the RHG at Cycle t containing unit i .

One attractive feature of the *saut-de-mouton* methodology is that it can use recent information to model the latest nonresponse episode. Thus, in the spirit of “explaining nonresponse” which we presented in [Section 5.1](#), we consider information that is closest in time to the nonresponse event “to be explained”. It is particularly appealing for monotonic designs since there are a wealth of data with which to model nonresponse: only respondents at Cycle $t-1$ enter into the nonresponse model and they were respondents at all previous cycles. The main problem with the *saut-de-mouton* methodology is that the nonresponse model for a given cycle is often developed independently of the previous cycles' nonresponse models. In particular, with the *saut-de-mouton* there is the temptation to model nonresponse at time t using only variables appearing at time $t-1$, in which case interaction terms between cycles can be all too easily missed. If the variables used to create the

RHG in previous cycles do not appear in any of the variables at $t-1$, then these variables cannot enter into the model.

Under a *saut-de-mouton* methodology, determining if interaction terms are important in describing cumulative nonresponse can be done by:

- searching at each cycle for interaction terms related to cumulative nonresponse and including them in the current cycle's nonresponse model;
- creating variables indicating the RHGs at all previous cycles and seeing if these are important predictors in the current cycle's nonresponse model (because these variables act as proxies for some of the variables the first approach would identify);
- nesting Cycle t RHGs into the ones of previous cycles (again, because these variables act as proxies for some of the variables the first approach would identify, but in a more rigid framework than the second approach).

All of these solutions keep the *saut-de-mouton* on track, so that not only are nonresponse models at each cycle appropriate, but the global picture that they create of cumulative nonresponse is also correct.

Note that in this context, the question of interaction terms amounts to determining if the cycles of nonresponse are independent. Verifying the independence assumption is critically important. If they are independent, then we can model nonresponse separately at each cycle and not worry about interaction terms across cycles. If they are dependent then interaction terms are important in describing the cumulative nonresponse and we need to account for this in our methodology (possibly through some mixture of the approaches listed above).

Nesting RHGs is possible whether the survey is monotonic or non-monotonic. Suppose at the first cycle we build ten RHGs: $RHG_1, RHG_2, \dots, RHG_{10}$. At Cycle 2, within each of the Cycle 1 RHGs, we could model the nonresponse occurring at Cycle 2 and build RHGs. For example, RHG_1 would be further broken down into, say, 10 RHGs. Repeating this process for all Cycle 1 RHGs gives us 100 Cycle 2 RHGs. Given how quickly the number of RHGs can grow with the number of cycles, nesting RHGs is a viable solution for surveys that last only a few cycles and that have only a few RHGs at each cycle. Moreover, as the number of RHGs increases, the number of units in each RHG decreases. This can quickly lead to very large weight adjustments across RHGs, not to mention their instability under bootstrap replication. Calculating the sampling variance can also become problematic due to small, or empty, RHGs. Finally, note that blindly nesting RHGs can also create more groups than are necessary since the Cycle t nonresponse may not be dependent on every RHG of all previous cycles.

To help sort this all out, let us consider an artificial example:

A cohort of 10 men and 10 women is selected from a population of 1,000 men and 1,000 women using a stratified simple random sample (*i.e.*, gender defines the strata). This cohort is followed through time and is surveyed at Cycles 1 and 2 using a monotonic nonresponse design. Our strategy for modelling nonresponse here is the *saut-de-mouton* using segmentation modelling, with Equations (5.1.1.1-1) and (5.1.1.1-2) in mind. We consider different applications of this strategy to this simple case to show how they yield different results regarding longitudinal consistency.

Because all of the sampled units have the same initial design weight, $w_i(0)$, a weighted and unweighted nonresponse model at Cycle 1 would yield equivalent response propensities. At Cycle 1, the only frame information available for modelling nonresponse is gender. If we were to ignore gender, the only alternative would be a uniform nonresponse adjustment. Since only 40% of women responded while 80% of the men did, this would clearly lead to longitudinally inconsistent estimates. In this example, if we do consider gender when modelling nonresponse then the male (M) and female (F) groups would be our Cycle 1 RHGs. The resulting Cycle 1 nonresponse adjusted weights, $w_i(1)$, rounded to the nearest integer, are shown in Table 5.1.1.1-1. The third column of Table 5.1.1.1-2 shows how these weights give us our desired estimated population totals by gender: 1,000 men and 1,000 women.

At Cycle 2, still using the *saut-de-mouton* approach, we wish to model nonresponse that has occurred between Cycles 1 and 2, using the Cycle 1 nonresponse adjusted weights, $w_i(1)$. These weights reveal a ratio of 1:1 males to females among Cycle 2 respondents (*i.e.*, Cycle 2 nonresponse does not affect women differently than men), but they reveal different response propensities by the Cycle 1 variable poverty. Suppose that we ignore gender and only use poverty to create two Cycle 2 RHGs: poor (P) and not poor (NP). Using the resultant Cycle 2 nonresponse adjusted weights, $w_i(2)$, Table 5.1.1.1-2 shows a shift in the estimated population ratio of men to women to 1,029:972. In other words, the Cycle 2 survey estimates of gender are longitudinally inconsistent. Ignoring interaction terms is problematic. In this case we did not consider interaction terms between the Cycle 1 and Cycle 2 RHGs.

	<i>Gender (Frame)</i>	$w_i(0)$	<i>Cycle 1 Response Status</i>	<i>Poverty (Cycle 1)</i>	$w_i(1)$ (M,F)	<i>Cycle 2 Response Status</i>	$w_i(2)$ (P,NP)	$w_i(2)$ (M,F&P, F&NP)
1	F	100	R	P	250	NR	0	0
2	F	100	R	P	250	R	400	500
3	F	100	R	NP	250	R	286	250
4	F	100	R	NP	250	R	286	250
5	F	100	NR		0			
6	F	100	NR		0			
7	F	100	NR		0			
8	F	100	NR		0			
9	F	100	NR		0			
10	F	100	NR		0			
11	M	100	R	P	125	R	200	167
12	M	100	R	P	125	R	200	167
13	M	100	R	P	125	R	200	167
14	M	100	R	P	125	NR	0	0
15	M	100	R	NP	125	R	143	167
16	M	100	R	NP	125	R	143	167
17	M	100	R	NP	125	R	143	167
18	M	100	R	NP	125	NR	0	0
19	M	100	NR		0			
20	M	100	NR		0			

Table 5.1.1.1-1. A *saut-de-mouton* example showing the dependence of RHGs across cycles and the importance of interaction terms when adjusting for survey nonresponse. Below each weight, the RHGs are listed in parentheses.

<i>Gender</i>	$\sum_S w_i(0)$	$\sum_{R(1)} w_i(1)$	$\sum_{R(2)} w_i(2)$ (P,NP)	$\sum_{R(2)} w_i(2)$ (M,F&P,F&NP)
M	1,000	1,000	1,029	1,000
F	1,000	1,000	972	1,000

Table 5.1.1.1-2. The problem of longitudinal consistency when the dependence of RHGs across cycles is ignored, using the weights in Table 2.

At Cycle 2, we could adopt a different approach. Suppose we considered all available variables and allowed for interaction terms across cycles. For comparison, let us again build a weighted model using $w_i(1)$. We notice three groups are affected differently by nonresponse at Cycle 2: male, female and not poor, and female and poor. Cycle 2 nonresponse equally affects poor men and men who are not poor. Defining three RHGs: male, female and not poor, and female and poor, gives us longitudinally consistent population estimates by gender (Table 5.1.1.1-2).

So in this example, longitudinal consistency is preserved if RHGs are nested across cycles. Let us now consider the alternatives to nesting. Suppose we were able to use both gender and poverty for the Cycle 2 adjustment. If the identification of the nonresponse predictors were done at each cycle in an unweighted fashion, we would pick up gender at Cycle 1 and only poverty at Cycle 2,

resulting in the same consistency problem as before. On the other hand, if we identified the Cycle 2 nonresponse predictors using the Cycle 1 nonresponse adjusted weights we would notice that gender does play a role, albeit a much smaller one than the poverty variable. Depending on the thresholds set for identifying the nonresponse predictors when building the RHGs, the gender variable still could have been missed at Cycle 2 even if the weights were used. So, an un-nested approach can also perform well if RHG variables from all previous cycles are involved in modelling (to protect against weak correlations between variables from different cycles) and if the modelling is carried out using the previous cycle nonresponse adjusted weights.

A few general remarks about nesting are in order. Nesting can be viewed as a type of calibration, that is, previous RHGs that are respected by those of the current cycle are essentially acting as calibration cells. (Further discussion of calibration as a way to treat nonresponse is given in [Section 5.2](#).) From the example, we see that longitudinal consistency is best guaranteed by explicitly nesting RHGs, since nesting ensures corrections made at previous cycles are maintained across cycles. Notice that as more cycles of RHGs are nested, tighter control over longitudinal consistency is achieved. Unfortunately, as pointed out earlier, nesting RHGs over time becomes impractical since the realized number of nested RHGs increases over the cycles, with an ever decreasing number of sampled units within each group. Therefore for a long-term longitudinal survey a different approach may be desirable. Furthermore, a nested approach may not work well in combination with segmentation modelling since the latter tends to produce many RHGs compared with the response propensity scoring method.

From this discussion and the more detailed one in [Section 6.0](#), we notice that the NLSCY nonresponse models at Cycles 1 through 5 did not incorporate explanatory variables reflecting the RHGs of previous cycles, nor did they nest RHGs across cycles. Only variables appearing at time $t-1$ were allowed to enter into the nonresponse model. These oversights hampered the ability of these methods to model nonresponse occurring in the NLSCY and ultimately correct for nonresponse bias.

Finally, generating bootstrap weights to estimate the sampling variance requires a careful tracking of the RHGs involved in correcting for cumulative nonresponse. The crux of the matter here (see [Perspective B](#) for an overview or Section 5.2 of [Girard \(2007\)](#) for the details) is this: at Cycle 3, say, of a longitudinal survey like the NLSCY, using a *saut-de-mouton* approach we essentially use Cycle 2 nonresponse adjusted weights as the starting point for Cycle 3 weighting. But what is the starting point for bootstrapping purposes? More specifically, what should be the starting point for replicate k ? There are two choices for any given unit: the Cycle 2 nonresponse adjusted weight for the sample or the k^{th} bootstrap weight for Cycle 2, adjusted for nonresponse. And since one usually expects that, on average (over the B replicates), the latter equals the former, one may be tempted to choose the least cumbersome: the nonresponse adjusted weights for the sample. But this is a mistake that should be avoided since while the first-order properties of the bootstrap weights may very well compare to those of the weights obtained for the sample, the comparison between the two that really matters are the second-order properties.

In general, the *saut-de-mouton* approach is appealing because it appears to easily incorporate recent survey information (from $t-1$). It can work well with monotonic designs when nonresponse is modelled independently at each cycle (for example, when the same variables are used at each cycle

to create RHGs). If, however, nonresponse is not modelled independently at each cycle, then the *saut-de-mouton* can perform poorly. And if the response pattern is non-monotonic (*i.e.*, there are nonrespondents at Cycle $t-1$, and consequently missing data), then it becomes cumbersome to implement. (The NLSCY developed a method for deriving weights for converted units, but this added to the complexity of the method – see [Section 6.3](#).) Another drawback of the *saut-de-mouton* is that it is difficult to correctly calculate the bootstrap weights to reflect the effect of the cumulative nonresponse.

5.1.1.2 Model building: considerations when applying the *saut-de-l'ange*

One way to circumvent some of the difficulties and complexities with modelling nonresponse separately at each cycle, while still having some degree of longitudinal consistency of the estimates, is to model the cumulative nonresponse up to the current cycle with one nonresponse model. This is equivalent to the approach used for cross-sectional surveys except that, in the case of longitudinal surveys, there is a lot more longitudinal information with which to model nonresponse. In other words, the nonresponse adjusted weight is simply:

$$w_i(t) = w_i(0) \times \frac{1}{\hat{\phi}_i(t)} \quad (5.1.1.2-1)$$

where here $\hat{\phi}_i(t)$ is the estimated cumulative response probability since the beginning of the survey. Longitudinal consistency is achieved by making sure that the variables for the cumulative model for Cycle t are also included in the cumulative model for Cycle $t + 1$. This approach greatly simplifies weighting and consequently simplifies calculating the bootstrap weights. However, as with the other methods presented, it too has its drawbacks.

Some of these drawbacks became clear when this approach was adopted for Cycle 6 of the NLSCY, using the response propensity scoring method to create RHGs. First, only frame information is available for everyone who is being modelled. In the case of the NLSCY, the sampling frame is a rich source of information because it includes information collected from the LFS, such as household characteristics. Notice that the Cycle t nonrespondents are missing all of their Cycle t variables, so incorporating these variables into a *saut-de-l'ange* model becomes a problem of modelling in the presence of item nonresponse. One way to proceed is through imputation and then feeding these imputed variables into the cumulative nonresponse model. This was the approach taken at Cycle 6 for the NLSCY. Although time only allowed for the incorporation of some frame and Cycle 1 information into the final model, using more cycles of data is currently being investigated in order to see if this improves the performance of the nonresponse adjustment. Notice, however, that as data from later cycles are incorporated into the model, the imputation rate increases (in fact it is at least as high as the cumulative nonresponse rate at that cycle), leading to higher variance in the model estimates if variance in the presence of imputation is taken into account. And, as we pointed out in [Section 4.2](#), imputation in a longitudinal setting is complicated. Many see the use of a nonresponse model with imputed data as flawed because it relies on imputing missing data, and the more recent the cycle of data, the higher the attrition and the greater the imputation rate. So, when using more recent data in the *saut-de-l'ange* model, try to restrict it to variables with low imputation rates. Note that the *saut-de-l'ange* approach is also well suited for surveys with a rich frame, as opposed to, say, a longitudinal survey where sampling was performed through Random Digit Dialling (RDD).

Recall that our goal is to produce a set of weights that corrects for the cumulative nonresponse up to the current cycle. The *saut-de-l'ange* addresses this directly by producing just one set of RHGs over all concerned cycles. The advantages of the *saut-de-l'ange* are that it is simple to implement, even in the non-monotonic setting, and it performs well when nonresponse bias is measured using the concept of longitudinal consistency (see [Section 5.0](#)). Finally, the bootstrap weights are easily calculated (see the following section). These factors and the fact that we found it performed better at reducing bias due to attrition – at least using our measure of longitudinal consistency – contributed to the NLSCY's adoption of the *saut-de-l'ange* at Cycle 6.

5.1.2 Variance estimation considerations about weighting

Before tackling the specific issue of how to conduct the bootstrap to appropriately account for the contribution to the variance of the weighting strategy, we need to address one fundamental question first: what does it imply about the inferential framework under which we work? Does our work then fall under the model-based approach? The answer is no: we are still design-based despite our heavy use of models.

Further exploration

To read more on inferential frameworks, see [Perspective I](#).

As we have argued already, the variance introduced by the nonresponse weighting strategy can be calculated because the assumed nonresponse mechanism behaves as if it were a sampling mechanism itself, on top of the one we already have that explains why we have a sample to begin with. This two-phase setting makes the variance challenge a more tractable one. But even then there is an issue with the bootstrap – it is known to fail to capture the total variance under a two-phase approach – only one of its two components is captured by the bootstrap (see [Perspective B](#) for an overview or [Girard \(2007\)](#) for the more complete story). Fortunately, the NLSCY operates in a setting where the missing portion to the total variance is small compared to what is captured, and this is due to the small sampling fractions used in the LFS.

As we saw in section 5.1, the nonresponse adjustment for a given unit is the ratio of two weighted sums (*i.e.*, the inverse of Equation (5.1-3)): one (the numerator) sums across all sampled units in the RHG that the unit belongs to, while the other (the denominator) sums across only the respondents in the RHG. With the bootstrap we need to compute these two weighted sums, using the *bootstrap* weights, one replicate after the other. So, while the main sample led to one adjustment, the 1,000 bootstrap replicates of the NLSCY lead to the computation of 1,000 adjustments.

A shortcut used in all cycles of the NLSCY prior to Cycle 6 was to replace the replicate-dependent weighted sum of bootstrap weights in the numerator by one corresponding sum using the sample weights. This shortcut circumvented the computation of 1,000 numerators, arguing possibly that the average of these 1,000 replicate-based numerators matched the sample-based numerator. This is not disputed. But that property of the replicate-based numerators of the nonresponse adjustments under the bootstrap is not relevant to the issue at hand: this is a first-order moment's property (the expected value) when what we really need is a correspondence based on second-order moments.

And that correspondence does not exist between the shortcut and the appropriate methodology, hence its inadequacy.

The impact of the use of this shortcut is not uniform for all cycles. Indeed, it led to variance estimates at Cycle 1 which over-estimate the true variance; to low variance estimates in some cases at Cycle 2; and to no visible effects at Cycles 3 to 5. For Cycle 2, in the Canadian Atlantic provinces, the RHG construction was driven by marital status as reported by the respondent at Cycle 1. Consequently, the constant-over-replicates' numerator that the shortcut used in the bootstrap nonresponse adjustments did not reflect the fact that this numerator was just an estimate, not a known value. Had there not been any post-stratification at Cycle 2 to blur the picture a bit, an analyst would have claimed a variance estimate of zero at Cycle 2 for the estimate of the population counts by marital status in that region, using the reported Cycle 1 data. (Indeed, all bootstrap estimates would have turned out to be precisely that one constant-over-replicates' numerator.) This is incorrect since the RHGs are domains, not post-strata. It had been suspected for some time that the Cycle 1 variance estimates were too large, but it was unclear what the problem was until the effect of the shortcut was clearly understood. After all, the shortcut is only known to decrease the variance if the domain of interest coincides with the RHG of the units of the domain. For Cycle 1, a cross-verification (described in [Perspective B](#)) was performed at each step of the derivation of the final bootstrap weights, and it was at this point that the shortcut was found to be the reason behind the overly high variance estimates. The same cross-verification was performed on the final bootstrap weights at Cycles 3 to 5, which were also subject to the shortcut, and the histograms appear normal.

As it turns out, the nonresponse methodology at Cycle 1 has one thing that sets it apart from that used at Cycles 3 and 5, which explains why the histogram for Cycle 1 differs so much from the others: the large number of RHGs. By creating a large number of RHGs at Cycle 1 – well over 400 of them – many of these turned out to be quite small in size. Consequently, for such small RHGs, the replicate-based numerators fluctuate greatly about their average, introducing variance that should not be there. It is only if the RHGs are large (as in Cycles 3 and 5) that the shortcut's effect on the variance is benign since in such cases the replicate-based numerators vary very little about their average, which the shortcut uses in their stead.

5.2 Calibration

One form of calibration used in a longitudinal setting is about matching the current cycle estimates of a kept-fixed-over-time characteristic to earlier estimates of that characteristic. The use of calibration in this manner seems to have first appeared in writing in [Singh et al. \(1995\)](#).

In general terms, calibration is about replacing a set of initial weights $\{w_k\}$ by another set of calibrated weights $\{w_{k,cal}\}$ such that

$$\sum_{k \in I} w_{k,cal} = T_I \quad (5.2-1)$$

for various sets of units I and in such a way that $\{w_{k,cal}\}$ is as close as possible to the original set $\{w_k\}$. This latter condition is spelled out using some distance function d_k between the weights $w_{k,cal}$ and w_k .

Further exploration

To read more on the choice of the distance, see [Perspective J](#).

In traditional calibration, the values T_i are known totals, usually from some external source of data (e.g., projections from census data). In NLSCY's longitudinal setting, the T_i are estimates from an earlier cycle for characteristics that are kept fixed over time. For example, one such value could be the Cycle 1 estimate, using both the Cycle 1 weight and the Cycle 1 reported value for the following marital status question: does the PMK live alone or not? At Cycle 2 we could then seek to calibrate the Cycle 2 weights so that by using the Cycle 2 set of respondents, but with the Cycle 1 reported marital status, we achieve the same estimate as before. Calibration performed this way in a longitudinal setting is a clear attempt to improve on the longitudinal consistency of the estimates, a concept explained in [Section 6.5](#).

Calibration, just like nonresponse weighting, replaces one set of weights with a revised set. The main difference is that calibration does not seek to explain how nonresponse occurred or to describe in any way the (hypothetical) nonresponse forces at work. Calibration, in a longitudinal setting, simply "observes and records" the effects that nonresponse has had on previously released (presumably, key) estimates and works to restore these estimates for the current cycle. More precisely, calibration takes the observed status of a given characteristic at a given previous cycle and computes a new weighted total using the current cycle's weights. Then, it finds the minimal way to tweak the current cycle's set of weights of the responding units so that when used in conjunction with the previously reported values for the characteristic it yields numerically the same result. If this has worked well, then (conceptually, at least) with the most current set of weights the analyst can reproduce any previous cycle's estimate. In other words, while an estimate originally released at Cycle t used Cycle t reported data and weights, with the current cycle's calibrated weights the analyst can reproduce that estimate by using the Cycle t reported data for the subset of units which responded at the current cycle.

Attractive as it is, calibration is accompanied by a lot of nontrivial practical issues that need to be addressed before we can even think of implementing it. These are briefly discussed in the next subsection.

5.2.1 Calibration in practice: the weighting steps

In [Tam et al.\(2007\)](#) the authors describe an attempt at calibration that was investigated for the NLSCY. The idea was to use software to perform the calibration based on variables that were identified as relevant characteristics. The choice of these variables and the restrictions that needed to be imposed on them in practice to ensure convergence to a solution were the main issues. For instance, in a multi-purpose survey like the NLSCY, it is difficult to identify key variables which should be used in the calibration. Indeed, finding variables is not in itself an issue since the NLSCY offers a myriad of variables to choose from; the issue is which variables are worth calibrating to, with the goal of improving the estimates for as many (un-calibrated) variables as possible.

5.2.2 Calibration in practice: the variance estimation process

To conduct variance estimation for calibration under the bootstrap is not, in principle, difficult. The basic idea is the following: whenever Cycle $t+x$ weights are calibrated to yield the Cycle t estimate for a given characteristic, Y , (both estimates using Cycle t values for Y), the k^{th} set of bootstrap weights for Cycle $t+x$ has to be calibrated to the k^{th} bootstrap estimate for Y , for Cycle t . Therefore, one does not calibrate each of the bootstrap weights to yield the Cycle t main estimate for Y , but to their respective bootstrap counterparts.

The reader will observe that for a calibrated variable, Y , the Cycle $t+x$ variance estimate will match that of Cycle t , even though the set of respondents has shrunk. Indeed, the bootstrap estimates for both cycles are point-wise the same (after all, we worked to force each of the Cycle $t+x$ bootstrap estimates to match its Cycle t counterpart). And this non-increasing variance estimate for the Cycle t estimate of Y using Cycle $t+x$ weights is legitimate. Indeed, we are entitled to the same variance estimate because we have essentially encrypted all of the information of Cycle t about Y into all of the weights so that, despite fewer respondents, the information is preserved fully from Cycle t to Cycle $t+x$. For variables that did not enter the calibration, the picture is not as rosy. Indeed, if the calibration does not bring any clear gains in precision to the un-calibrated estimates, then chances are the effect of calibrating on the variance will be detrimental because of the increased heterogeneity in the calibrated weights. (Calibrated weights tend to be unique: each one differing from the others since calibration is a micro adjustment unlike nonresponse weighting where all units belong to the same RHG share the same nonresponse adjustment.)

A big issue in practice is the fact already mentioned in *Perspective B* that a bootstrap replicate contains only about 60% of the sampled units. Plainly put, there are 40% fewer weights for the calibration to tweak in a given bootstrap replicate than there were with the main sample. Consequently, a set of calibration variables which led to a convergent methodology for the main sample may very well not lead to a convergent one under the bootstrap. Therefore, in choosing the proper variables (and aggregates of their values) for calibration, one must think of the important restrictions the bootstrap imposes on it all.

A preliminary study which adapted the variables identified in [Tam et al.\(2007\)](#) to the bootstrap found that the gains in efficiency were not obvious for un-calibrated variables. (Actually, in the paper the selection of variables to be used in the calibration was not made with any consideration given to the bootstrap issue just presented, but only with the main sample in mind. Consequently, to perform the bootstrap it was necessary to revisit the aggregation originally made of the variables to allow for convergence.) It was found that while the variance estimate at Cycle 6 for a calibrated variable is about the same as that of Cycle 1, this gain in precision does not appear to extend to un-calibrated variables. Indeed, for the un-calibrated variables, the variance at Cycle 6 compares to the variance obtained with the logistic model. In other words, while variance estimates do convey the retaining information exercise that was performed for all calibrated variables, it appears that for variables not explicitly calibrated to, no real gain in variance was achieved. Since it is impossible to pick beforehand a set of calibration variables in the NLSCY suitable for all possible analyses, and based on the lukewarm findings regarding the variance performance of calibrated weights, the decision was made not to implement a pure calibration methodology in the NLSCY.

6.0 The nonresponse weighting methodologies of Cycles 1 through 7

This section contains the nonresponse weighting details of Cycles 1 through 7. We describe the methodology that was implemented at the time, and summarize the advantages and disadvantages of the approach. These details are useful to know because they explain why in practice (based on NLSCY) one approach is preferable to another.

As explained in [Section 5.0](#), the NLSCY uses the RHG method at each cycle to produce a set of longitudinal weights that corrects for the cumulative total nonresponse. The longitudinal consistency of the NLSCY estimates was assessed after Cycle 5 ([Statistics Canada \(2007\)](#)) and the conclusion was that some estimates might be biased. This finding prompted a review of the NLSCY nonresponse weighting methodologies used from Cycles 1 through 5 with a view to developing a new Cycle 6 methodology that would better minimize nonresponse bias.

The structure of this section loosely follows the chronology of events surrounding the development of the Cycle 6 weighting strategy. Once possible nonresponse bias was detected, we re-examined the methodologies of Cycles 1 through 5, identified their shortcomings (which were largely those of the *saut-de-mouton* approach), and documented other considerations when developing a longitudinal nonresponse model ([Section 6.1](#), [Section 6.2](#) and [Section 6.3](#)). We then developed the Cycle 6 methodology, which uses response propensity scores estimated from logistic regression model ([Section 6.4](#)), and evaluated it using a measure of longitudinal consistency which was also used to compare it with the previous approaches ([Section 6.5](#)).

To provide a complete description, some notation is required. Let the original sample selected be denoted by S . Each selected unit i has a design weight, $w_i(0)$, which is the inverse of the unit's probability of selection. In the case of the NLSCY, this probability of selection has been adjusted for LFS nonresponse. At Cycle t , the set of respondents is $R(t)$ and nonrespondents is $NR(t)$. Moreover, unit i belongs to the RHG denoted by $RHG_i(t)$, has an estimated response probability of $\hat{\phi}_i(t)$, and has a nonresponse adjusted weight of $w_i(t)$.

In the subsections that follow, we describe how $w_i(t)$ is calculated at each Cycle t . However, note that prior to their release the NLSCY weights are post-stratified to ensure that demographic projections based on the Canadian Census are respected. Province of residence, age, and gender define the post-strata. The post-stratified weight is:

$$w_{i, \text{post-stratified}}(t) = w_i(t) \times \frac{N_P}{\sum_{k \in P} w_k(t)} \text{ for } 1 \leq t \leq 6 \text{ and } i \in P \quad (6.0-1)$$

where N_P is the population count for the post-stratum P .

6.1 Cycles 1 and 2

The *saut-de-mouton* approach was employed at Cycles 1 and 2 and is easy to apply because of the monotonic nonresponse design at both cycles: everyone surveyed at Cycle t was a respondent at all previous cycles and thus has data for Cycle $t-1$, and $w_i(t-1) > 0$. With these two pieces of information, Cycle t nonresponse was modelled using the segmentation method. The Cycle t weights were then calculated using the following formulae:

$$\hat{\phi}_i(t) = \frac{\sum_{k \in R(t) \cap RHG_i(t)} w_k(t-1)}{\sum_{k \in RHG_i(t)} w_k(t-1)} \text{ for } t \in \{1, 2\} \quad (6.1-1)$$

and

$$w_i(t) = w_i(t-1) \times \frac{1}{\hat{\phi}_i(t)} \text{ for } t \in \{1, 2\}$$

$$w_i(t) = w_i(0) \times \frac{1}{\hat{\phi}_i(1)} \times \dots \times \frac{1}{\hat{\phi}_i(t-1)} \times \frac{1}{\hat{\phi}_i(t)} \quad (6.1-2)$$

Notice that $\hat{\phi}_i(t)$ is simply the weighted response rate of $RHG_i(t)$. A schematic representation of the Cycle 1 and 2 methodology is shown in Figure 6.1-1. The main drawback of this approach is that the Cycle t nonresponse model and RHGs were built without considering previous cycle RHGs, and only variables from the previous Cycle, $t-1$, were considered during model building, which may have overlooked other important explanatory variables ([Section 5.1.1.1](#)). Note also that segmentation modelling produced a large number of RHGs at Cycle 1, but far fewer at Cycle 2 (as shown in Figure 6.1-1).

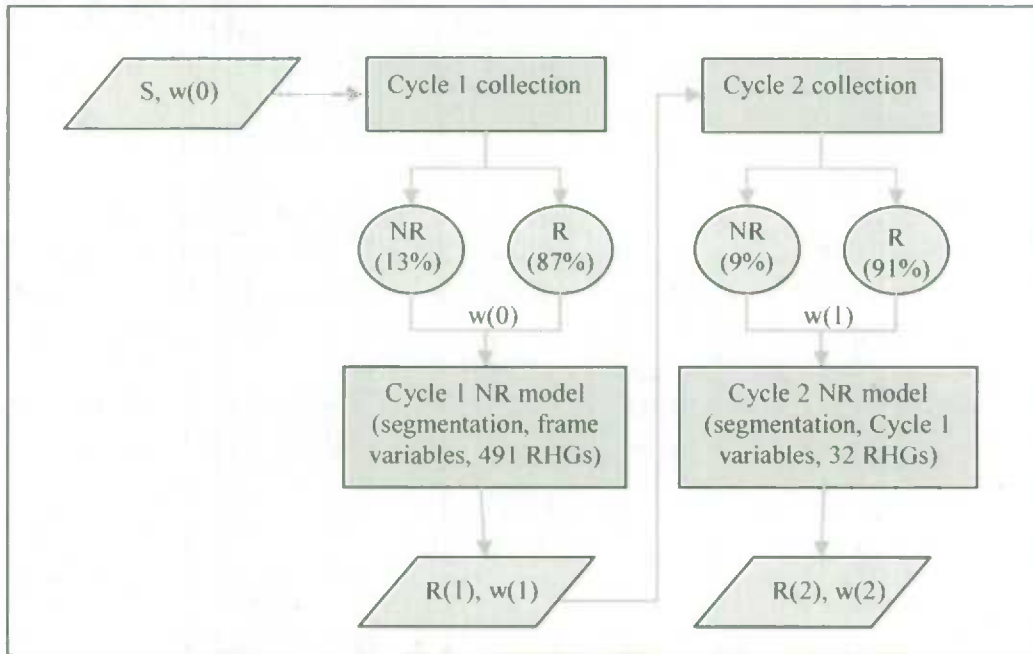


Figure 6.1-1: Schematic description of the Cycle 1 and 2 nonresponse methodology (classic *saut-de-mouton*)

6.2 Cycle 3

The nonresponse weighting methodology for the Original Cohort changed at Cycle 3 when a non-monotonic design was adopted. For the Original Cohort, any unit that was a nonrespondent at Cycle 1 was never again sent out for collection. Thus at Cycle 3, five patterns of nonresponse were possible: $\{r r r, r r nr, r nr nr, r nr r, nr nr nr\}$.

Units that responded at Cycle 3 but not at Cycle 2 (*i.e.*, $\{r nr r\}$) are called converted units. These converted units are missing two pieces of information required to implement the *saut-de-mouton* approach of Cycle 2: Cycle 2 data and Cycle 2 nonresponse adjusted weights (*i.e.*, $w_i(2) = 0$ for these units).

However, all respondents at Cycle 3 were also respondents at Cycle 1, and consequently had Cycle 1 data and Cycle 1 nonresponse adjusted weights, $w_i(1) > 0$, so a hybrid *saut-de-mouton* and *saut-de-l'ange* approach was implemented. The Cycle 1 *saut-de-mouton* model was used to account for Cycle 1 nonresponse and a *saut-de-l'ange* model, using Cycle 1 variables and segmentation modelling, was used to construct RHGs that would account for all of the cumulative nonresponse occurring after Cycle 1. The weighting formulae are:

$$\hat{\phi}_i(t) = \frac{\sum_{k \in R(t) \cap RHG_i(t)} w_k(1)}{\sum_{k \in RHG_i(t)} w_k(1)} \text{ for } t = 3 \quad (6.2-1)$$

and

$$\begin{aligned} w_i(t) &= w_i(1) \times \frac{1}{\hat{\phi}_i(t)} \\ w_i(t) &= w_i(0) \times \frac{1}{\hat{\phi}_i(1)} \times \frac{1}{\hat{\phi}_i(t)} \end{aligned} \text{ for } t = 3 \quad (6.2-2)$$

Figure 6.2-1 is a schematic of the Cycle 3 approach. This methodology suffers the same drawbacks as the Cycle 2 *saut-de-mouton* approach, namely, the Cycle t nonresponse model and RHGs were built without considering previous cycle RHGs and the search for explanatory variables was restricted to only one cycle ([Section 5.1.1.1](#)): two nonresponse adjustments were multiplied together without accounting for any dependence between them.

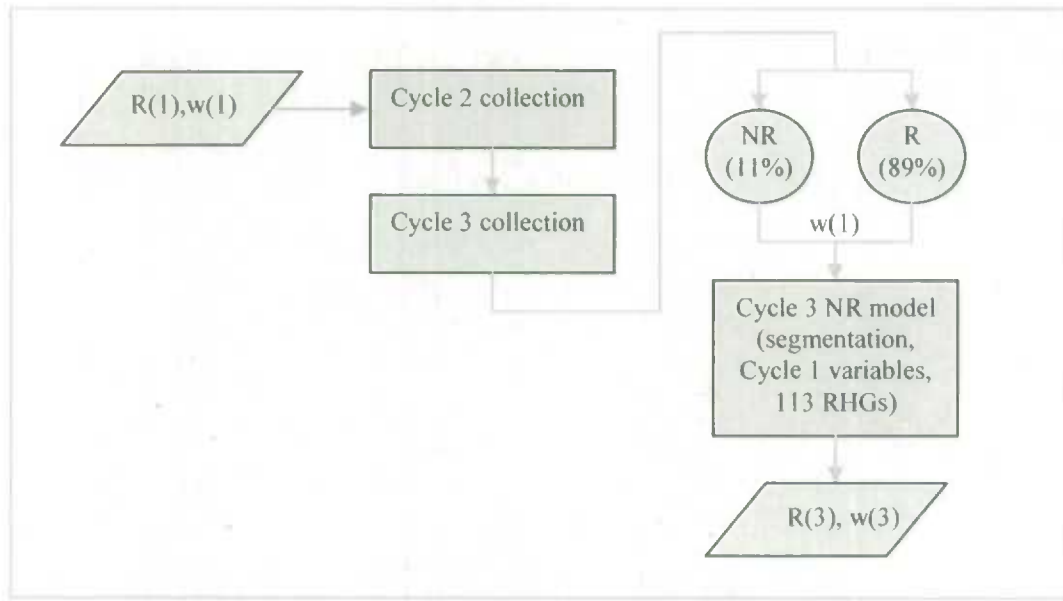


Figure 6.2-1: Schematic description of the Cycle 3 nonresponse methodology (hybrid *saut-de-mouton* and *saut-de-l'ange*).

6.3 Cycles 4 and 5

At Cycle 4, the nonresponse weighting methodology was changed once more. Cycles 4 and 5 also had non-monotonic response patterns and converted units. To continue with the Cycle 3 approach would have meant forever being restricting to Cycle 1 variables for nonresponse modelling. At the time, this constraint was believed to be a major drawback since it was felt that data from more recent cycles could help model more recent nonresponse. So, instead, at Cycle 4 the *saut-de-mouton* approach was adapted to accommodate converted units.

Recall that the goal of the *saut-de-mouton* approach is to correct for Cycle t nonresponse. This requires that respondents at time t have nonresponse adjusted $t-1$ weights, a requirement that is not met by the converted units, $NR(t-1) \cap R(t)$. So, at Cycle 4, nonresponse adjusted $t-1$ weights, $w'_i(t-1) > 0$, were created for these converted units.

The new weights were calculated by taking a unit's design weight, $w_i(0)$, and multiplying it by a factor that accounts for all nonresponse up until time $t-1$. This factor is equal to the sum of the final nonresponse adjusted $t-1$ weights divided by the sum of all the design weights, for all respondents at time t . For the converted units:

$$w'_i(t-1) = w_i(0) \times \frac{\sum_{k \in R(t-1)} w_k(t-1)}{\left(\sum_{k \in R(t-1)} w_k(0) + \sum_{k \in NR(t-1) \cap R(t)} w_k(0) \right)} \quad (6.3-1)$$

for $i \in NR(t-1) \cap R(t)$ and $t \in \{4, 5\}$

Note that since the newly created weights, $w_i'(t-1) > 0$, must respect the same total as the original Cycle $t-1$ weights, new $t-1$ weights must be created for those who really were respondents at time $t-1$. Thus, to create $w_i'(t-1)$ for units in $R(t-1)$, we must ensure that

$$\sum_{i \in R(t-1) \cup (NR(t-1) \cap R(t))} w_i'(t-1) = \sum_{i \in R(t-1)} w_i(t-1) \quad (6.3-2)$$

This condition is met by defining the new weights of $R(t-1)$ by:

$$w_i'(t-1) = w_i(t-1) \times \frac{\sum_{k \in R(t-1)} w_k(0)}{\left(\sum_{k \in R(t-1)} w_k(0) + \sum_{k \in NR(t-1) \cap R(t)} w_k(0) \right)} \quad (6.3-3)$$

for $i \in R(t-1)$ and $t \in \{4,5\}$

In other words, those who were respondents at $t-1$ must have their original $t-1$ nonresponse adjusted weights deflated. This deflation factor is calculated as the ratio of design weights.

Now that all respondents at time t have nonzero nonresponse adjusted $t-1$ weights, the *saut-de-mouton* modelling of Cycle t nonresponse can begin. Since converted units are, by definition, respondents at time t , they do not require a nonresponse adjustment at time t . Thus,

$$w_i(t) = w_i'(t-1) \quad (6.3-4)$$

for $i \in NR(t-1) \cap R(t)$ and $t \in \{4,5\}$

For respondents at time $t-1$ who were respondents or nonrespondents at time t , a *saut-de-mouton* approach was used to model nonresponse at time t using segmentation modelling. The weighting formulae are as before, this time using the new $t-1$ weights, namely:

$$\hat{\phi}_i(t) = \frac{\sum_{k \in R(t) \cap RHG_i(t)} w_k'(t-1)}{\sum_{k \in RHG_i(t)} w_k'(t-1)} \quad (6.3-5)$$

for $i \in R(t-1)$ and $t \in \{4,5\}$ and where the nonresponse adjusted weight for the i^{th} respondent is,

$$w_i(t) = w_i'(t-1) \times \frac{1}{\hat{\phi}_i(t)} \text{ for } i \in R(t-1) \text{ and } t \in \{4,5\} \quad (6.3-6)$$

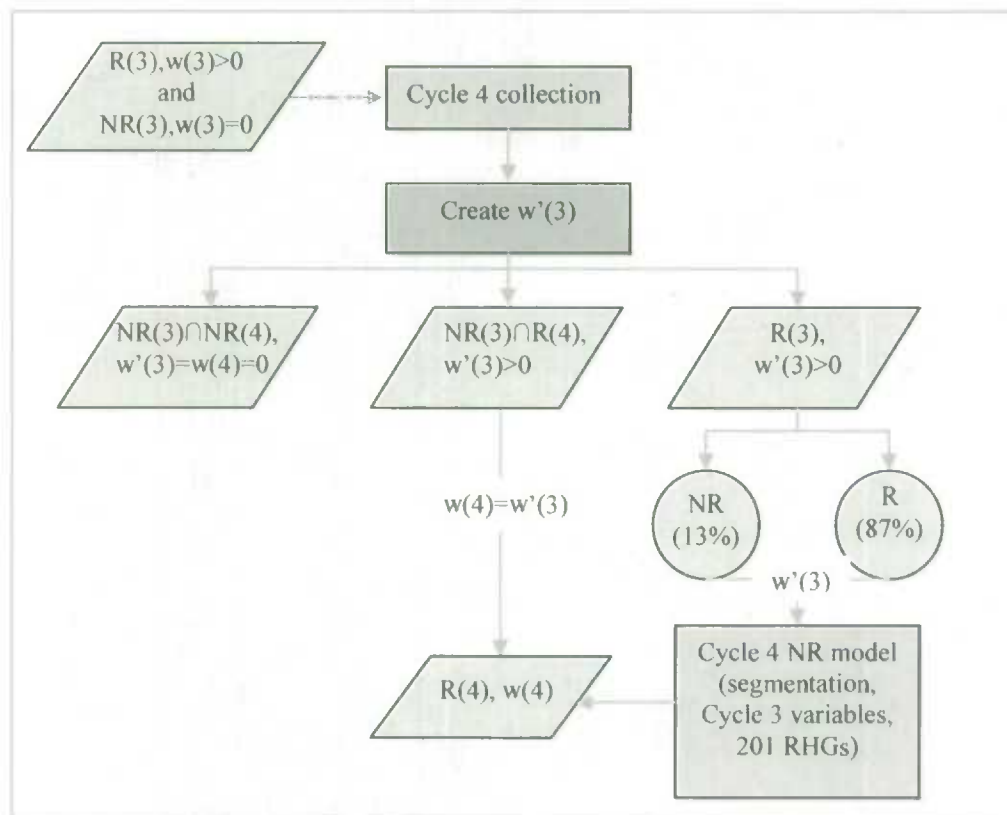


Figure 6.3-1: Schematic description of the Cycle 4 nonresponse methodology (*saut-de-mouton* adapted to handle converted units)

A schematic of the approach used at Cycle 4 is given in Figure 6.3-1. An analogous diagram exists for Cycle 5 with the NR group making up 12% of R(4) and the Cycle 5 nonresponse model specifications: segmentation, Cycle 4 variables, and 180 RHGs. As with the earlier approaches, the Cycle t nonresponse model and RHGs were built without any consideration of previous cycle RHGs and the search for explanatory variables was restricted only to the previous cycle (Section 5.1.1.1). Moreover, there were problems with how the converted units were handled (as described in the following section). For these and other reasons cited in the following, we decided to adopt a new approach at Cycle 6.

6.3.1 Handling converted units in a *saut-de-mouton* framework

Notice that nowhere in the formulae for the new $t-1$ weights (Equations (6.3-1) and (6.3-3)) is $RHG_{i,t-1}$ mentioned. That is, we do not respect the Cycle $t-1$ RHGs. Applying these formulae to the simple example in Table 6.3.1-1 yields the $w_i'(t-1)$ shown in the table. Notice how these weights differ from the design weights. Applying this strategy over time would lead us further and further away from the original design, effectively departing from a probabilistic framework. This phenomenon could also contribute to longitudinally inconsistent estimates.

On the other hand, if we were to respect the Cycle $t-1$ RHGs, then we would re-write (6.3-1) as

$$w'_i(t-1) = w_i(0) \times \frac{\sum_{k \in R(t-1) \cap RHG_i(t-1)} w_k(t-1)}{\left(\sum_{k \in R(t-1) \cap RHG_i(t-1)} w_k(0) + \sum_{k \in NR(t-1) \cap R(t) \cap RHG_i(t-1)} w_k(0) \right)} \quad (6.3.1-1)$$

for $i \in NR(t-1) \cap R(t)$ and $t \in \{4,5\}$ and consequently to preserve (6.3-2), we also need to rewrite (6.3-3) as:

$$w'_i(t-1) = w_i(t-1) \times \frac{\sum_{k \in R(t-1) \cap RHG_i(t-1)} w_k(0)}{\left(\sum_{k \in R(t-1) \cap RHG_i(t-1)} w_k(0) + \sum_{k \in NR(t-1) \cap R(t) \cap RHG_i(t-1)} w_k(0) \right)} \quad (6.3.1-2)$$

for $i \in R(t-1)$ and $t \in \{4,5\}$

Suppose unit i did not respond at Cycle $t-1$, but did respond at Cycle t . At Cycle $t-1$, the weight of unit i is distributed only to those respondents at time $t-1$ who were in the same RHG, (i.e., $RHG_i(t-1)$). At Cycle t , we assign unit i a value of $w'_i(t-1)$. In effect, we are pretending he was a respondent at time $t-1$, in which case he should take back the share of his weight from those – and only those – who received it (i.e., those in $RHG_i(t-1)$). These weights are also shown in Table 6.3.1-1.

Unit i	$w_i(0)$	Cycle $t-1$ response status	Cycle $t-1$ RHGs	$w_i(t-1)$	Cycle t response status	$w'_i(t-1)$ (Not respecting Cycle $t-1$ RHGs)	$w'_i(t-1)$ (Respecting Cycle $t-1$ RHGs)
1	2	R	1	6	R	$\frac{6 \times (2+6)}{(2+4+6)} = 4$	$6 \times \frac{2}{6} = 2$
2	4	NR	1	0	R	$\frac{4 \times (16+6)}{(2+4+6)} = 7.3$	$4 \times \frac{6}{(4+2)} = 4$
3	6	R	2	16	R	$\frac{16 \times (2+6)}{(2+4+6)} = 10.7$	$16 \times \frac{6}{6} = 16$
4	10	NR	2	0	NR	0	0
Total	22			22		22	22

Table 6.3.1-1: Example of how to handle converted units in a *saut-de-mouton* approach, illustrating the importance of respecting RHGs at $t-1$.

Obviously this requires good documentation of the RHGs used at each cycle so that they can be retrieved at subsequent cycles.

This discussion shows further flaws with the Cycle 4 and 5 weighting approach (a *saut-de-mouton* approach adapted to handle converted units). Even when correctly done, handling converted units with a *saut-de-mouton* methodology is complicated. By contrast, the *saut-de-l'ange* does not require a special process for converted units: they are handled in the same way as all other units.

6.4 Cycles 6 and 7

Here, we describe the weighting methodology developed for Cycle 6, which is the current nonresponse weighting strategy used by the NLSCY. At Cycle 7, improvements upon the Cycle 6 approach were explored (namely, the use of calibration), but none was found due largely to our constraint of producing one-size-fits-all survey weights ([Tam \(2008\)](#)). The NLSCY methodologists continue to explore ways to further improve the nonresponse weighting strategy developed at Cycle 6.

6.4.1 Segmentation modelling versus logistic modelling

Recall from [Section 5.1](#) that segmentation modelling can yield a large number of RHGs because of the branching nature of the chi-square trees used to build the groups. Since each branch is isolated from the others, this can result in distinct RHGs with the same response rate (and consequently the same weight adjustment). This is what we found for the NLSCY (Cycles 1 to 5): segmentation modelling yielded a large number of RHGs relative to the overall sample size, with often small differences between the response rates of different RHGs.

The response propensity scoring method, by contrast, offers greater control and flexibility over the number and character of the RHGs produced. (Recall from [Section 5.1](#) that response propensities are estimated using a logistic regression model and then similar response propensities are grouped together. This prevents the proliferation of RHGs with similar response rates.) Recent studies at Statistics Canada have compared segmentation modelling to the response propensity scoring method when treating nonresponse (for example, [Haziza and Beaumont \(2007\)](#)). These studies found the response propensity scoring method to be more robust for dealing with nonresponse than the segmentation modelling approach. For these reasons, we decided to abandon segmentation modelling and switch to the response propensity scoring method at Cycle 6.

6.4.2 Using response propensity scores at Cycles 6 and 7

It was at the start of our production period for Cycle 6 that we realized that changes to the nonresponse weighting methodology were in order. Therefore, we were only able to implement a subset of the changes that we had in mind. The first change was to adopt a *saut-de-l'ange* approach (which works well with both monotonic and non-monotonic designs), modelling the cumulative total nonresponse up to and including Cycle 6 in one step. This enabled us to retain the RHG approach but prevented us from having to deal with previous cycle RHGs and devising a special methodology to handle converted units. The second major change was switching from segmentation modelling to the response propensity scoring method.

As described in [Section 5.1](#), the response propensity scoring method uses a logistic regression model to estimate the response probability of each sampled unit. The RHGs are created by sorting

the estimated probabilities and using the deciles to delineate 10 groups. The number of groups is adjusted downwards as necessary, so that a monotonic response rate pattern across groups is achieved. Within each RHG, the response rate is calculated and used to construct the nonresponse adjustment. This methodology gives one nonresponse adjustment at Cycle t :

$$w_i(t) = w_i(0) \times \frac{1}{\hat{\phi}_i(t)} \text{ for } t = 6 \quad (6.4.2-1)$$

where

$$\hat{\phi}_i(t) = \frac{\sum_{k \in R(t) \cap RHG_i(t)} w_k(0)}{\sum_{k \in RHG_i(t)} w_k(0)} \text{ for } t=6 \quad (6.4.2-2)$$

Figure 6.4.2-1, below, illustrates the Cycle 6 nonresponse methodology.

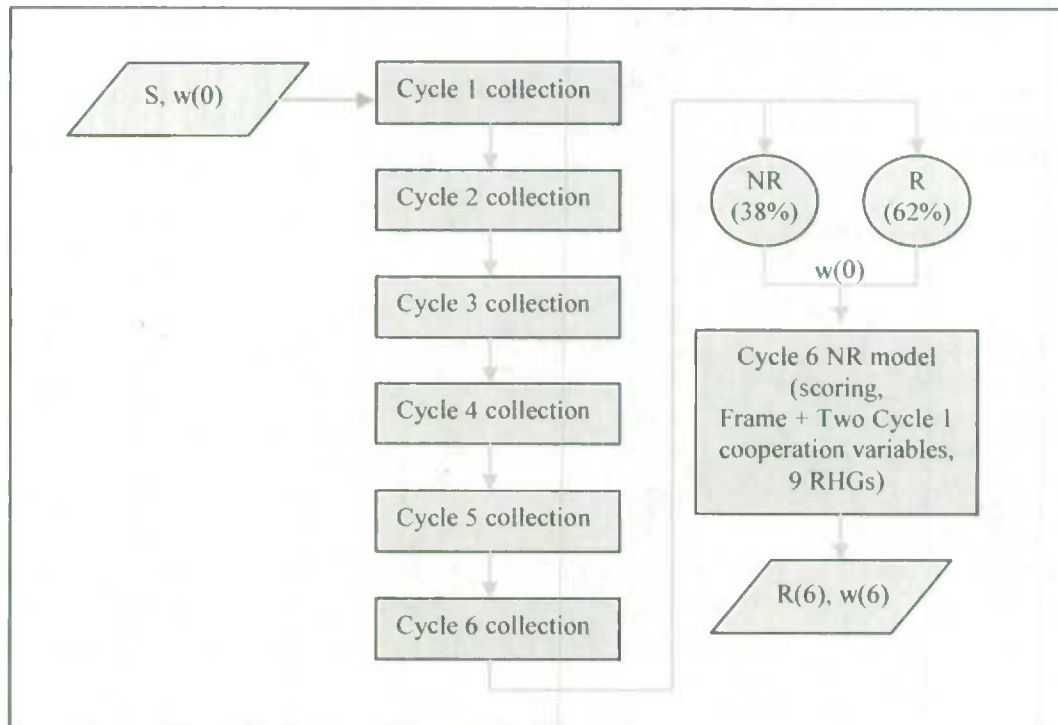


Figure 6.4.2-1: The classic *saut-de-l'ange* approach applied at Cycle 6.

To choose the variables for modelling, we created an automated program that runs chi-square tests on any set of categorical variables. Some variables we had at our disposal for modelling came from the sampling frame and six cycles of NLSCY data. We submitted 70 frame variables into the chi-square program which identified 20 key variables that were entered into the logistic regression model. When it came to the NLSCY variables, there was a confounding of item and total nonresponse that interfered with the performance of the chi-square test; essentially the 'missing value' category dominated the chi-square test. However, since we had Cycle 1 information for 86% of our sample, we considered only the Cycle 1 respondents and ran the automated chi-square program to identify key Cycle 1 variables related to nonresponse. Our goal was then to impute

these variables for the Cycle 1 nonrespondents. Because of time constraints we were only able to use two imputed variables in the final logistic model: two Cycle 1 cooperation scores which are discussed at length in the next section. The NLSCY logistic regression nonresponse model was built by a stepwise selection of variables from among the group of key variables. Although the model is unweighted, we did include the province, a key stratification variable. The Hosmer-Lemeshow test was used to assess model fit.

The final Cycle 6 nonresponse model accounts for the nonresponse from the start of the survey and contains 10 LFS variables, two cooperation variables and a total of nine RHGs. The LFS variables are: province of residence, respondent's highest level of education ever completed, total number of employed persons in the economic family, type of dwelling, class of worker (main job), number of children aged 0 to 6 years old, highest grade of elementary or high school completed, total overtime hours worked in reference week (paid and unpaid), spouse's education group and LFS rotation group.

We adopted the same approach for Cycle 7, where we investigated the use of a hybrid logistic calibration model, which performed well in terms of reducing overall bias but was ultimately rejected for two reasons: NLSCY has over 1,000 variables and we could not find one set of calibration variables that would be suitable for all possible analyses, and, as mentioned earlier, for variables that were not calibrated to, the variance estimates were similar to those obtained using the considerably simpler logistic model (so, there were no efficiency gains).

6.4.3 The cooperation variables

The idea behind a cooperation variable is to combine data that indicate how enthusiastically a respondent has participated in the survey, and then to categorize the person according to his or her level of cooperation. For example, if an individual ever refused to participate, but was finally persuaded to do so (*i.e.*, was a converted refusal), then this person is likely to refuse again in the future. Similarly, if a respondent rescheduled the interview many times before finally agreeing to carry out the interview, this could indicate a lack of interest and future nonresponse.

The form of the cooperation variables used at Cycle 6 simply categorizes units according to the percentage of questions that they did not answer, of all the questions that were posed. Specifically, a cooperation score CS_i was calculated for each Cycle 1 respondent as an item nonresponse rate using the following equation:

$$CS_i = \frac{m_i}{n_i}, m_i \leq n_i \quad (6.4.3-1)$$

where n_i is the number of questions that were asked to respondent i and m_i is the number of questions that were not answered.

The scores were then categorized by decile: the first decile 0-10% indicates individuals who were the most cooperative in the first cycle; the last decile 90-100% indicates those who were the least cooperative.

We were able to derive two cooperation variables from the Cycle 1 data. The Cycle 1 questionnaire has several components: household-level questions answered by the PMK (e.g., household revenue), child-level questions answered by the PMK, child-level questions answered by the child, questions answered by the child's teacher and principal, and items completed by the interviewer. From this information, the following two cooperation scores were derived:

- i. A cooperation score calculated that excluded both the school component and questions completed by the interviewer:

The respondents are the PMK and the child. When the PMK is reporting for more than one child, the household-level questions are only counted once and their item nonresponse is added to the item nonresponse for each child (for the child-specific questions answered by either the PMK or child).

- ii. A cooperation score including the school component and questions completed by the interviewer:

The cooperation score for Cycle 1 respondents was calculated as above, but also includes the item nonresponse for the teacher, principal and interviewer. Figure 6.4.3-1 compares this Cycle 1 cooperation score with cumulative, unit response rates at Cycles 2 through 6. The graph illustrates that respondents who were less cooperative at Cycle 1 have higher subsequent unit nonresponse rates than those who were more cooperative at Cycle 1.

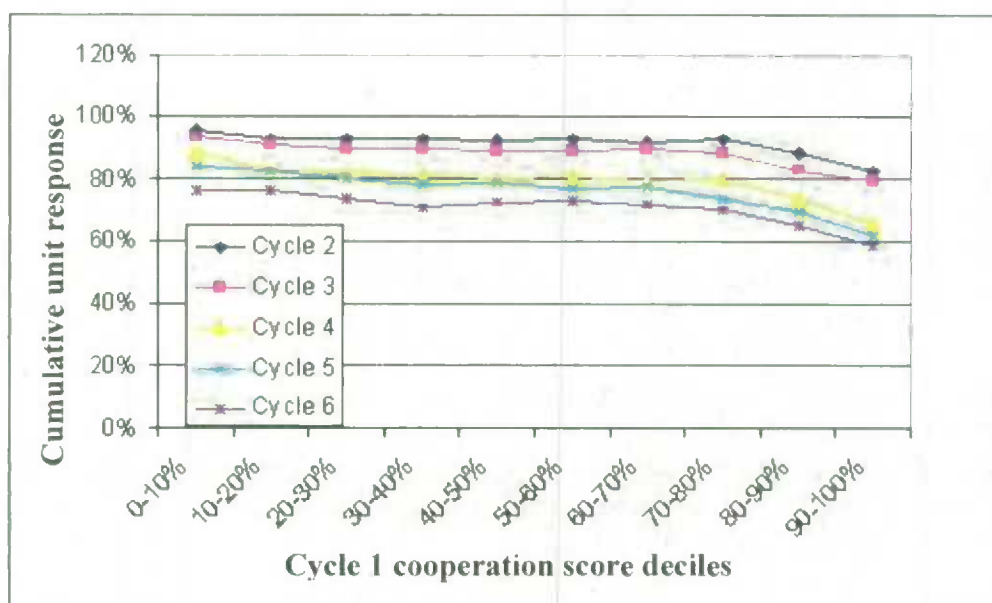


Figure 6.4.3-1: Cumulative, unit response rates across Cycles 2 to 6 for different Cycle 1 cooperation scores (the score includes the school component and questions completed by the interviewer).

Both variables were found to be significant in the Cycle 6 nonresponse model. From this we infer that in some cases it is the PMK's reaction to (cooperation with) the survey that drives their future nonresponse. In other cases the PMK could be somewhat cooperative but poor cooperation from the principal and teacher may influence the PMK's attitude towards the survey at subsequent cycles.

In addition to being used in the nonresponse model, a cooperation variable was also tested at Cycle 7 collection to flag “at risk” respondents, in order to prevent nonresponse. Future plans for the cooperation variable include further development to incorporate more paradata (e.g., converted refusals, rescheduled interviews, *etc.*) which should provide even better indications of a person's cooperativeness.

Other longitudinal surveys have also found item nonresponse at one cycle to be a useful predictor of total nonresponse at a subsequent cycle ([Burkam and Lee \(1998\)](#); [Loosvelt *et al.* \(2002\)](#); [Watson *et al.* \(2003\)](#); [Watson *et al.* \(2006\)](#)).

6.5 Comparing the nonresponse weighting methodologies of Cycles 1 to 6 using longitudinal consistency

Now that we have thoroughly described the nonresponse weighting methodologies of Cycles 1 through 6, we are ready to compare them via an empirical study. In particular, we compare the longitudinal consistency of each method, with respect to the Cycle 1 NLSCY data. Recall that with a longitudinal cohort, attrition erodes the sample over time. If we are doing an adequate job of adjusting for nonresponse at each cycle, then the longitudinal nonresponse adjusted weights at each cycle should produce consistent estimates for static characteristics. This approach was developed in [Statistics Canada \(2007\)](#) where it was used to identify nonresponse bias in some estimates, and is generalized here to compare nonresponse weighting methodologies. To evaluate a weighting methodology, we take the post-stratified weights the method generates at Cycle t and compare the relative absolute differences between the actual Cycle 1 estimates and the Cycle 1 estimates computed using these Cycle t weights. The exercise is carried out for Cycles t from 2 to 6.

To help us evaluate the weighting methodologies, we establish two baselines. The first is a uniform nonresponse adjustment, meaning that the nonresponse mechanism is assumed to be the same throughout the sample. So, the nonresponse adjustment is simply the inverse of the weighted response rate for the entire sample. This is equivalent to using a single RHG. The second baseline is the new Cycle 6 *saut-de-l'ange* methodology, which we apply to the cumulative nonresponse at Cycles 2 through 5. The modelling process and creation of RHGs was done independently for each cycle.

There are many ways to measure the longitudinal consistency of a categorical variable. For instance, we could calculate the relative absolute difference for each category and sum over all categories. We shied away from this approach because it more heavily weights those variables that have many categories. Instead, we computed the relative absolute difference at the variable level. This method allows categories with larger estimates to be given more importance.

In general terms, define

$$I_{ijk} = \begin{cases} 1 & \text{when unit } i \text{ is in the category } j \text{ of variable } k \\ 0 & \text{otherwise} \end{cases} \quad (6.5-1)$$

where $j = 1, \dots, J_k$ and $k = 1, \dots, K$.

Then, $\hat{Y}_{jk}(t)$, the Cycle 1 estimate of variable k computed using the Cycle t weights (that were derived using a given weighting methodology) is

$$\hat{Y}_{jk}(t) = \sum_{i \in R(t)} w_{i, \text{post-stratified}}(t) I_{ijk} \quad (6.5-2)$$

Using $\hat{Y}_{jk}(1)$, the actual Cycle 1 estimate calculated using the methodology in [Section 6.1](#), we can compute the relative absolute difference for variable k as

$$r_k(t) = \frac{\sum_{j=1}^{J_k} |\hat{Y}_{jk}(1) - \hat{Y}_{jk}(t)|}{\sum_{j=1}^{J_k} \hat{Y}_{jk}(1)} \quad (6.5-3)$$

Since the denominator represents the combined target population of all J_k categories, (6.5-3) can be seen as the total absolute difference as a percentage of the target population. Alternatively, (6.5-3) can be interpreted as a weighted average of the category relative absolute differences when written as

$$r_k(t) = \frac{\hat{Y}_{1k}(1)}{\sum_{j=1}^{J_k} \hat{Y}_{jk}(1)} \times \frac{|\hat{Y}_{1k}(1) - \hat{Y}_{1k}(t)|}{\hat{Y}_{1k}(1)} + \dots + \frac{\hat{Y}_{J_k k}(1)}{\sum_{j=1}^{J_k} \hat{Y}_{jk}(1)} \times \frac{|\hat{Y}_{J_k k}(1) - \hat{Y}_{J_k k}(t)|}{\hat{Y}_{J_k k}(1)} \quad (6.5-4)$$

As an example, refer back to Table 5.1.1.1-2 and the gender variable, $r_{\text{Gender}}(2)$. The relative absolute difference is

$$r_{\text{Gender}}(2) = \frac{|1,000 - 1,029|}{2,000} + \frac{|1,000 - 972|}{2,000} = 2.9\% \quad (6.5-5)$$

According to our view of longitudinal consistency, estimates of static characteristics from Cycle 1 should not deviate substantially between cycles. Therefore, once we have determined $r_k(t)$ for all K variables, we seek to measure this deviation for each characteristic by looking at the parameters of the distribution of the $r_k(t)$.

For the evaluation here, we let K be the entire set of categorical variables from Cycle 1. We chose to use all the variables available to us rather than a subset of characteristics since it would have been possible to find a subset of variables for which the average relative absolute difference under one nonresponse adjustment method would be greater or smaller than the total under another method. By using all the variables we felt that the subjectivity of choosing any particular set of variables was eliminated, and that this ensured that an adequate picture of the overall performance of the various nonresponse adjustment methodologies was given.

In Table 6.5-1, the mean, first quartile (Q1), median, third quartile (Q3) and maximum value of $r_k(t)$ are computed. This gives a comparison of Cycle 1 estimates using weights from Cycles 2 to 6 derived using different nonresponse adjustment methodologies. The tighter the distribution of $r_k(t)$ is around zero, the more longitudinally consistent are the estimates. From the table, notice that the uniform nonresponse adjustment is the worst at maintaining longitudinal consistency, with the mean increasing faster than the other methods as the nonresponse rate grows. The mean and median of the other approaches are fairly similar, but with the *saut-de-l'ange* method generally doing better. The third quartile under the *saut-de-l'ange* method is lower than the third quartile under the hybrid and *saut-de-mouton* methods. For all methods, we also see that the longitudinal consistency deteriorates as the nonresponse rate increases, shown by the ever increasing median. However, estimates under the *saut-de-l'ange* method are losing longitudinal consistency at a slower rate than the other methods.

Cycle	Cumulative nonresponse (%)	Nonresponse weighting methodology	$r_k(t)$ (%)				
			Mean	Q1	Median	Q3	Max
2	21	<i>Classic saut-de-mouton</i>	1.0	0.4	0.8	1.7	5.1
		<i>Saut-de-l'ange</i>	1.0	0.4	0.8	1.3	5.2
		Uniform adjustment	1.6	0.5	1.2	2.4	7.1
3	23	Hybrid <i>saut-de-mouton</i> / <i>saut-de-l'ange</i>	1.2	0.5	0.9	1.9	6.7
		<i>Saut-de-l'ange</i>	0.9	0.4	0.8	1.3	4.7
		Uniform adjustment	1.7	0.6	1.2	2.7	6.2
4	31	<i>Saut-de-mouton</i> adapted to handle converted units	1.6	0.6	1.1	2.9	7.7
		<i>Saut-de-l'ange</i>	1.2	0.6	1.0	1.6	7.2
		Uniform adjustment	2.5	0.7	1.8	4.3	8.8
5	34	<i>Saut-de-mouton</i> adapted to handle converted units	2.0	0.7	1.2	3.4	8.3
		<i>Saut-de-l'ange</i>	1.6	0.6	1.2	2.1	8.4
		Uniform adjustment	3.0	0.9	1.9	5.0	10.9
6	38	<i>Saut-de-l'ange</i>	1.6	0.8	1.4	2.3	8.7
		Uniform adjustment	3.5	1.0	2.3	5.9	12.5

Table 6.5-1: Comparison of various nonresponse weighting methodologies, measuring the longitudinal consistency, $r_k(t)$, for all Cycle 1 categorical variables.

While improvements in longitudinal consistency have been realized with the *saut-de-l'ange* methodology, further development of the weighting methodology is needed. Additional improvements to the modelling process are presented in the following section. One of the possible improvements, the incorporation of interaction terms, had been investigated at the time of Cycle 6 production. However, the frame variables used to develop the interaction terms did not appear to improve either the model fit or the longitudinal consistency arising from the resulting nonresponse adjustment. With the use of more longitudinal data, however, interaction terms involving NLSCY survey data may turn out to play a more important role.

7.0 Post-stratification and other issues

In this section we consider post-stratification issues that may arise in a longitudinal survey, along with other related issues.

7.1 Post-stratification

Post-stratification ensures that the final survey weights match certain known (usually demographic) population counts. The framework for this idea is best conveyed by the concept that the French school of sampling uses for post-stratification: *redressement*. It is not an easy word to translate in this context. The verb *redresser* means to “to straighten up”, so *redressement* is akin to “straightening up”, as in straightening up a young tree by tying it to a stake. This is the idea behind *redressement*: straightening up something that was not going as planned, without necessarily expecting that the end result would be exactly as if everything went according to plan from the start. (In horticultural terms, we are not necessarily trying to turn the young tree into something as straight as the stake it is tied to.) If the end result is exactly as expected then this is a bonus; otherwise, with *redressement* there is the idea that we are content to have a very close approximation of it. This is not something that we are used to seeing in practice, as post-stratification is mostly used to force some estimates to match exactly known totals.

In the case of the NLSCY, the 240 crossings of age, gender and province constitute the post-strata, and the demographic projections of these are considered to be known counts. Post-stratification in the NLSCY is used specifically to address undercoverage of the target population that we know to exist for the survey population, but the improvement on performance of the estimates is global. (Even when there is no reason to suspect either under or overcoverage, post-stratification usually is beneficial.)

Incidentally, while most surveys post-stratify like the NLSCY to information not contained on their frame, a survey could very well do just that. In other words, the information could be there on the frame (and used for stratification purposes, for instance) but be ignored up until the sample is selected. This is called by some (for further emphasis) post-sampling stratification, and it is a technique that allows a multi-purpose survey to select a sample under a “non-committal” design like SRS and only incorporate the (appropriate) frame information in the estimator to be used. If analysts were to build their own survey weights (instead of getting them from Statistics Canada as it is the case), then such an approach could be envisioned.

Some people are concerned about reporting a zero variance estimate when the survey estimate is a post-stratum total or, more generally, about the possibility that by using too many post-strata, the variance estimate for a given domain can be artificially reduced.

Further exploration

To read more on post-stratification and its role under variance estimation, see [*Perspective K*](#).

For cross-sectional surveys, determining the post-stratum size using the original sample size is normally sufficient because nonresponse is relatively low. On the other hand, some longitudinal cohorts are long-lived: they have to contend with nonresponse that can accumulate to high levels over time. Over time, this can lead to under-populated post-strata (*i.e.*, few respondents) and consequently unstable weight adjustments. It is important that the post-strata be created at the beginning of a cohort with ample effective size so that they can remain adequately populated by respondents throughout the cohort's life because no solution to underpopulated post-strata is satisfactory. For instance, collapsing sparsely populated post-strata is a possibility, but it is preferable that the same post-strata are the same for every cycle of a longitudinal survey. Consequently, care should be taken when choosing the post-stratification cells at the start of any longitudinal survey. Not only should the original sample size be taken into account, but so should the impact that the anticipated nonresponse will have on the bootstrap replicates (and not just the main sample) over the survey's lifespan.

This is what we are witnessing with the Original Cohort. Just like all the other NLSCY cohorts, the Original Cohort is post-stratified by age, gender of the child and province of residence. For the Original Cohort this represents 240 post-strata (10 provinces by two genders by 12 one-year age groups). While at Cycle 1 the available sample size may have justified such a fine partition of the sample into post-strata, it may no longer be the case by Cycle 6 and beyond. More precisely, while the set of respondents at Cycle 6 is large enough to ensure non-empty post-strata, empty groups are encountered in the bootstrap. As explained in [*Perspective B*](#) on the bootstrap, only about 60% of the originally sampled units enter a replicate (some more than once) while the others are simply not selected by the with-replacement scheme used by the bootstrap. Now, if there are 7 or 8 units in a post-stratum, then we can expect on average 1 out of every 1,000 replicates to be devoid of post-stratum units (*i.e.*, by leaving about 40% of units out, once in a while all 7 or 8 units will happen to not be selected). This rough combinatorial argument is not to be taken too seriously since other considerations are at work (like re-sampling PSUs, not necessarily units *per se*) but still it sheds some light onto the issue. For the NLSCY, empirically we find that a post-stratum is at risk of turning up empty for at least one replicate whenever it contains fewer than 10 respondents.

Although none of the 240 post-strata was empty as of Cycle 7, some were empty for at least one bootstrap replicate. This is the case for 2-year-old boys in PEI for the Original Cohort. This results in difficulties in the estimation of the variance of such post-strata totals since one or more replicates are lost. This problem will only be exacerbated as cycles go by and may need to be permanently addressed, but how?

If we were to start a new cohort of the magnitude and longevity of the Original Cohort, then we would need to rethink the construction of post-strata to avoid empty post-strata from surfacing at some point in the bootstrap replicates. This could be done by either choosing a smaller set of post-strata or borrowing from the idea of *redressement* introduced above and have something like quasi-post-stratification. In this latter case, we would start with a preliminary set of post-strata, possibly the same 240 post-strata that we have for the Original Cohort, and group these into larger final post-strata based on similar adjustments. For example, two post-strata that have nothing to do with one another other than that they lead to an adjustment by 5%, say, would get merged. Since merged post-strata would have slightly different preliminary post-strata adjustments, the larger final post-strata would lead to an adjustment that is possibly different. In other words, the preliminary post-

strata totals would not be exactly met, but only approximately. (In practice, the main nuisance would be to have preliminary post-strata survey estimates that are non-integers, and therefore come with a variance estimate that is tiny but still non-zero.) Quasi post-stratifying this way would certainly be beneficial to the overall performance of the estimates just like an equivalent but more demanding post-stratification would be, provided we can live with the fact that known totals are not necessarily matched exactly. Another approach is to realize first that post-stratification is the finest form of calibration: marginal totals, like provincial totals, are met because the provincial subtotals by province, age and gender, are themselves met. Here, a coarser calibration could be used where, for instance, only the provincial, age and gender marginals are met, not their combinations: age and gender and province.

But what about an existing cohort like the Original Cohort which is struggling with this problem already? In such a context a natural option is to collapse post-strata. We can collapse based on characteristics, for example, ignoring gender, or by grouping post-strata with similar adjustments. For the latter option, we can resort to advanced clustering algorithms to implement it in practice. But no matter how it is implemented, strata collapsing will introduce a break in the cohort's methodology. Furthermore, and perhaps even more importantly, will it solve the issue with respect to variance estimation? The answer to this question is unfortunately "no". Indeed, any revised set of post-strata will have the potential of leading to inappropriate variance estimates if we are not careful, just like with the original set we have. In other words, if we are careful, then the issue of empty post-strata in some replicates will not compromise in itself the results: it is at this point more an issue of how the replication data are analyzed than anything else.

In *Perspective C* we discuss at length what the analyst needs to do to address the situation appropriately, and what follows assumes that the reader is familiar with this. Specifically, we argue in *Perspective C* that the analyst may need to discard some replicates in a given analysis if the estimates they give rise to are found to be outlying values. Unless the problem is severe, only a handful of replicates will be ignored and the loss of analytical power will be minimal; it is certainly less harmful to ignore some replicates than to ignore outlying estimates and use all replicates.

Collapsing the current post-strata into any revised set of post-strata will not relieve the analyst of his or her responsibility to critically analyze the replicate information at hand. Indeed, some domains require close monitoring by the analyst to ensure that the variance estimate makes sense. This would be the case still of the old defective post-strata under the revised methodology. Consequently, if there is no way to alleviate the burden on the analyst by taking care entirely of defective post-strata by collapsing, then it is probably wiser not to collapse. Indeed, collapsing does introduce issues of its own, like endangering the variance estimates of not only the problematic post-strata but also of the well-populated post-strata that are part of the merger.

7.2 Weighting sub-components of interest of the NLSCY

In addition to the master files that NLSCY releases at each cycle – these are about the entire cohort they pertain to (*i.e.*, the entire sample) – some specific subsets are also, or have also been, released. These are the so-called restricted-access file and the education component file. The restricted-access file, released at each cycle, only contains those respondents of the Original Cohort who explicitly agreed to share their data with the survey's sponsor, currently known as Human

Resources and Social Development Canada (HRSDC). The restricted-access file represents about 97% of the full sample. To give a flavour of the issues that may arise with the release of chosen sub-components of a survey, the restricted-access file is used in the following example.

The biggest danger with the restricted-access file is to take too lightly the issue of weighting because of the very low proportion of non-sharers: about 3%. One might be tempted to think that it is not important how weighting (or re-weighting) is done to account for the missing 3% because the proportion is so small. It is true that various re-weighting methodologies will look the same, based solely on the point estimates each will give rise to. But here is the trap – and one which is too easily overlooked – the associated variance estimates may nonetheless differ significantly. In other words, the issue does not only lie in the first-order properties of the methodology used, but also in the second-order properties.

For the restricted-access file, there are several possibilities for re-weighting. One possibility is to inflate uniformly the weight of children in the restricted-access file by the inverse of the weighted proportion it represents; another is to inflate their weights within post-strata (*i.e.*, let post-stratification do the re-weighting); another is to re-do the weighting from start to finish (*i.e.*, nonresponse and post-stratification). In the first scenario, the restricted-access file is treated as a domain of the whole population and, as such, the post-strata estimates based on the members of the restricted-access file need not agree with the demographic projections on them (this is not desirable). With the other two options, the restricted-access file leads to estimates of post-strata population counts which match the demographic projections.

Either way, the variance estimation has to be conducted adequately. This requires re-weighting within each replicate, either uniformly or within post-strata accordingly, and not use the weighting adjustment obtained for the main sample in their stead.

In the case of the uniform adjustment, this means that if we need to increase the final weights by 3%, then we have to re-calculate the uniform adjustment for each replicate, and not simply increase automatically the bootstrap weights by the 3% determined from the full sample. For the adjustment within post-strata, the same approach is to be used, but instead of the whole file, the adjustment is re-computed within each post-stratum and replicate. The adjustment obtained based on the sample is not to be used. While it may appear simpler to take the final file and subset it to the restricted-access file, and re-post-stratify again to obtain the final restricted-access file, we recommend going through the entire process already established for the whole sample (with the exception of re-constructing the RHGs; use the same RHGs, but re-calculate the nonresponse adjustment). The advantage of using the same computer programs but on a subset now is that it does not introduce another process altogether to treat the restricted-access file. Though it looks simple enough to simply subset the final file and post-stratify it again, looks can be deceiving, and in our experience it is better to go with a well-established process.

8.0 Conclusion

The 10-year review of the NLSCY methodology ([Statistics Canada \(2007\)](#)) resulted in a re-evaluation of the entire weighting and variance methodologies. This working paper covers what we learned in developing a suitable weighting methodology for a longitudinal survey like the NLSCY. (For everything we learned on the variance estimation front, see the bootstrap working paper [Girard \(2007\)](#).)

Specifically, we discussed in detail the impact that different nonresponse models can have on the survey estimates. We compared imputation with nonresponse weight adjustments, and in the case of a weight adjustment:

- how to create the RHGs (segmentation versus response propensity scoring),
- whether to model all of the accrued nonresponse in one step (*saut-de-l'ange*) or cycle-by-cycle (*saut-de-mouton*),
- whether to calibrate to previous cycles' estimates,
- how to evaluate nonresponse bias using a measure of longitudinal consistency,
- how to handle post-stratification and related weighting issues.

We tried to address these weighting issues in an integrated fashion by considering their impact on the whole survey, including sampling but especially variance estimation using the bootstrap.

In addition, the various *Perspectives* detail the discussions and conclusions we (the NLSCY methodologists) reached on various topics that influenced our decisions regarding the final choice of a weighting methodology, for example:

- issues behind trying to produce one-size-fits-all survey weights,
- an overview of various bootstrap issues,
- complications arising from unequal design weights (which the NLSCY inherits from the LFS),
- whether to use weights when constructing RHGs, and having constructed the RHGs, whether to use weights to calculate the response propensities,
- a discussion of modelling, and whether nonresponse models affect our inferential (design-based) framework.

The NLSCY methodologists continue to make improvements to the weighting methodology, and in particular, to search for ways to further mitigate nonresponse bias. We hope that the lessons learned by NLSCY methodologists benefit those working on other (particularly longitudinal) surveys.

Acknowledgements

The editors and contributors would like to thank Joanne Moloney and Martin Renaud for their active participation in the review process of this paper.

References

- Alavi, A. and Simard, M. (2006). Weighting and Estimation in the Presence of Nonignorable and Other Nonresponse. Household Survey Methods Division. *Statistics Canada internal document*.
- Ardilly, P. (2000). *Les techniques d'enquête*. Éditions Technip, Paris.
- Ardilly, P. and Tillé, Y. (2003). *Exercices corrigés de methods de sondage*. Éditions Ellipses, Paris.
- Beaumont, J.-F. and Mitchell, C. (2002). A System for Variance Estimation Due to Non-Response and Imputation (SEVANI). *Proceedings of the Statistics Canada Symposium 2002*, Statistics Canada.
- Birta, L.G. and Arbez, G. (2007). *Modeling and simulations: exploring dynamic system behaviour*. Springer, London.
- Brick, J.M. and Kalton, G. (2003). Handling Missing Data in Survey Research. *Statistical Methods in Medical Research*, 5, 215-238.
- Burkam, D.T. and Lee, V.E. (1998). Effects of Monotone and Nonmonotone Attrition on Parameter Estimates in Regression Models with Educational Data: Demographic Effects on Achievement, Aspirations, and Attitudes *The Journal of Human Resources*, Special Issue: Attrition in Longitudinal Surveys, 33(2), 555-574.
- Deville, J.-C. and Särndal, C.-E. (1992). Calibration Estimators in Survey Sampling, *Journal of the American Mathematical Association*. 87(418), 376-382.
- Dudewicz, E. J. and Mishra, S.N. (1988). *Modern Mathematical Statistics*. Wiley, New York.
- Eltinge, J.L. and Yansaneh, I.S. (1997). Diagnostics for formation of nonresponse adjustment cells with an application to income nonresponse in the U.S. Consumer Expenditure Survey. *Survey Methodology*, 23, 33-40.
- Fitzgerald, J., Gottschalk, P. and Moffitt, R. (1998). An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study of Income Dynamics. *The Journal of Human Resources*, 33(2), 251-99.
- Girard, C. (2007). How to avoid getting all tied up bootstrapping a survey: A walk-through featuring the Canadian National Longitudinal Survey of Children and Youth. *Statistics Canada Methodology Branch Working Paper*, HSMD-2007-001E.
- Graubard, B. (2002). Analysis of Surveys and Applications to Health Data. Workshop given at the Statistics Canada Symposium.
- Haziza, D. and Beaumont, J.-F. (2007). On the construction of imputation classes in surveys. *International Statistical Review*, 75, 25-43.

- Kass, G.V. (1980). An explanatory technique for investigating large quantities of categorical data. *Applied Statistics*, 29, 119-127.
- Kim, J.K. and Kim, J.J. (2007). Nonresponse weighting adjustment using estimated response probability. *The Canadian Journal of Statistics*, 35, 501-514.
- Lafortune, Y. (2008). Workshop given at Toronto Research Data Centre, <http://www.utoronto.ca/rdc/events.html#NLSCY2008>
- Little, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54, 139-157.
- Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.
- Lohr, S.L. (1999). *Sampling: Design and Analysis*. Duxbury Press, New York.
- Loosveldt, G., Pickery, J. and Billet, J. (2002). Item nonresponse as a predictor of unit nonresponse in a panel survey. *Journal of Official Statistics*, 18, 545-557.
- Michael, R.T. and O'Muircheartaigh, C.A. (2008). Design priorities and disciplinary perspectives: the case of the US National Children's study. *Journal of the Royal Statistical Society. Series A*, 171(2), 465-480.
- Michalewicz, Z. and Fogel, D.B. (2000). *How to solve it: modern heuristics*. Springer, New York.
- Picot, G. and Webber, M. (2005). Taking stock: The future of longitudinal surveys. *Proceedings of the Statistics Canada Symposium 2005* [CD-ROM], Statistics Canada.
- Rao, J.N.K. and Wu, C.F.J. (1988). Re-sampling Inference With Complex Survey Data. *Journal of the American Statistical Association*, March 1988, Vol. 83 (401), 231-241.
- Rao, J.N.K., Wu, C.F.J. and Yue, K. (1992). Some recent work on re-sampling methods for complex surveys. *Survey Methodology*, 18(2), 209-217.
- Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley, New York.
- Rust, K. and Rao, J.N.K. (1996) Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, 5, 283-310.
- Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.
- Shao, J. and Sitter, R.R. (1996). Bootstrap for Imputed survey Data. *Journal of the American Statistical Association*, 91, 1278-1288.
- Scott, A. (2006). Population-Based Control Studies. *Survey Methodology*, 32(2), 123-132.

Singh, A., Wu, S. and Boyer, R. (1995). Longitudinal survey nonresponse adjustment by weight calibration for estimation of gross flows. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 396-401.

Statistics Canada (2007). The National Longitudinal Survey of Children: 10-year review of the methodology. (ed. S. Franklin). Household Survey Methods Division. *Statistics Canada internal document*.

Statistics Canada (2008). *Methodology of the Canadian Labour Force Survey*. Catalogue no. 71-526-X.

Tam, M., Tremblay, M., Franklin, S. and Girard, C. (2007). Improving the unit nonresponse adjustment in the NLSCY using logistic regression modelling and calibration. *2007 Proceedings of the American Statistical Association, Survey Research Methods Section [CD-ROM]*, American Statistical Association, 3316-3322.

Tam, M. (2008). Weighting Issues at Cycle 7 of the National Longitudinal Survey of Children and Youth. Presentation to the Household Surveys Methods Division Technical Committee, Statistics Canada, http://method/BiblioStat/Research/TechCom/HouseholdSurveys/Minutes/index_e.htm

Tremblay, M. and Franklin, S. (2008). The Rationale for the NLSCY Design Weight Methodology. Household Survey Methods Division. *Statistics Canada internal document*.

Tremblay, M., Franklin, S., Meyer, S., Tam, M. and Zheng, A. (2009). The NLSCY Samples for Cycles 1 to 8 and their Design Weights. Household Survey Methods Division. *Statistics Canada internal document*.

Valliant, R., Dorfman, A.H. and Royall, R. M. (2000). *Finite Population Sampling and Inference – A Prediction Approach*. Wiley, New York.

Watson, D. (2003). Sample attrition between waves 1 and 5 in the ECHP. *European Sociological Review*, 19, 361–378.

Watson, N. and Wooden, M. (2006). Identifying Factors Affecting Longitudinal Survey Response. Paper presented at the Methodology of Longitudinal Surveys International Conference, University of Essex, Colchester, UK, July 12-14, 2006.

Perspective A: On the challenges of producing multi-purpose weights

The NLSCY, like many other household surveys at Statistics Canada (if not all), releases a single set of weights with its data. More precisely, the NLSCY releases one set of weights for each population of interest and definition of a longitudinal respondent, as there are different sets of weights for longitudinal and cross-sectional analyses for instance. Contrary to what is often seen on the business surveys' side, the scope of analytical studies spawned by social/household surveys like the NLSCY is huge and not driven by just a few characteristics (like business income is to many business surveys). Consequently, with a unique set of weights, everyone's expectations have to be kept at a reasonable level with regard to the quality of the inferences that can be built from them. Indeed, with a unique set of weights to be released, one may feel compelled to create them in such a way as to address all analytical needs at once, or at least, *try*. But this is like trying to dress a bunch of men with a one-size-fits-all suit when clearly separate tailored suits would be in order. There is a limit to what a one-size-fits-all approach can accomplish and as long as we are aware of this, everything should be okay. If there is no compromise to be made with respect to the goals to be attained, then a tailored approach is in order. The same is true for weights: it is a difficult (if not impossible) task to design them as to address *a priori* just about any analytical need out there.

Methodologically speaking, to have a unique set of released weights is a very restrictive setting to work from, one from which the survey statistician cannot do wonders, especially when attempting to reduce the impact nonresponse is having on the inference. The lack of a main variable of interest makes it difficult to channel the efforts into one adjustment; the survey statistician is left trying to compensate for pretty much *all* of the effects of nonresponse. This is what has been driving the NLSCY nonresponse weighting strategy in the past and likely will in the future.

Ideally, there would be a separate set of weights issued for any given analysis to reflect the fact that different situations call for different methodologies to be used. But since there is no way with social surveys to know in advance how the data will be used, we cannot commit ourselves too strongly to any given specialized methodology. Indeed, this would compromise the universality of application of the weights produced. (In other words, if one knows the weights will be used in a variety of situations, do not over-tailor them to one specific situation but rather stick to the one-size-fits-all approach as much as possible.)

If we cannot produce the tailored weights beforehand, what about leaving it to the analysts themselves? For example, the survey statistician could produce a set of generic weights to start with (and let us purposely keep vague what "generic" means) that analysts would turn into final weights to suit their own analytical needs. One issue with this approach is the burden it puts on analysts (since they would need to create bootstrap weights on their own, for instance). Another issue has to do with the fact that with user-defined weights, the results obtained can hardly be reproduced by others unless a significant amount of documentation is provided by the analyst about the methodology used.

Perspective B: An overview of the bootstrap as used in the NLSCY

In order for the construction or the refurbishment of a weighting methodology to be successful, considerations must be given at the same time to the variance estimation methodology. Since the bootstrap is the method of variance estimation for the NLSCY (as it is for many other household surveys at Statistics Canada), we need to understand its key features. We thus propose here an overview of the bootstrap as NLSCY uses it, which is more succinct than the account given in [Girard \(2007\)](#) upon which much of what we present here is based. Some features presented here are new and are consequently presented in more detail.

Context

To grasp what the bootstrap is about, a few words on variance estimation are in order. Suppose we are interested in estimating a population characteristic through some estimator which channels the information obtained from a sample of units, the sample having been drawn using a given sampling design. An estimator (also called by some a statistic) is a mathematical function of the sample obtained under the sampling design which assigns to a given sample a numerical value, the estimate. The estimate can be seen as a very condensed, yet pertinent, summary of the information contained in the sample about the characteristic of interest. Going through all possible samples that the sampling design can create, we obtain all possible estimates the estimator can possibly give rise to. The probability distribution on the various outcomes, the estimates, is the sampling distribution (of the estimator). Though discrete, the sampling distribution is built from far more samples than can actually be drawn using a computer given the current state of computer technology. The basic statistical properties of the estimator, like its bias and variance, are derived from the relevant properties of its distribution (*e.g.*, its bias is the difference between the expected value of the distribution and the value of the parameter of interest). It is common in practice to hear of the bias or the variance of an estimate which is misuse of language since the estimate is not a random quantity *and* bias and variance are attributes of a random process. It is preferable then to speak of the bias or variance associated with the estimator, which makes it more explicit that it is not an attribute of the estimate but rather is inherited from the underlying estimator. Incidentally, because the stochastic behaviour of the estimator (and thus not of the estimate) is described here as inherited from the stochastic nature of the sampling selection, this means we are design-based. (For more information about the design-based inferential framework, consult [Perspective I.](#))

Figure B-1 below shows what a sampling distribution of the estimator can look like. Strictly speaking, in this case it is the Monte Carlo approximation of the sampling distribution which is portrayed since, as we have mentioned above, even in a simulation setting where we are able to draw multiple samples we cannot usually hope to draw *all* possible samples: there are usually just way too many of them. The distribution, which happens to be normally distributed in this case, is shown with its best-fit curve, the continuous line. The value of the parameter of interest in this fictitious case is 5; from this, the symmetry and where the axis of symmetry is located, it is clear that the estimator used is unbiased for the parameter value of interest.

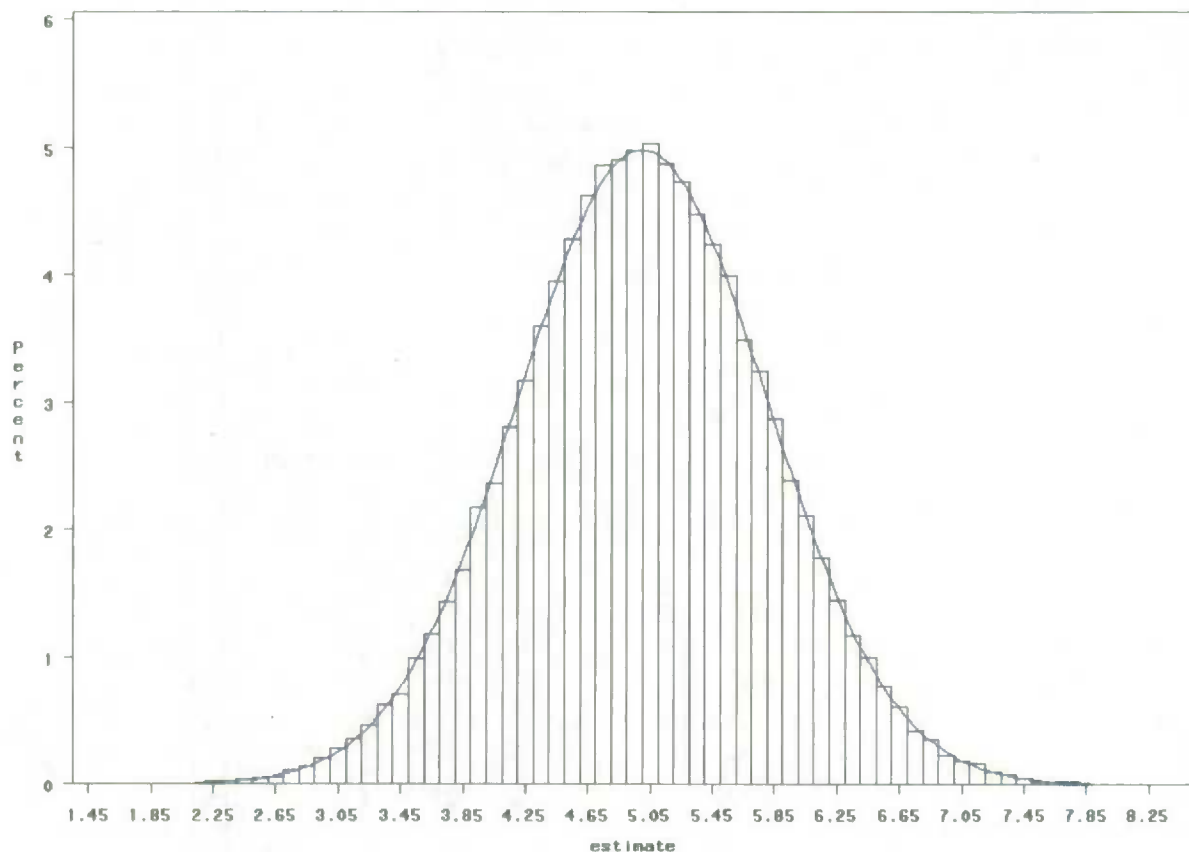


Figure B-1: The sampling distribution of an estimator

To obtain in practice an estimate of the variance of the estimator appears impossible since, by definition, we only have the benefit of observing one sample when all of them (or at least some large number of them, as in a Monte Carlo simulation) are needed. This is where the replication underlying the bootstrap comes into play. By drawing (sub-) samples from the observed sample in a clever way (called bootstrap replicates or simply replicates if it is clear from the context that the replication method used is the bootstrap), and getting from them the appropriate replicate-based estimates, it is often possible to assign a probability distribution to the replicate-based estimates such that its shape matches the shape of the sampling distribution. And if the two distributions are close enough in shape (*i.e.*, if the bootstrap distribution does converge to the sampling distribution of the estimator, then it means in particular that their variances are also close).

When the bootstrap distribution converges to the sampling distribution for a given estimator *and* sampling design, then we say in day-to-day parlance that the bootstrap works (for that estimator *and* design). This is, incidentally, why the properties of the bootstrap described in the literature have an asymptotic ring to them since it is usually in the limit that the two distributions really match. While the bootstrap can be shown to work in a given setting for two different estimators, the notion of convergence is key in practice to explain why the bootstrap may not work equally well in both cases. This is simply because the convergence may be slower for one estimator than for the

other. For example, under Simple Random Sampling With Replacement, simulations clearly show that for the same (small) sample size, the bootstrap distribution matches the sampling distribution much better in the case of the mean than that of the median.

Suppose now a sample was drawn, leading to an estimate of, say, 6 for the characteristic of interest. Figure B-2 below shows the probability distribution on the replicate-based estimates obtained bootstrapping (in a way to be made more precise below) the *one* observed sample that is available. The best-fit curve of the *sampling* distribution displayed in Figure B-1 above is carried over to Figure B-2 to facilitate comparisons between the two distributions. It is clear from Figure B-2 that while the bootstrap histogram is not located under the best-fit curve, it has about the same shape as the sampling distribution. This is not a coincidence but an instance of a general principle: when the bootstrap works, the bootstrap distribution is centered on the sample estimate value (and not the true population value) and, asymptotically, has the same shape as the sampling distribution.

In practice, as argued already, from the correspondence of the shapes of the two distributions, we can use the variance of the bootstrap distribution, which is known, as an estimate of the variance of the sampling distribution. The bootstrap variance calculated would match exactly the variance estimate obtained from the sample if we could obtain the complete bootstrap distribution, which we cannot in practice, as only B of all possible replicates can be handled by today's computers.

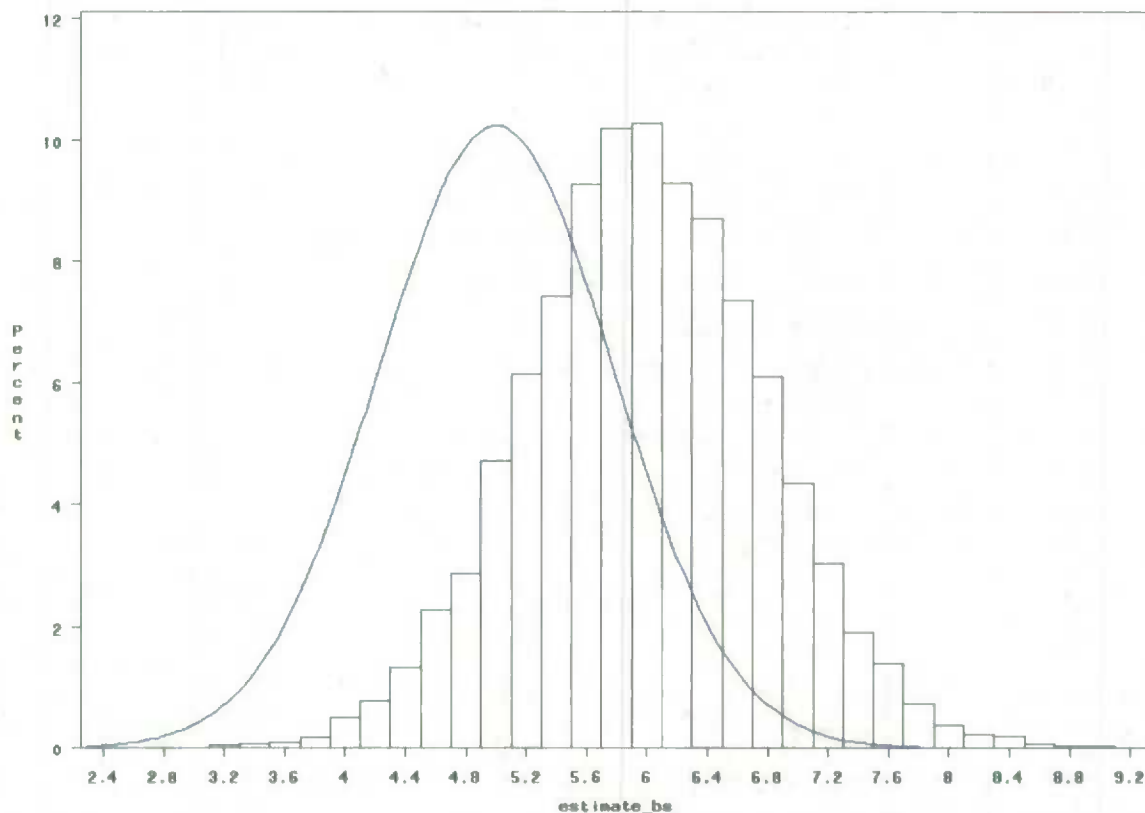


Figure B-2: The bootstrap distribution associated to the estimator, obtained from bootstrapping the sample

The various forms of bootstrap

The bootstrap method was introduced for the case of independent and identically distributed random variables which is well-suited to classical statistics, but not to survey sampling *per se*. This is why over the past years several methods have appeared in survey sampling which are variants of the original bootstrap method, all attempting to address one particular aspect or another of survey sampling. And this is still today a very much active field of research. One such variant of the bootstrap is of particular interest to us since it is the one most used at Statistics Canada and is actually employed by the NLSCY: the Rao-Wu rescaled bootstrap method ([Rao and Wu \(1988\)](#)). (In what follows, when we speak of the bootstrap, we refer to the Rao-Wu bootstrap variant.)

There are other replication methods which can be used by surveys, a popular one being the delete-1 jackknife (described briefly below); they are not to be mistaken one for the other.

Rao-Wu bootstrap: the implementation

To perform the bootstrap, first select with replacement from each stratum one less primary sampling unit (PSU) than what was selected to obtain the sample, and repeat this independently B times. (Observe the subtlety that we do not resample the number of PSUs in the sample which turned out to be useful but the whole set of PSUs, which includes the out-of-scope units.) In NLSCY, we take $B=1,000$. Some surveys use fewer replicates (*e.g.*, 500 or even 250). Generally speaking, the choice of B should be as large as possible while keeping well in mind the limited computer resources at our disposal.

While B may appear large (especially if we have run bootstrap estimates already and have seen how demanding the processing is on computers!), keep in mind that the entire re-sampling space is gigantic in size. Indeed, note that if one has just one stratum comprised of n units, then there are n^{n-1} possible replicates in the bootstrap re-sampling space. This number is huge and will far exceed whatever value is chosen for B in practice: for $n=50$, for example, the number 50^{49} has 84 *digits*! Compare that to the delete-1 jackknife re-sampling space which consists of only n replicates obtained by taking out one PSU from the sample at a time. Because with the bootstrap we do not usually come even close to exhausting all possibilities by taking B of them, we introduce in the grander picture of variance estimation a second source of error. Indeed, in addition to the sampling error which already exists, there is the additional error due to the entire set of replicates not being fully used: it is the bootstrapping error. We can reformulate what we have said about the jackknife with respect to the size of its re-sampling space by saying that there is usually no jackknifing error introduced in practice.

For a given sampled unit k and given replicate b we record the number of times k was chosen with replacement; this is called the unit's multiplicity (in the replicate) and is noted as $mult_{k,b}$. While the multiplicity of a unit has a range of values from 0 and $n_h - 1$, in practice a multiplicity larger than 5 rarely occurs since the low multiplicities are by far the most common events. More specifically, multiplicities of 0, 1 and 2 are the most likely ones to be observed in practice, with occurrence propensities of about 37%, 37% and 18% respectively, thus accounting for over 90% of multiplicities likely to be observed. The occurrence propensity associated to the multiplicity of zero has one important consequence: a given replicate is made of only about 63% of the initial sample.

Consequently, any small domain in light of the sample is likely to be much smaller in a given replicate. Therefore, domains which can be “controlled” by the survey statisticians, such as Response Homogeneous Groups (RHGs) and post-strata, should be built with this in mind as well (in addition to bias-reducing considerations) to avoid variance estimation complications later on. In other words, these groups should be of sensible sizes not with respect to the sample (which is what is usually done), but rather with respect to the replicates – which is harder to achieve.

For each replicate b and unit k , compute the following weight, called the b^{th} bootstrap weight of unit k using the design weight of the unit and some replication information:

$$bsw_{k,b} = w_k \times \frac{n_h}{n_h - 1} \times mult_{k,b} \quad (\text{B-1})$$

(We assume here that the number of PSUs selected is small compared to the number of eligible PSUs, which is the case with NLSCY; see [Girard \(2007\)](#) for details when this assumption is not reasonably met.)

This is the bootstrap counterpart of the design weight. It is interesting to note that the initial Rao-Wu bootstrap was not expressed in terms of weights (which came only later through [Rao et al. \(1992\)](#)) but in terms of a “one-piece” bootstrap estimator for the mean.

Bootstrap: nonresponse and post-stratification

Since at this point we have not addressed specifically how the nonresponse and post-stratification are to be handled, we can only describe the bootstrap approach with respect to these in general terms. The more specific details relative to each of the approaches will be made precise when they arise in the main text.

The idea is to re-evaluate, using the bootstrap weights, the adjustments obtained on the basis of the sample. Assuming that we have a set of weighting classes, or RHGs, to work from, we would compute within each replicate the adjustment that was obtained using the main sample. This means that the construction of the RHGs themselves is not revisited in each of the replicates; it is only the adjustments within the *existing* sample-based RHGs which are revisited.

The same goes for the handling of the post-stratification.

Bootstrap: cross-verification

At this point, the final bootstrap weights are available for release for analysts to use. But before the survey statistician responsible for them actually does so, there is a verification that can be performed to check the validity of the whole bootstrap implementation. (This came to our attention after [Girard \(2007\)](#) was written so it is described here in detail.) To help introduce it, consider the following snapshot illustrating what a file of multiplicities looks like.

VIEW TABLE: Work.Multiplicities										
	psu	M1	M2	M3	M4	M5	M6	M7	M8	
93	93	1	2	1	1	1	1	0	0	
94	94	0	0	1	1	1	0	2	2	
95	95	1	1	0	1	2	0	0	1	
96	96	0	0	0	1	1	1	1	1	
97	97	1	3	1	1	0	0	0	1	
98	98	2	0	0	4	1	1	2	0	
99	99	0	0	0	0	1	2	0	2	
100	100	2	2	0	0	1	1	0	2	
101	101	0	0	3	0	1	0	0	0	
102	102	0	0	1	0	1	1	3	0	
103	103	0	1	0	1	0	1	2	0	
104	104	0	0	2	0	1	0	0	2	
105	105	1	1	0	0	0	2	2	1	
106	106	1	0	1	1	4	2	0	1	
107	107	1	1	2	2	0	0	1	0	
108	108	2	2	0	1	0	0	0	1	
109	109	1	1	1	0	0	1	1	0	
110	110	2	1	0	2	0	1	2	0	
111	111	0	2	1	1	1	1	0	0	

The diagnostic, as applied to that file of multiplicities, is as follows:

1. Compute for each row (*i.e.*, for each PSU) the empirical variance of the 1,000 multiplicities:

$$\text{row variance} = \frac{\sum_{b=1}^{1,000} (m_b - \bar{m})^2}{1,000 - 1} \quad (\text{B-2})$$

2. Compute for each row the following relative difference (in percentage points) of that variance to its expected value under the bootstrap, which is of $(n-1)^2 / n^2$, n being the number of PSUs in the same stratum as the PSU itself:

$$\text{ratio of the variances} = \frac{\text{row variance}}{(n-1)^2 / n^2} \quad (\text{B-3})$$

Note: To explain the variance figure of $(n-1)^2 / n^2$, let m represent random variable of the bootstrap multiplicity, of which m_1 to m_{1000} depicted in the snapshot above are realizations. By definition of the sampling with replacement of $n-1$ among n which underlies the bootstrap, it follows that m is a Bin($n-1, 1/n$) variable. Indeed, there are $n-1$ "trials" for which a given unit has a chance of $1/n$ of being selected at each of the trials and independently from one to the other. Therefore, m has an expected value of $(n-1) \times 1/n$ and a variance of $(n-1) \times 1/n \times (1 - 1/n) = (n-1)^2 / n^2$.

3. For a given set of multiplicities, plot the histogram of these ratios obtained over all rows. The histogram should be centered on 1 and show little spread.

This diagnostic can actually be used at all the stages of the bootstrap weighting, from the creation of the multiplicities to the delivery of the final bootstrap weights. (Use bootstrap weights in lieu of the multiplicities in B-2 and use the final weight in the denominator of B-3 instead of $(n-1)^2 / n^2$. The ratio of the variances should still be centered on 1 and show little spread, though more with bootstrap weights than with the multiplicities themselves.)

To illustrate, let us compare three cases:

- multiplicities of the Original Cohort,
- final bootstrap weights for Cycle 1,
- final bootstrap weights for Cycle 6.

The final bootstrap weights at Cycle 6, just like those of Cycle 1, are built directly from the multiplicities of the Original Cohort. In the case of the Cycle 6 bootstrap weights, this is significant since they could very well have been built from the previous cycle bootstrap weights, not straight from the multiplicities themselves. This is a consequence of the *saut-de-l'ange* weighting methodology that was implemented for the Original Cohort starting at Cycle 6. It has the benefit of not making the Cycle 6 bootstrap weights dependent on earlier cycles' bootstrap weights, especially if concerns exist on how these have been obtained.

In Figures B-3 and B-4, we present the histogram of the ratio of row-variances for the multiplicities of the Original Cohort and the final bootstrap weights for Cycle 6, respectively. Observe how tight both are around 1, the central value, though the Cycle 6 histogram (Fig. B-4) shows a bit more spread and lack of symmetry than the histogram built from the multiplicities (Fig. B-3).

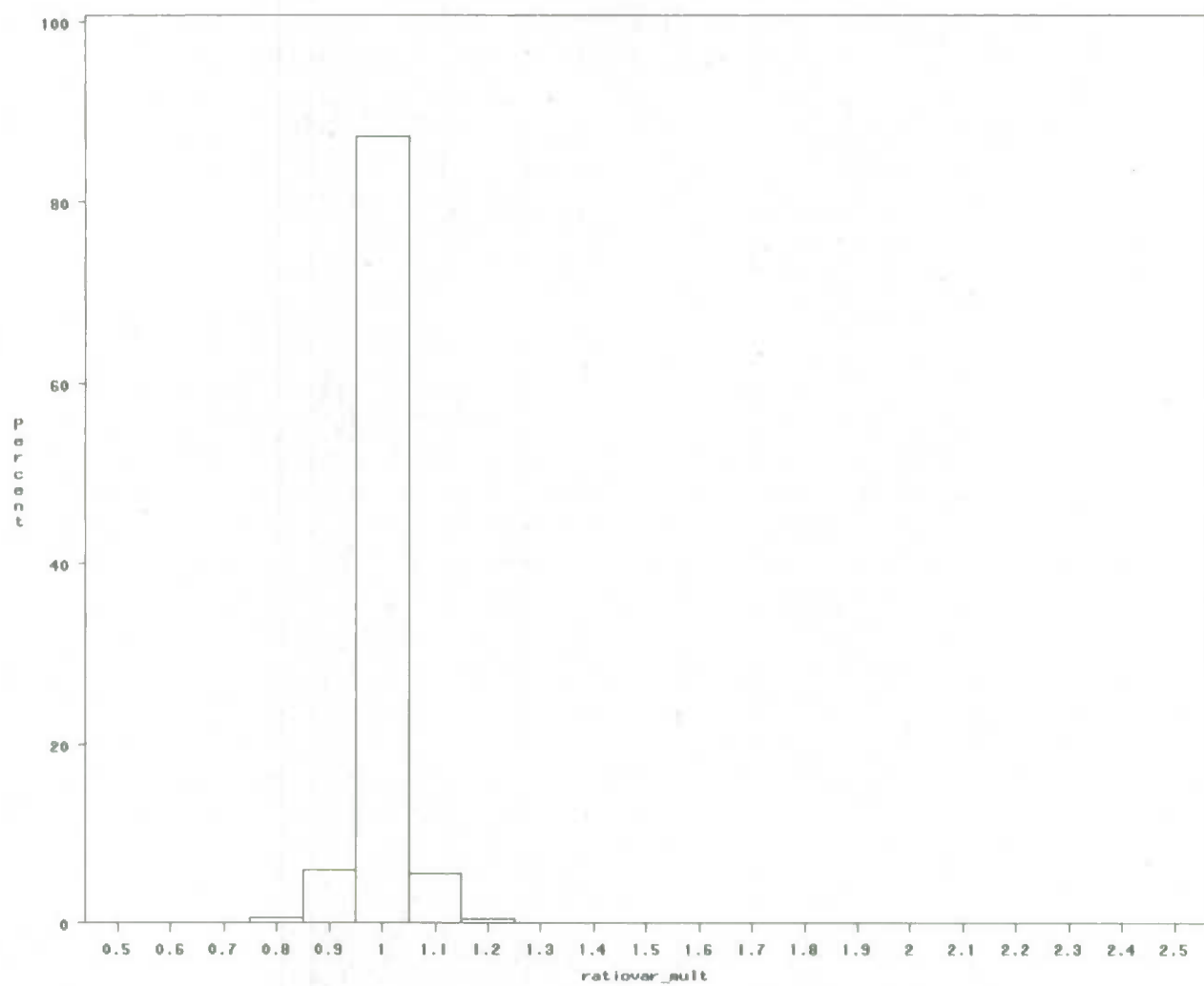


Figure B-3: Ratio of row-variances in the case of the multiplicities file for the Original Cohort.

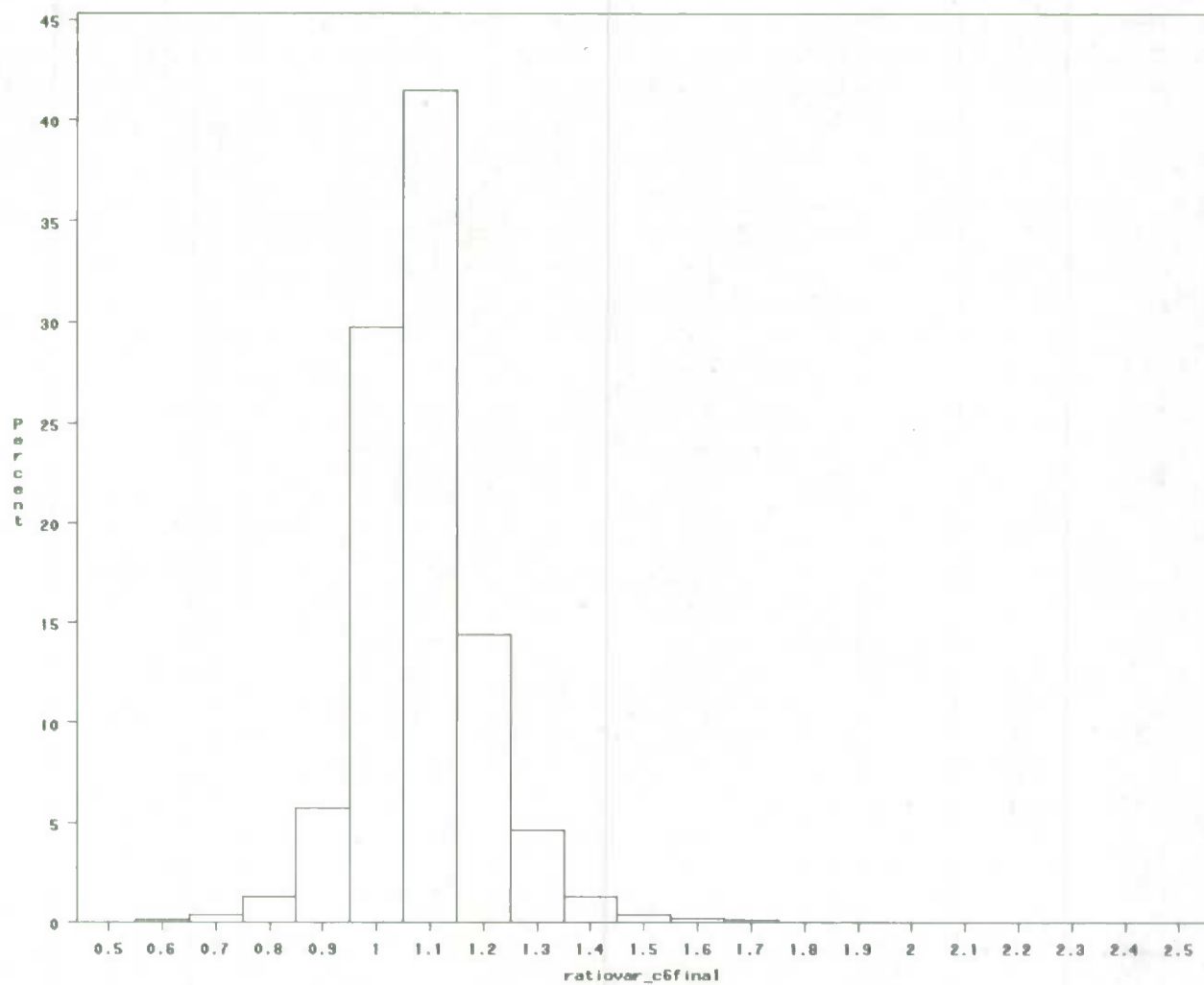


Figure B-4: Ratio of row-variances in the case of the final weights for Cycle 6 of the Original Cohort.

In Figure B-5, we present the histogram of the ratio of the variances for the final bootstrap weights for Cycle 1 of the Original Cohort: it is significantly different than the one based on multiplicities (Fig. B-2). Observe the range of ratios used in B-5, which extends on the right beyond 40 while the two histograms above did not reach 2.6. Furthermore, if we zero in on the (sub-) range 0 to 2.6 of Figure B-5, then we obtain the histogram depicted in Figure 2.6. It is clear that even for the same range, the histogram at cycle 6 does not compare to the other two.

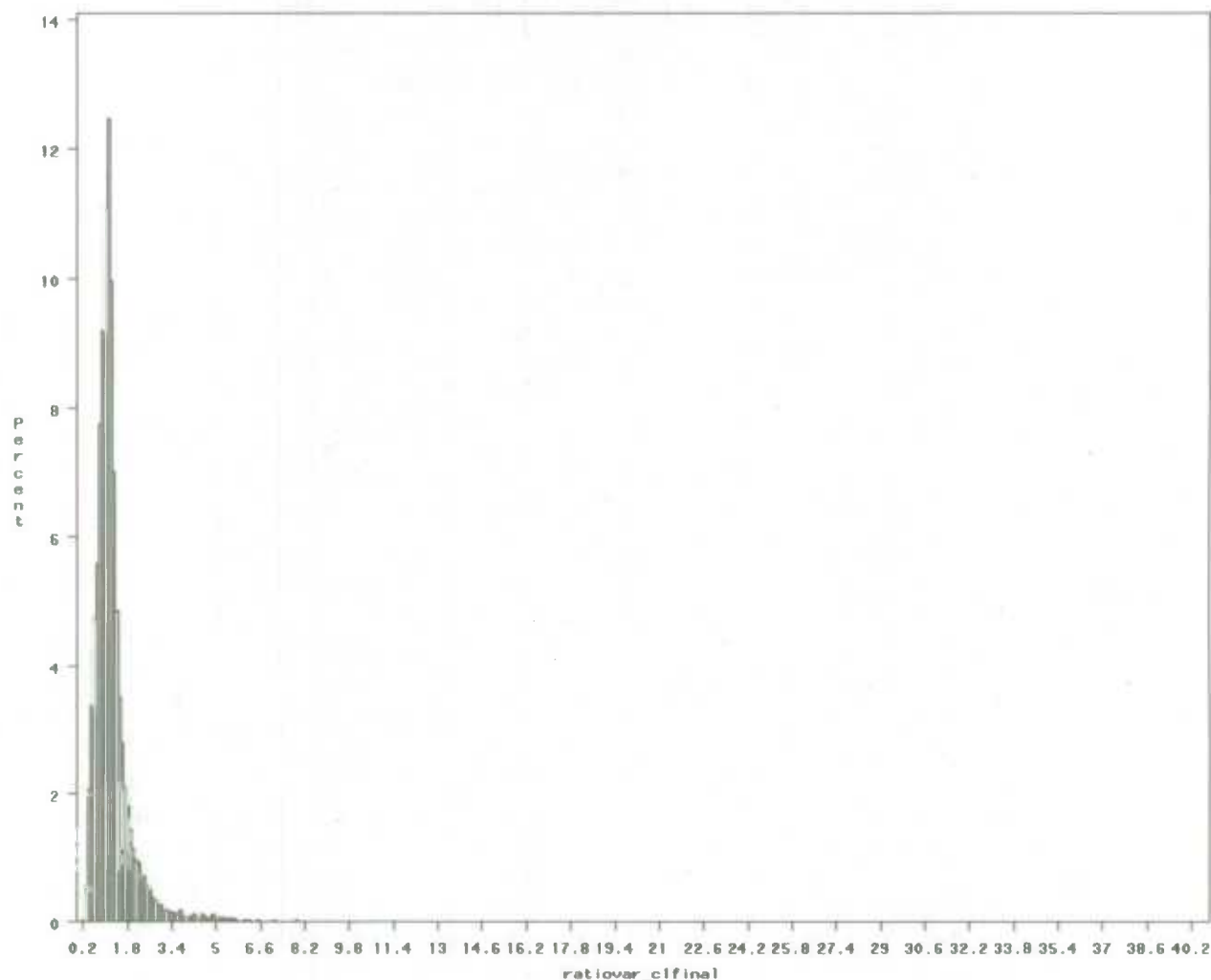


Figure B-5: Ratio of row-variances in the case of the final weights for Cycle 1 of the Original Cohort – full range.

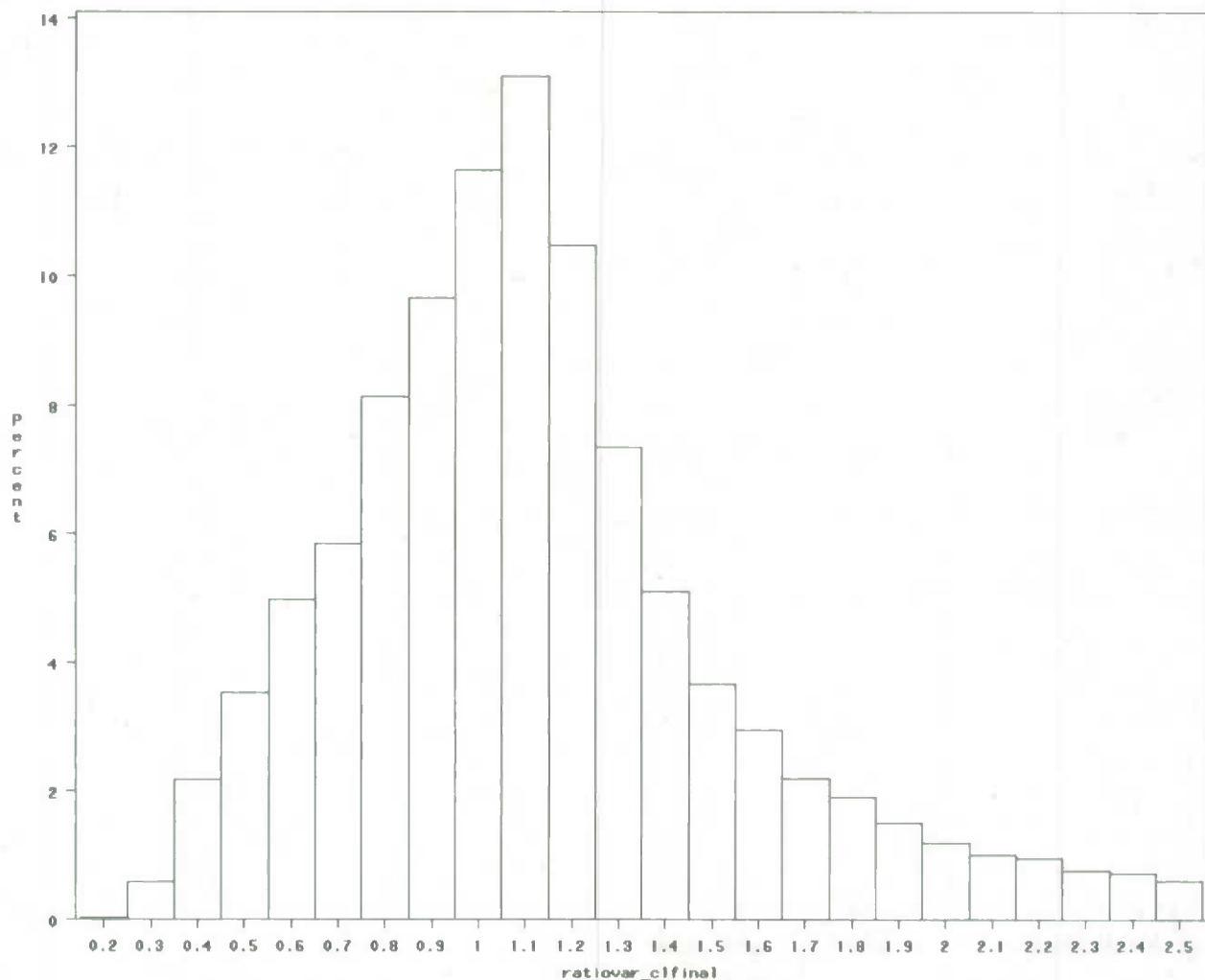


Figure B-6: Ratio of row-variances in the case of the final weights for Cycle 1 of the Original Cohort – limited range.

We describe in [Section 5.1.2](#) how this diagnostic applied to the Cycle 1 final bootstrap weights has come to explain a defect in NLSCY's implementation of the bootstrap at Cycle 1 of the Original Cohort which had remained an open question in [Girard \(2007\)](#).

Bootstrap: the computation

In practice, all the bootstrap steps outlined above are taken care of by the survey methodologist and not the end user. Actually, the latter is provided with a set of B final bootstrap weights ready for analytical use. What is left for the user is to make use of the bootstrap weights provided to compute the bootstrap variance estimate. Let us illustrate this in the context of the estimator of the mean. For a domain D , this estimator will have the following form for the sample:

$$\hat{\bar{y}} = \frac{\sum_{k \in I} y_k \times w_{k, final}}{\sum_{k \in I} w_{k, final}} \quad (B-5)$$

where $\{w_{k, final}\}$ is the set of final (or released) survey weights which include the nonresponse and post-stratification adjustments.

The user needs to compute the following B estimates:

$$\hat{\bar{y}}_b = \frac{\sum_{k \in I} y_k \times w_{k, b, final_bs}}{\sum_{k \in I} w_{k, b, final_bs}} \quad (B-6)$$

where $\{w_{k, b, final_bs}\}$ is the bootstrap counterpart of the set $\{w_{k, final}\}$.

The bootstrap variance estimate is then obtained as:

$$v_b(\hat{\bar{y}}) = \frac{1}{B} \sum_{b=1}^B (\hat{\bar{y}}_b - \hat{\bar{y}})^2 \quad (B-7)$$

(Simply replace $\hat{\bar{y}}_b$ and $\hat{\bar{y}}$ in Equation (B-7) with $\hat{\theta}_b$ and $\hat{\theta}$, respectively, to obtain the bootstrap variance estimate for any estimate $\hat{\theta}$.)

Computer-processing efficiently the B bootstrap estimates can be cumbersome and resource-intensive in some cases; this is why users usually rely on existing software like SUDAAN[®] to carry out these calculations instead of handling them themselves. With most such software, the bootstrap is available, but only indirectly: the user needs to specify BRR (which stands for Balanced Repeated Replication, a replication method used by some instead of the bootstrap or the jackknife) as the replication mechanism used to obtain the weights. While as a replication method the BRR is very different from the bootstrap, it remains that once the replication weights have (somehow) been obtained (BRR or bootstrap), the variance calculation process that follows is the same for both replication methods though it is referred to in software under the BRR label.

Bootstrap: does it capture all the variance introduced by nonresponse and post-stratification?

One attractive feature of the Rao-Wu bootstrap (which is not necessarily a feature of every bootstrap method used in a survey sampling setting) is its ability to account for *most* of the effect that the nonresponse and post-stratification strategies have on the sampling variance – most, but not all. This is essentially due to the fact that the Rao-Wu bootstrap has an explicit weight-based form, whereas other forms of bootstrap usually have not.

The idea here is to identify the nonresponse strategy based on RHGs as acknowledging that nonresponse is acting as if it were a second phase of sampling, one on top of the sampling which has yielded the sample that was drawn (the so-called first phase). It can be argued that this two-phase variance *for the mean* splits up into two additive components, only one of which the bootstrap captures adequately. A careful analysis of this decomposition of the variance for the mean reveals that the component not captured by the bootstrap is negligible compared to the other *provided* the first-phase sampling fraction is small. (And this is the case in the NLSCY.) Unfortunately, we do not know to what extent this is also true of estimators like the median for which such a decomposition is not known to exist. (Most people just assume that the same conclusion holds and thus use the bootstrap with the median, and other estimators for that matter, when there is nonresponse.) The same conclusion is reached as to the ability of the bootstrap to capture all of the variance arising from a two-stage design, since for variance estimation purposes the two-stage design falls within the same framework as the two-phase design. In other words, if the sampling fraction on PSUs is small then the bootstrap will capture the significant part of the multi-stage variance (for the mean, that is!). When the first-phase sampling fraction can not be considered small, then the component to the variance missed by the bootstrap usually is too big to ignore.

On the other hand, the case of post-stratification is straightforward since the contribution of post-stratification to variance is captured adequately by the bootstrap. The reader with general knowledge of RHGs and post-strata may be puzzled by these divergent findings about the ability of the bootstrap to capture completely or not the contribution to variance of nonresponse and post-stratification methodologies. After all, a given RHG and post-stratum have this in common: from the outset, the share of the sample each group will get is random in size.

The key feature that sets the two groups apart as far as variance estimation goes is this: *given* the sample, the effect of post-stratification is entirely determined, but the effect of nonresponse is not. In other words, if we are given the sample, then we have all the information we need to constitute the post-strata: once the sample is obtained, the post-stratification loses all its "apparent randomness". This is not the case with nonresponse usually, especially if modelled as a second phase of sampling: knowing the outcome of the first-phase is not enough information to determine entirely the outcome of the second-phase, which here is the nonresponse. It is basically this "residual randomness", given the sample, which is missed in our evaluation of the total variance using the bootstrap. (Though sad, this makes sense since the bootstrap "lives and breathes" in the realm of the selected sample, the first-phase of sampling in the grander picture which includes nonresponse. If the nonresponse mechanism was simple enough to be determined indeed entirely by the sampling – a very naïve and unlikely scenario-, then the bootstrap would capture all of the variance.) The general idea is this: as the sampling fraction gets larger, and thus brings the observed sample closer to a census, the error due to the initial sampling is marginally small compared to the error introduced by the nonresponse. And the way the Rao-Wu bootstrap is implemented, it fails to capture the error outside of what is introduced by the first phase of sampling: it does not "know" how to capture the additional error introduced by the subsequent sampling (the nonresponse really) from the information contained in the observed sample. There are forms of the bootstrap that were introduced recently specifically in an attempt to integrate into the variance estimate the sampling error introduced by both phases of sampling, but for the time being none appears truly satisfactory from a practical point of view.

Bootstrap: carefully analyzing the data

At this point an analyst can use the bootstrap weights provided to perform analyses and must be made aware of the pitfalls of analyzing bootstrap data for variance estimation purposes. As we have just seen in the previous section, bootstrap variance estimates are obtained in practice routinely through existing software (like SUDAAN[®], for instance). As with any statistical analysis of data, there is a danger of performing variance estimation purely mechanically, that is without a careful inspection of all the key steps. Indeed, while most analytical software have built-in (simplistic) rules to handle some common thorny issues, many important ones are not covered. This in itself would not be a problem if the software used allowed for a certain degree of interaction with the analyst and/or provided intermediate results to the analyst for close “independent” inspection. But most (if not all) software capable of variance estimation under the bootstrap are not built that way. Consequently, to obtain a variance estimate from them is to some degree (depending on how problematic the situation is) a leap-of-faith on the part of the analyst.

The point is this: software is no substitute for common sense on the analyst’s part which is needed for the statistical analyses performed to be of any value. The analyst must be critical of each and every one of the steps involved in the calculation of the variance estimates and be able to detect issues when they arise, because the analytical software cannot be trusted to handle these by themselves.

A good example is the following, taken from the NLSCY: suppose the analyst is interested in the count of male kids in PEI who were 2 years old back in 1994. Using Cycle 6 data, the analyst gets mechanically a point estimate of 981 (notice, it is an integer and no rounding was involved). If the analyst is not careful, a variance estimate of 962.361 will be reported with this estimate. The fact that the estimate 981 is an integer (and not some fractional number) means that the domain of interest is actually a post-stratum or a direct sum of post-strata: the estimate *has* to be a demographic projection based on Census data or a sum of such projections. (A look at the NLSCY’s documentation confirms that this domain is a post-stratum.) Consequently, the variance estimate reported with the estimate of 981 ought to have been zero, not 962.361. Indeed, since the domain is a post-stratum, a domain for which we actually know the real count, there is no sampling variability associated with that estimate: sample after sample, our estimate would always be the same (under the same methodology, of course) – that one projection total.

To see why the analyst got this variance estimate let us take a closer look at the bootstrap estimates for that domain. The following snapshot depicts for that domain (see oval) a critical part of the 1,000 bootstrap estimates, those from 720 to 727, along with the number of respondents in it (the variable “count”).

	Post-stratum	count	bsw720	bsw721	bsw722	bsw723	bsw724	bsw725	bsw726	bsw727
49	02F8	56	19502	19502	19502	19502	2E4	19502	19502	19502
50	02F9	42	23140	23140	23140	23140	23E3	23140	23140	23140
51	02M0	23	3405	3405	3405	3405	3405	3405	3405	3405
52	02M1	10	981	981	981	981	0	981	981	981
53	02M2	35	5906	5906	5906	5906	5906	5906	5906	5906
54	02M3	32	4756	4756	4756	4756	4756	4756	4756	4756
55	02M4	112	48377	48377	48377	48377	48E3	48377	48377	48377
56	02M5	146	77346	77346	77346	77346	77E3	77346	77346	77346
57	02M6	41	7715	7715	7715	7715	7715	7715	7715	7715
58	02M7	48	7149	7149	7149	7149	7149	7149	7149	7149
59	02M8	54	20566	20566	20566	20566	21E3	20566	20566	20566
60	02M9	46	24221	24221	24221	24221	24E3	24221	24221	24221
61	03F0	17	3407	3407	3407	3407	3407	3407	3407	3407
62	03F1	12	941	941	941	941	941	941	941	941
63	03F2	42	5828	5828	5828	5828	5828	5828	5828	5828

One bootstrap estimate stands out: the one associated with replicate 724 (see arrow). That estimate is zero because none of the 10 units (see the variable “count”) from that post-stratum was selected in bootstrap replicate 724. (Remember, we have argued above that only about 60% of the units in a domain are part of a bootstrap replicate. Thus, for replicate 724 here, all 10 units were in the 40% of the units that were not selected.)

If the variance estimate is computed directly without analyzing the data carefully, then we get as claimed:

$$v_{bs} = \frac{1}{1,000} \sum_{b=1}^B (estb - 981)^2 = \frac{1}{1,000} \left[(981 - 981)^2 + \dots + \underbrace{(0 - 981)^2}_{b=724} + \dots + (981 - 981)^2 \right] = \frac{981^2}{1,000} = 962.361$$

This is the variance estimate the analyst will get if SUDAAN[®] or BootVar is used. To get the variance estimate of zero the analyst is entitled to here, the analyst needs to discard the information provided by replicate 724 and just rely on the other 999 “pieces” of bootstrap information:

$$v_{bs} = \frac{1}{999} \sum_{\substack{b=1 \\ b \neq 724}}^B (estb - 981)^2 = \frac{1}{999} \sum_{\substack{b=1 \\ b \neq 724}}^B (981 - 981)^2 = 0$$

The issue here is not solely about replicates with an estimate of 0; it is bigger than that. The issue arises whenever there is among the replicate estimates at least one outlying value (which may be zero or not). To prove this point, consider the province of PEI as the domain which contains as a sub-domain the problematic post-stratum “02M1”. A critical portion of the replicate-based estimates is shown in the snapshot below. Observe that for all replicates but 724, the replicate-based estimate for the whole of PEI is the demographic projection used in post-stratification, which is 23,148, whereas for replicate 724 it stands at 22,167, which is 981 units short of it. An estimate

of 22,167 is obtained for replicate 724 because the sub-domain "02M1" contributed zero to the weighted sum instead of the 981 it contributed to all the other replicates.

VIEWTABLE: Work.Province_totals_production

	province	Final, for release, long. weight after all adjustments, incl. NR & post-stratification	est720	est721	est722	est723	est724	est725	est726	est727
1	1	88996	88996	88996	88996	88996	88996	88996	88996	88996
2	1	23148	23148	23148	23148	23148	22167	23148	23148	23148
3	2	144088	144088	144088	144088	144088	144088	144088	144088	144088
4	3	115131	115131	115131	115131	115131	115131	115131	115131	115131
5	4	1090582	1090582	1090582	1090582	1090582	1090582	1090582	1090582	1090582
6	5	1773616	1773616	1773616	1773616	1773616	1773616	1773616	1773616	1773616
7	6	182869	182869	182869	182869	182869	182869	182869	182869	182869
8	7	173611	173611	173611	173611	173611	173611	173611	173611	173611
9	8	489913	489913	489913	489913	489913	489913	489913	489913	489913
10	9	576125	576125	576125	576125	576125	576125	576125	576125	576125

One must not conclude from this that replicate 724 is bound to be discarded for anything that regards PEI. To prove this point we have created a *random* characteristic among the children in PEI (i.e., a domain which does not coincide with any of the post-strata) and got the following snapshot for the critical area in the bootstrap file around replicate 724:

VIEWTABLE: Work.Domain_totals_production

	province	Final, for release, long. weight after all adjustments, incl. NR & post-stratification	est720	est721	est722	est723	est724	est725	est726	est727
1	0	24438.144675	26934.038125	23945.320916	26063.606592	22820.578712	23523.287525	24393.29323	22799.71432	24970.677126
2	1	5857.1821518	5949.9430062	4757.8714563	5794.774445	5453.2312575	5764.8392687	6587.2658882	5150.916155	6408.033345
3	2	41563.078301	43344.640961	36710.232969	38633.899185	38453.310067	39532.749513	42315.706516	40110.513012	40224.967941
4	3	33983.208274	33052.359951	36160.351675	34654.08966	33339.242463	38677.012279	35180.968761	36688.436563	31903.116843
5	4	329061.48019	323386.83427	347076.42786	335977.94483	311544.05369	323117.8566	303778.78255	325862.6897	353098.6913
6	5	533410.94015	555828.44175	583748.70942	571196.51421	504209.40482	547247.92612	524719.64123	534318.12159	552026.38875
7	6	55502.035416	56011.967778	55987.516617	53610.431829	54143.506022	61478.299246	56920.797548	56538.432287	54195.039331
8	7	49489.575045	52239.608468	48037.885714	47836.162291	48246.324391	47306.885506	48046.642173	53380.293241	50759.678758
9	8	133621.2117	139879.56887	128321.46029	134642.39042	138713.16152	150217.48898	145568.47577	134179.18995	146793.09915
10	9	194964.82752	202022.21997	188142.42175	179709.68062	183919.69426	192730.61579	208591.64548	180145.29898	188491.58023

Contrarily to what we have seen in the two previous cases, the bootstrap estimate obtained from replicate 724 for this domain no longer stands out as a sore thumb. Indeed, smaller and larger values can be found among the eight bootstrap estimates displayed here, not to mention the entire spectrum of 1,000 bootstrap estimates (not shown). Of course, it would remain to be seen whether there are replicates which yield estimates for the size of this domain which are clearly outlying; but one thing is for sure: replicate 724 will not be a part of them.

The analyst thus has to be careful analyzing bootstrap data and cannot entrust the software used to *always* make the appropriate inference decisions on the analyst's behalf: software is programmed to handle basic cases only.

Variance estimation: stability over small domains

All things considered it is quite easy to obtain variance estimates in practice using the bootstrap, to a point where it brings issues of its own. Indeed, variance estimates are obtained using the bootstrap in an amazing array of situations, which often go beyond what the supporting theory covers, and when the theory does cover it, the usage made is pushed to the extreme (*e.g.*, case of the mean but for tiny domains). Therefore, despite its ease of use, the analyst must learn to be critical of the situations where it is actually used, and make use of good judgment to set limits: if a result appears to be too reliable (*i.e.*, comes with a small variance estimate) to be true given the domain size, then it probably is!

The case of tiny domains is certainly the most frequent issue which presents itself in practice with respect to variance estimation. With a design-based approach, the user needs to realize that the variance estimation method uses strictly sample information as its input – nothing comes from the outside, such as models (model as in a vehicle for external-to-sampling-design information). Roughly speaking, the smaller the sample size, the less there is of the much-needed information for the method to turn into an appropriate assessment of the variance. (While the sample size is an easy way to summarize the information content of the sample, it is a crude assessment which ignores other important factors like the covariance inner-structure of the sample: the analyst should not gauge the worth of a sample solely based on its size.) This is thus an issue the analyst will face with *any* design-based variance estimation method. On top of that with the bootstrap the analyst must deal with the fact already stated earlier that a given replicate is made of only about 60% of the sample. So a tiny sample to begin with results in an even tinier replicate. If a domain is too small, then it may not have representing units in one or several replicates, thus reducing the effective number of replicates B actually available. In other words, in such cases, the bootstrap variance estimation method becomes very unstable, to a point where it may be altogether unreliable.

In [Girard \(2007\)](#), an asymptotic result on the variance of the variance for a mean under SRSWR is quoted (from [Dudewicz and Mishra \(1988\)](#)) and used to evaluate the stability of the variance estimation process. This result was used because the exact result provided in [Ardilly and Tillé \(2003\)](#) was believed to hold only for a normally distributed variable of interest when actually its scope is quite general:

Result (Exercise 2.21 of Ardilly and Tillé (2003))

If Y is a variable of interest with mean \bar{Y} finite variance σ^2 and kurtosis κ , then under SRSWOR the variance of the variance estimator of the estimator of the mean is

$$V(\hat{V}(\hat{Y})) = \left(1 - \frac{n}{N}\right)^2 \frac{1}{n^2} \left[\frac{N(N-n)}{n(n-1)(N-1)^2(N-2)(N-3)} \right] \times \\ \sigma^4 (\kappa(N-1)(N(n-1) - (n+1)) - N^2(n-3) + 6N - 3(n+1)) \quad (\text{B-8})$$

Let us use this result (which is valid for all sample sizes) directly to the following case which is of importance in practice in social surveys: the estimation of a proportion. In other words, let us assume that Y is binomially distributed with success p .

From statistical handbooks (or www.wikipedia.com) we get in this case that:

$$\begin{aligned}\sigma^2 &= p(1-p) \\ \kappa &= \frac{3p^2 - 3p + 1}{p(1-p)}\end{aligned}\tag{B-9}$$

(Observe that both expressions remain the same if we make the change $p \mapsto 1-p$.)

Table B-1 lists a few scenarios of sample and population sizes, along with the corresponding Coefficient of Variation (CV) of the variance of the variance estimator which is:

$$CV = \frac{\sqrt{V(\hat{V}(\hat{Y}))}}{V(\hat{Y})}\tag{B-10}$$

(It makes sense to use the CV even though we work with proportions since both the numerator and denominator are invariant under the change $p \mapsto 1-p$.)

N	n	P (in %)	CV (in %)
10,000	1,000	2.5	18.26
		5	12.39
		10	8.00
	100	2.5	60.57
		5	41.12
		10	26.57
1,000	100	2.5	57.83
		5	39.26
		10	25.37

Table B-1: Stability of the variance estimate for proportions under SRSWOR.

As we can see from Table B-1, the variance estimator is very unstable for extreme proportions (big or small). This puts a damper on anyone's ambitions to obtain statistically valid results in the case of small sample sizes.

Note: all of this can be obtained numerically through simulations as well – see [Girard \(2007\)](#). In fact, simulations offer the opportunity to cover more complicated cases like that of the median under the same design, or other basic designs.

The difficulty though with this result is its restricted applicability: it concerns only the simplest design and estimator. What can be done in the more realistic cases that occur in practice? One

option is to use synthetic variables in conjunction with the bootstrap. The idea is as follows. For a (real) characteristic of interest which has led to a point estimate of \hat{p} , generate in the bootstrap file of respondents a set of, say, ten (or more, if one is patient enough to wait for the computations to be made!) random Bernoulli variables with probability of success p . For each of these synthetic variables, obtain the corresponding bootstrap variance estimate and then compute their empirical variance. (The variables are called "synthetic" because unlike a real characteristic they do not exhibit any correlation with the weights: they were generated independently of them.) This gives in practice a good idea of the stability of the variance estimate, indicating whether it can be trusted or not.

Variance estimation: confidence intervals for small proportions

Suppose an analyst is in the following situation: a sample under SRS, say, was obtained which has led to estimates of about 4% and 19% for two distinct characteristics of interest (which are held by 5% and 20% of the units in the population, respectively). The standard errors of these two estimates are used to obtain the following 95% confidence intervals in the usual way (*i.e.*, assuming normality of the underlying sampling distribution):

(-0.2%, 8.2%)

and

(11.2%, 26.8%)

The first interval, the one about 4%, is clearly suspect since it extends into the negative numbers when actually a proportion is always non-negative. It is tempting to rid the interval of that inconsistency by capping off the lower bound to zero (*i.e.*, reporting instead a confidence interval of (0, 8.2%)), but the issue with it is more serious than that.

The two intervals were constructed assuming the normality of the underlying probability distribution on the estimates. In this simulation setting we have obtained an approximation of that sampling distribution for both estimators by simply sampling *ad nauseam*. (In other words, we carried out a Monte Carlo simulation.) We have obtained the following distribution for the proportion of 20% - the continuous line is the best-fit-to-the-histogram normal curve:

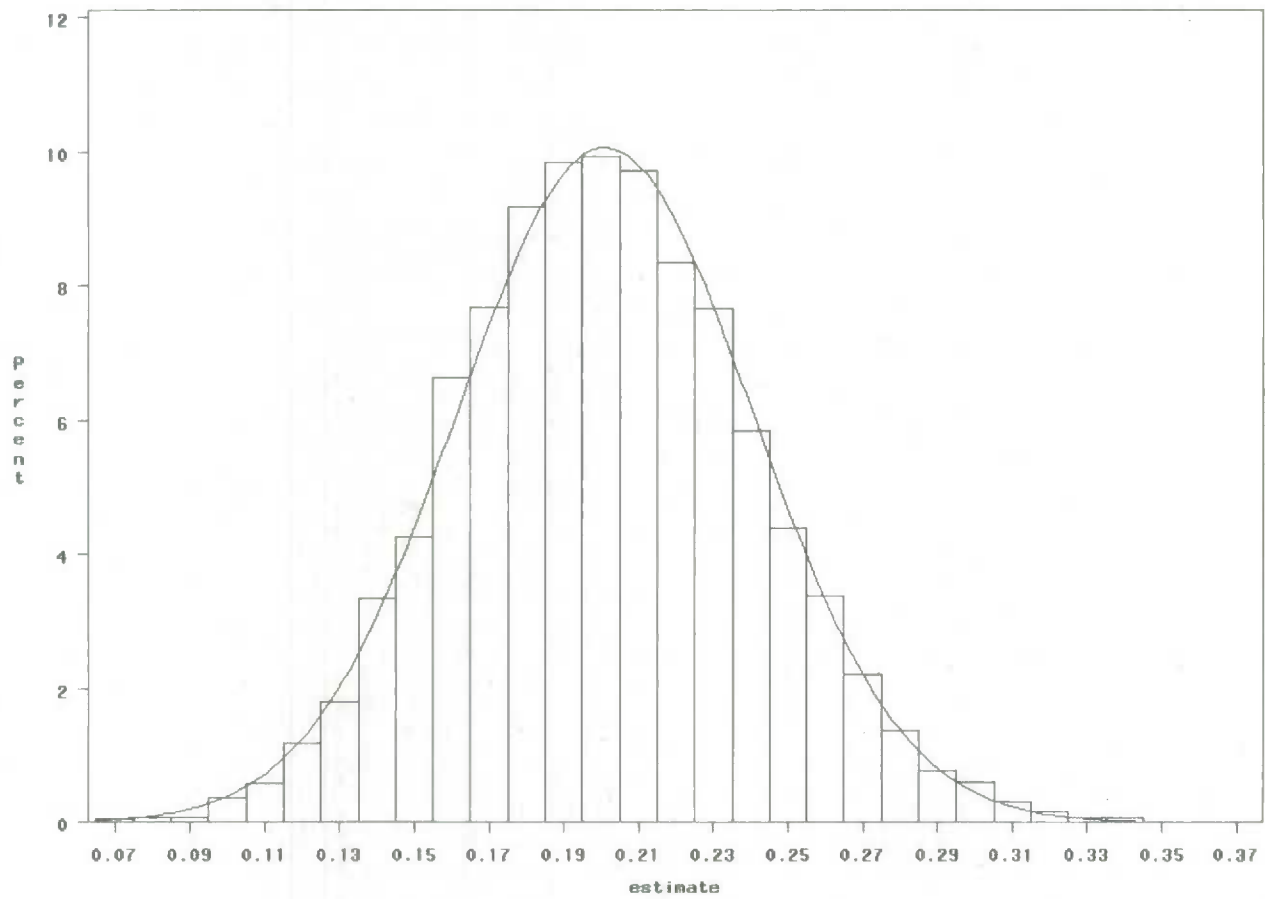


Figure B-7: Sampling distribution of the estimator of a proportion of 20% under SRSWOR with best normal fit

Observe just how good the overall fit of the continuous line with the histogram is in this case. Observe that while the sampling distribution is by definition discrete, the (continuous) normal distribution can still be used as a fit in this case because its tails vanish for all practical purposes at the extremities of the range of the histogram. Now let us turn our attention to the proportion of 5%, below:

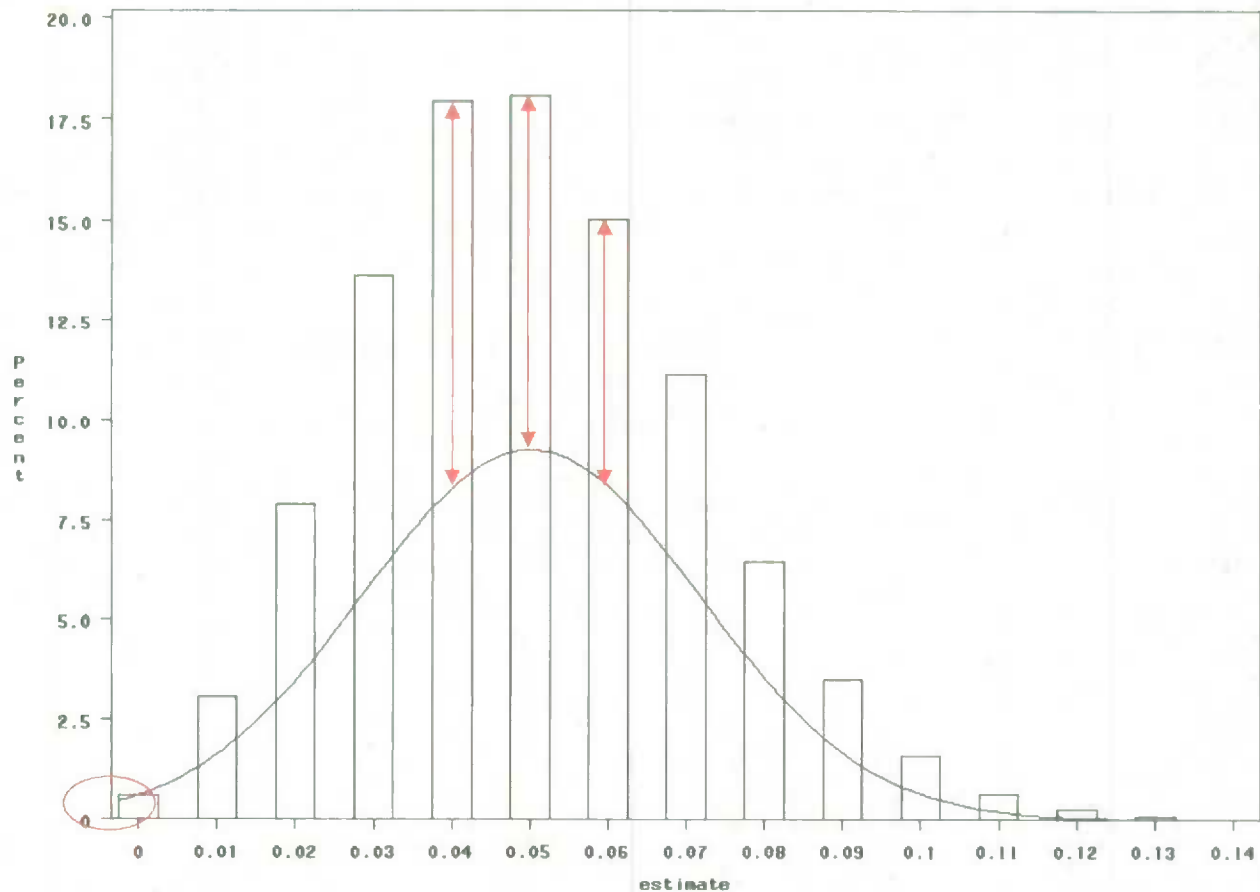


Figure B-8: Sampling distribution of the estimator of a proportion of 5% under SRSWOR with best normal fit

The lack of fit is obvious here, as pointed out by the various arrows and the oval near the origin. The latter explains why we first obtained a negative lower bound on the confidence interval: the best-fitted normal curve extends here a bit in the negative numbers, beyond the range of the histogram, *before* becoming “zero for all practical purposes”. The direct consequence is that the coverage claimed with the use of this methodology, which is provided by the normal distribution, is inaccurate.

This clearly shows that normal-based confidence intervals should not to be built blindly, especially in the case of bounded quantities like proportions. Indeed, small proportions may not lead to an estimator which is normal-looking. Some may object to this entire line of argument in practice because in our context of a simulation we were able to approximate the true sampling distribution by repeatedly sampling the population, something we cannot obviously do usually. Therefore, they

claim, there is no way for us in practice to draw conclusions about the non-normality of the sampling distribution. Not quite. Recall: the bootstrap distribution is a suitable approximation of the true sampling distribution! So, by plotting the histogram of the bootstrap estimates of a proportion, say, the analyst can tell if the normal assumption on the sampling distribution is reasonably met. The problem, in practice, is that analysts rely on software to compute their variance estimates rather than hard-coding it and none of them provides the histogram of the bootstrap estimates as a side-product, nor gives access to the replicates themselves for the analyst to subsequently plot them on the side.

Bootstrap and normalized (or standardized) weights

It is a fact that many analysts resort to normalized weights (a.k.a. standardized weights) to obtain variance estimates through their model-based software. Though disparaged by many survey statisticians, this practice and the context of its use have to be better understood by everyone involved. We feel it is important that analysts understand the limitations of this approach and that survey statisticians realize why it is used despite its limitations, since many analysts do rely on normalized weights to carry out their inferences. We present here the general idea. (Our account follows [Lafortune \(2008\)](#)).

Normalized weights are essentially an attempt to introduce the concept of survey weights into the model-based framework which underlies typical statistical software. It is a step toward achieving a proper design-based inference but in some instances it falls considerably short of doing so. A normalized set of weights $\{w_k^{norm}\}$ is obtained from a set of survey weights $\{w_k\}$ with average weight \bar{w} the following way:

$$w_k^{norm} = \frac{w_k}{\bar{w}} \quad (B-11)$$

The set of survey weights $\{w_k\}$ cannot be used directly in a model-based software because the weights will be interpreted as frequency weights. And since the sum of the “frequency weights” provided often enters the calculation of the variance in the form of an available “sample size”, the variance estimates obtained in such a manner are grossly under-estimating the true (design-based) variance. The normalized weights are an attempt to bring the available “sample size” computed that way to a more appropriate order of magnitude. Actually, the sum of the normalized weights yields the sample size.

A common fallacy with normalized weights is to think that their use acknowledges the features of the design needed for proper (design-based) variance estimation. Though traces of it are certainly found in the survey weights, the structure of the sampling design is only depicted by the bootstrap weights. The idea is that *one* set of weights does not reveal the whole structure, the inter-dependencies; to fully reveal the structure, one needs many, many samples at once (which we never have) or, the second-best, many replicates taken from the one observed sample. Practically, this means that significance levels reported about statistical conclusions obtained through normalized weights are not necessarily valid since some components to the overall variance are not taken into account appropriately.

Though the normalized weights are inadequate substitutes to the bootstrap weights, they are useful to have when the bootstrap weights cannot be used and variance estimates are nonetheless needed. This is the case actually of some analyses using advanced modelling (*e.g.*, hierarchical modelling). In such instances the analyst needs to be extra careful about how best to convey in statistical terms the scope of the conclusions reached.

Perspective C: Dissemination procedure including the RDC network and the ATT

The main preoccupation with releasing data files for the NLSCY is the protection of the confidential information they contain about the respondents to the survey. This is why there are only a few ways for an analyst to access the data, each involving a tight control process. There is the basic request, remote access and the Research Data Centres (RDC) network.

With the basic request approach, the analyst describes to a contact employee of Statistics Canada the request to be performed on the data. This can be done over the phone, but usually by email, and is done on a cost-recovery basis. For example, what proportion of mothers breast-fed their 1-year-old in 1994 compared to now? The employee will first investigate the possibility of carrying out the analysis (for instance, does the NLSCY have the information needed?) and then cost the request. In practice, the employee may need to carry out the analysis in part or in whole to actually answer this question fully. Consequently, resources may be spent with no end results since an analysis proposal which is deemed impossible to be carried out usually is free of charge. If the employee gets the permission to proceed with the analysis, results are subsequently obtained and vetted for disclosure risks. It is only when the results are found to comply with the Agency's strict rules on the disclosure of information that they are released to the analyst. Depending on the complexity of the analysis, a request may take anywhere between an hour and a few working days. This way of proceeding is mainly used by analysts with very basic requests (like tabulations) or by those who are not in a position to use the other two channels described below.

There is another avenue for the analyst with good programming skills (or at least access to help on that front) to obtain the desired results: remote access. This approach calls for the analyst to submit to a contact employee associated with the NLSCY dissemination team a fully-working computer program (usually written in SAS) which carries out the intended analysis on the data. In order to get a program to a working level, the analyst has access to NLSCY's synthetic files which are used for testing and debugging purposes. The synthetic files exhibit essentially the same structure as the released files they are associated with but the data they contain have been altered beyond recognition, even more than they are for a Public-Use Microdata File (PUMF). For example, only a portion of the records of the released file are used in a synthetic file and the data associated with each record is not entirely its own. Once that is done, the program is transmitted to Statistics Canada where it is run on real data. The various outputs of the program are scrutinized to ensure they represent no risk of disclosure of confidential information. If such a risk is believed to exist, then the outputs are altered (*e.g.*, through the suppression of categories originally asked for by the analyst) to close the gap *before* the outputs are handed back to the analyst. Aside from the time and resources spent by the analyst to program the request and the computer-resources by Statistics Canada to process it, this procedure requires greater care with respect to disclosure risks than the basic requests above. Indeed, the requests sent through remote access usually are much more complex in nature than the basic requests carried out over a few phone calls or emails, and consequently the assessment of potential breaches in the (complex) outputs is not as straightforward as with, say, a basic tabulation.

The third option for an analyst is to visit one of the RDCs. There are more than 15 RDCs located in 20 universities across Canada. To access the microdata housed in a RDC, an analyst must first submit a project proposal. Once the project has been approved, a number of security procedures

must be followed. For example, Statistics Canada performs an Enhanced Reliability Check and the analyst must take The Oath or Affirmation of Office and Secrecy. This means that the analyst is sworn-in: he/she is lawfully bound by the Oath not to reveal any information he/she may come across through his/her research activities. The analyst consequently has the same responsibilities with regard to the protection of confidential information as any Statistics Canada employee. Even after taking the Oath, the analyst's work is still under the close scrutiny of the RDC employee who acts as an extra layer of protection to minimize further risks of accidental, or unintentional, disclosure of confidential information (For more information about the RDC program see, <http://dissemination.statcan.ca/english/rdc/network.htm>.)

The basic issue with the RDC is that both the data and the analyst must physically be moved from their "preferred" spot. Indeed, the analyst would greatly prefer to remain in his/her office at all times rather than have to visit an RDC (one may not be close by at all) and Statistics Canada would prefer that the data remain in its head office, under lock and key. For the analyst this may prove to be so great a burden as to avoid the RDC option altogether.

Considering the array of options available to the analyst, we came up with an innovation, that of an on-line tabulation tool to facilitate access to NLSCY's data (and ultimately to other surveys' data). The idea of this tool is to address the most glaring issues the basic analyst is faced with when trying to access the data. The typical analyst we had in mind was a program manager in some Canadian organization who needs a basic set of basic tabulations, quickly. While it is to be expected that complex statistical analyses require both time and resources to be performed, it is unfortunate that basic requests cannot be met in a very timely manner under the current framework. Indeed, no matter what option an analyst chooses to get his/her tabulations done, it is most likely that several hours, if not days, will elapse before he/she can receive the sought-after results.

The on-line tool, when available, would allow an analyst to specify from his/her office and through a series of capture screens the information needed to describe the tabulations needed. The information would be sent to a Statistics Canada server through a secured connection to be processed *centrally*. In other words, the data and the analyst, contrarily to the RDC option, would not need to "physically meet half-way". Since the analyses involved are basic tabulations, the tool would have the capacity to analyze the request, and later the results, for disclosure purposes. It is only when the outputs required meet the stringent rules implemented by the tool that the results would be sent back to the analyst, through the same channel already used. We envision a basic tabulation request to be processed in only a matter of minutes. Such a tool is attractive to many analysts since it would be easy of access and use, it would handle the methodology adequately on its own (*e.g.*, would use the bootstrap weights) and it would send back the results vetted for disclosure risks in a timely manner.

At this point and time of its development, the on-line tool is in the form of a prototype called the Automated Tabulation Tool (ATT) which is used to assess the functionalities such a tool would need to have to meet the needs for tabulations of our typical analyst.

Another initiative we introduced is to hold workshops which help analysts bridge the gap there is between analytical needs and the practical and adequate use of survey methods to longitudinal surveys, and beyond ([see Lafortune \(2008\)](#)).

To give the flavour of what is typically discussed in the workshops, here are the methodological issues covered in the workshop held for RDC analysts in Toronto in the fall of 2008 in the day and a half it lasted:

- populations of interest: longitudinal and cross-sectional,
- sampling,
- survey estimates and survey weights (longitudinal – funnel versus non-funnel – and cross-sectional),
- sampling error: variance, standard error (versus standard deviation), CVs and confidence intervals,
- variance estimation using bootstrap weights,
- issues with small area estimation (*i.e.*, when the sample size for a domain of estimation is too small),
- nonresponse (why it is an issue for analysts and techniques to deal with it – ignore it, report it as a category, profile the nonrespondents, calculate a weight adjustment, impute),
- normalized or standardized weights and why they are an incomplete application of the design-based approach,
- pooling data (combining multiple cohorts),
- a real example of an analysis dissected to illustrate all the previous points.

Perspective D: Unequal weights in a survey like the NLSCY

As discussed in [Perspective A](#), a given set of weights about a cohort at any given cycle of the NLSCY, be they cross-sectional, longitudinal funnel or longitudinal non-funnel, is multi-purpose: the adjustments put into the weights are not the best for any given analysis but they are meant to perform well overall.

The issue of using a set of multi-purpose weights to handle various analyses at once, instead of distinct sets of tailor-made weights, becomes truly problematic when the weights are unequal. Indeed, if the NLSCY had a self-weighted design, then by definition (at least before nonresponse and post-stratification issues arise) the weights would provide the user with the same performance. When the weights are (approximately) equal, there is a symmetry that prevents, to a significant extent, any extreme estimate from arising: the symmetry is very forgiving that way. To illustrate the point, consider a domain whose size is to be estimated using a sample drawn from a SRSWOR design. Given the sample, for the ensuing Horvitz-Thompson estimate to miserably fail for the domain under consideration, we would need some extreme phenomenon to happen: there are either too few or too many units from the domain in the sample than expected. Such a “bad sample” can certainly happen, and it is a possibility that the survey sampler has to live with, but it is quite rare. Consider now the alternative of a sample drawn from some other design leading to largely unequal weights. In this scenario, in addition to the possibility of a bad composition of the sample for the domain, there is the issue of which units get the extreme weights, big or small. For instance, is there a pattern in the sample such as all units from the domain happen to have the larger weights? Or the smaller weights? If the weights are not “well-balanced” between units inside and outside of the domain, then even if the sample is a pretty good one, things can go awry pretty quickly.

The lack of symmetry brought on by the unequal weights amplifies any odd behaviour of the sample. Indeed, any “hiccup” in a characteristic will be amplified if it occurs in conjunction with a pattern in the weights. That interaction between “odd values” of the characteristic of interest and unequal weights hurts when one is working with only one set of multi-purpose weights for the whole survey; such an interaction is difficult (if not straight out impossible) to compensate for *a priori* when building the weights, which generally happens *before* the interaction can even be noticed. If we could produce a set of tailor-made weights for any analysis then we would have the leisure of observing a potential interaction between a characteristic and various features such as the design weight, the nonresponse, *etc.* *before* committing ourselves to a *final* set of weights.

As we saw in [Perspective A](#), the NLSCY is based on the LFS, hence the unequal weights. In addition, the NLSCY introduces other adjustments which tend to increase the heterogeneity of the weights.

Therefore, with unequal weights, weird estimates are more likely to be observed than had the weights been (nearly) all the same. Consequently, the *true* variance associated with an estimate (usually) is larger under unequal weights than it would be under a self-weighted design of comparable size (*i.e.*, the design effect is greater than 1). The fallacy in practice with unequal weights is to think that the variance estimation always is so stable that whatever statement made using the true variance holds also for all of its estimates. But it is a fact that unequal weights also affect the whole variance estimation process itself, rendering it quite unstable: in the presence of

unequal weights the variance estimator is subject to non-negligible variance of its own. Consequently, a given variance estimate may differ significantly from the true variance it is estimating, contrary to common belief. What does this mean? It means that in practice we can only associate a variance estimate, not the true variance, to a given point estimate. Usually, that extra layer of uncertainty is dismissed because variance estimation is perceived as very stable: the variance estimate is deemed to be so good as to be the true variance itself. That may very well be the case for self-weighted designs and well-behaved estimators, but for every other scenario there is a reason to be concerned with the reliability of the variance estimates themselves.

That being said, it is not at all easy to quantify the variability that exists in the variance estimates, especially under a complex design. It is indeed known that the variance of the variance of the estimator of the mean is difficult to evaluate except in the simplest of situations (*i.e.*, a SRSWOR design). But when we are able to we find that it is stable enough for all practical purposes provided the sample size is not too small (*i.e.*, the square root of the variance of the variance is deemed small compared to the variance estimate itself). Unfortunately, that the variance estimator fares rather well under SRSWOR does not mean this nice behaviour extends to more complex designs, since the symmetry provided by the equal weights contributes a lot to the stability observed.

To drive home the point, here is an interesting parallel that needs to be taken with a grain of salt as to its real meaning in a complex survey setting:

Claim

If X_1, \dots, X_M are M i.i.d. variables with mean μ , finite variance σ^2 , kurtosis α_4 and w_1, \dots, w_M are given constants, then the following holds about the estimator

$$\begin{aligned}\tilde{X} &= \frac{\sum_{i=1}^M w_i X_i}{\sum_{i=1}^M w_i} & (D-1) \\ V(\tilde{X}) &= \frac{\sum_{i=1}^M w_i^2}{\left(\sum_{i=1}^M w_i\right)^2} \sigma^2 \\ \hat{V}(\tilde{X}) &= \frac{\sum_{i=1}^M w_i^2}{\left(\sum_{i=1}^M w_i\right)^2} S_M^2 \\ V(\hat{V}(\tilde{X})) &= \frac{\left(\sum_{i=1}^M w_i^2\right)^2}{\left(\sum_{i=1}^M w_i\right)^4} \frac{(\alpha_4 - 1)\sigma^4}{M}\end{aligned}$$

Corollary to the claim: Case of constants all equal to 1 and normally distributed variables

In the case where $w_i = 1$ for all i and variables are normally distributed, we get the known result for $\tilde{X} = \bar{X}$, the sample mean:

$$V(\hat{V}(\bar{X})) = \frac{2\sigma^4}{M^3}$$

since $\alpha_4 - 1 = 2$ for normal variables (whose kurtosis $\alpha_4 = 3$).

Note: Usually this corollary's result is obtained as a consequence of the variance statistic being chi-square distributed.

A more telling expression for the variance of the variance estimator is:

$$V(\hat{V}(\tilde{X})) = \left[1 + \frac{\sum_{i=1}^M \left(\frac{w_i - \bar{w}}{\bar{w}} \right)^2}{M} \right]^2 \frac{(\alpha_4 - 1)\sigma^4}{M^3} \quad (D-2)$$

Note: [Graubard \(2002\)](#) used this type of reasoning (*i.e.*, using weights as constants) to obtain a “design effect for weights” by considering the ratio of $\hat{V}(\tilde{X})$ to that of the variance without the weights. Therefore, it serves as a measure of the inefficiency, variance-wise, of the survey weights used in a given inferential context. We went one step further because our interest mainly lay in characterizing the stability of the variance as a function of the weights (*i.e.*, the variance of the variance).

It is clear from (D-2) that the *variance* of the variance estimator is smallest when all the constants w_i are equal to their average \bar{w} (*i.e.*, they are all equal since in that case the squared factor is 1), and that any significant departure from that one scenario only contributes to the instability of the variance estimate.

To illustrate what it can reveal in practice, we applied this result to the NLSCY weights. For Cycle 1 of the Original Cohort, the longitudinal weights led to a squared bracket factor value of over six. For Cycle 6 of the Original Cohort we calculated that measure at each of the weighting steps, from the LFS-inherited weight to the NLSCY final weight. Doing so we got a value of 4.32 for the LFS inherited weight, 5.37 for the LFS weight adjusted for eligible kids, 6.84 after nonresponse and 7.11 for the final weight. So, as to be expected, each new adjustment adds a new layer of heterogeneity to the weights.

Again, a direct interpretation of that value as a meaningful measure of instability of the variance estimate is very limited. What this precisely means is this: if the constants are taken to be the weights of the sampled units in the Original Cohort NLSCY, then the estimator built in (D-1) would have a variance for its variance estimator over six times that had the constants all been equal. Consequently, the "variance" figures provided above involved here have nothing to do with the complex design variance of the NLSCY. Indeed, none of the covariance present in a complex survey is taken into account here. The sole feature that is somewhat represented here is the weights, but only through sampled X s' independent constants. It is mainly provided here as a way (one among others) to show just how the weights are indeed unequal in the NLSCY and that their impact on variance stability is non-negligible. But still it nonetheless shows the impact unequal weights can have on variance stability. To a lesser extent, it also reminds us that the smaller the sample size, the greater the instability of the variance estimate.

Furthermore, there is yet another issue to be discussed here with respect to variance estimation. Not only have we argued already that unequal weights render the variance estimator unstable, but they may also jeopardize the applicability of the bootstrap method to obtain an (approximation of the) estimate for the variance in practice. Indeed, [Rust and Rao \(1996\)](#) make it clear on page 294 (see beginning of their Section 2.5) that it is difficult to study the properties of replication methods theoretically. And when this was done (at least at the time of writing their paper), one of the assumptions was that no weight was disproportionately large.

In a longitudinal setting unequal weights that are very disproportionate can create inconsistencies in longitudinal variance estimates. Consider the units having a given characteristic of interest for which we got an estimate at Cycle 1 along with a variance estimate for it. At Cycle 2 we can elect to re-estimate that characteristic using the Cycle 1 reported data but with the Cycle 2 respondents (and their weights) this time. In an ideal world we would get exactly the same estimates, point and variance. Indeed, if nonresponse has perfectly been addressed for that one characteristic, then nothing is lost by using the respondents of Cycle 2 (which are assumed here to be a strict subset of those of Cycle 1) in lieu of the original respondents, and thus the same estimates can be obtained again. But in practice things are not that simple.

To see why, let us consider two consecutive cycles' sets of weights, Cycle 1 and Cycle 2, say, and the same Cycle 1 dichotomized characteristic of interest. Suppose we are interested in estimating the number of children that had the characteristic of interest at Cycle 1. The basic difference between the two cycles is that Cycle 2 children are a proper subset of Cycle 1 since nonresponse occurred between the two cycles. If we were successful at eliminating all impacts of nonresponse, then both cycles' estimates would be equal. If we use the Cycle 1 set of bootstrap weights we will compute the 1,000 bootstrap estimates of "the domain size" and plot their histogram, then we get something looking like this:

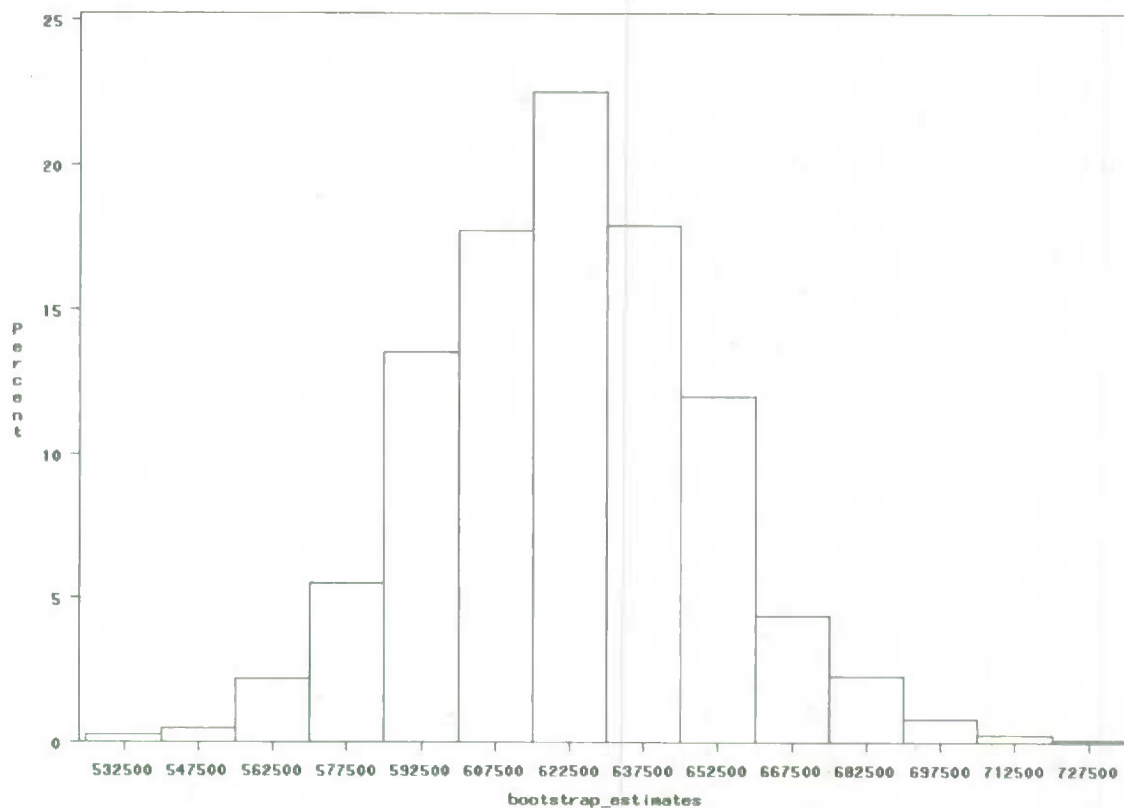


Figure D-1: Bootstrap distribution with emphasis on the tails.

Let us focus for the sake of the argument on the rightmost tail. Since in the context of a dichotomized variable of interest the estimate is just the sum of the weights of the children involved, it is not too far fetched to think that the extreme right bootstrap estimates, those falling in the 727,500 bin, are due to the presence of a certain number of children with very large survey weights. (There can of course be other reasons for those extreme bootstrap estimates to arise since the survey weight is only one of two key ingredients – the other being the multiplicity – but we assume for our purposes here that the large survey weights are indeed responsible for the large bootstrap estimates.)

Now, at Cycle 2 we have lost some of the Cycle 1 respondents that contributed to producing that Cycle 1 bootstrap distribution through their weight. While the nonresponse strategy *en vigueur* does compensate to some extent for the loss incurred, it cannot stop some weird “local” patterns of nonresponse from happening. And one such weird local pattern could be the loss to nonresponse of the very children who contributed to those extreme Cycle 1 bootstrap estimates. If that were the case, then nonresponse methodology would help compensate for the loss “in some generic way” but would not compensate adequately for the loss of some of the largest weights. Consequently, the Cycle 2 largest bootstrap estimates would not be large enough to match their Cycle 1 counterparts. In other words, in drawing the Cycle 2 bootstrap distribution we would observe that the right-most tail of Cycle 1 has shrunk. And because it would have shrunk, the ensuing variance estimate would not be large enough. This is how a sound and well-implemented bootstrap methodology can still end up yielding variance estimates which longitudinally do not appear to be consistent: the domain

is small and therefore the extreme bootstrap estimates of a cycle depend on a few chosen “extreme” units which may disappear entirely at the next cycle.

Furthermore, unequal weights are an issue with respect to variance estimation because there are longitudinal analyses (like those involving hierarchical models) for which we do not know how to make appropriate use of weights in the calculation of the variance estimate. This ties in with our discussion in Perspective B on the use of normalized weights.

Perspective E: On the mathematical activity of modelling

We have already used the word “model” quite a few times by now and yet we have not talked at all about the intellectual activity of modelling. Actually, as pervasive as the activity of modelling is in applied science, it is a fact that very few statisticians have had any exposure to it as part of their training. As a result, most go from theory to practice with a discomfort, if not a distrust, for modelling which they find too subjective an activity (one in stark contrast with the “objectivity” of science in academia). Applied scientists like field biologists and engineers, on the other hand, are all too familiar with modelling and have come to accept it as being synonymous with day-to-day science.

Our growing use of models in survey statistics requires that we get more familiar with the mathematical activity of modelling. Since it is very difficult to get a suitable introduction to modelling from the usual textbooks a statistician consults, we propose here a succinct overview of the topic which will provide the reader with the basic ideas.

A great place to start a self-study on the topic is to read about it through www.wikipedia.com using the keyword “model” (the items “conceptual model” and “scientific modelling” identified by the site as a result of the search are a good starting point). The website (through its relevant items) also contains basic references for further investigation. Two useful (and more conventional) introductions to models are [Michalewicz and Fogel \(2000\)](#) and [Birta and Arbez \(2007\)](#).

The whole modelling activity is summarized neatly in [Michalewicz and Fogel \(2000\)](#):

Every time we solve a problem we must realize that we are in reality only finding the solution to a model of the problem. All models are a simplification of the real world, otherwise they would be as complex and unwieldy as the natural setting itself. [...] If our model has a high degree of fidelity, we can therefore have more confidence that our solution will be meaningful. In contrast, if the model has too many unfulfilled assumptions and rough approximations, the solution may be meaningless, or worse.

Let us expand on this a bit. Survey statisticians, just like any applied scientist facing a real-life problem of non-trivial complexity, make assumptions in order to make the problem more tractable: they thus choose (whether they realize it or not) to trade a complex problem for a solvable one. In other words, they abandon the original and intractable problem for some simplified one, called the auxiliary problem, which can be solved. The danger with this trade is to lose the relevance of the initial formulation by solving an ill-chosen auxiliary problem whose solution is so irrelevant as to be of any use in the end. To convince the reader that modelling - which is to judiciously simplify a problem as to make it more tractable, yet still relevant - is required in just about any practical situation of anything but trivial complexity, consider the following amusing challenge.

Imagine a truckload of good old Canadian (!) snow being driven due south from Ottawa. We get to monitor at every moment the amount of un-melted snow there is left in the truck bed. Where will the truck be at when the snow has all melted? Somewhere in the state of New York? Or in North Carolina? Or all the way down to the Keys, south of Miami?

To answer this, we will need to know how heat from the sun is absorbed by the snow and how it travels through the truckload, leaving melted snow in its wake. This knowledge is provided by physics' heat equations. So far, so good. But knowing what the equations are and finding their solution in practice though are two separate things. So, to help solve the problem, we may want to make the continuous snow load *discrete*; this is the finite element method which assumes that the (continuous) object actually is a finite grid. We thus replace the problem of finding out the temperature at a given time at *any* point of the load to the simpler one of knowing it at any of the vertices of a finite grid. Of course, the quality of the approximation will depend directly on the number of squares the grid contains: the finer the grid the better the approximation though, computationally speaking, the coarser the grid the better. A compromise will have to be reached. And one way is not to assume a uniform (or equally dense) grid throughout the snow load but a grid with sub-areas of greater/smaller density in sub-divisions: in areas where the temperature is thought to be (approximately) locally uniform use a coarse local grid and a finer one otherwise. But the snow... will we assume it has no impurities? (Can we otherwise account for the effects the sun will have on various degrees of dirty snow?) Will we assume that the sun is the only heat provider through radiation? What about the truck itself (*i.e.*, its transmission heats up the truck bed from underneath)? Will we assume it is not contributing any heat at all? And will we assume heat from the sun only hits the uncovered snow on top but not warming up the sides of the truck? Warming up *both* sides of the truck evenly? Will we assume some constant mean input of sun radiation over a 24-hour period or will we go through the pain of figuring out the varying level of radiations hitting the snow as the truck travels from broad daylight to night time? What about the cooling effect of the wind as the truck runs at higher speed?

One thing ought to be clear by now: there is no way we can solve an interesting real-life problem like this one *without* some extensive assumption-making. But this is what modelling is all about! And the extent of the assumptions made, from realistic to overly simplistic, gives us an idea of how much faith we should have in the solution. After all, we are not interested in the model's solution (read: the simplified problem's solution) *per se*, but rather what it says about the real situation's elusive solution.

It is a fact (sad to some) that different modellers will come to different solutions to the truckload problem above (and more generally to any problem) because they will have made different assumptions along the way. Some will not make any assumptions but will be in the unsettling position of not being able to provide *any* solution at all, albeit an imperfect one.

The idea behind successful modelling is that not *all* the known factors need to be accurately depicted to have any hope of using the solution of the simplified problem to meet our needs. And making assumptions is not something we should be "ashamed" of: we should instead state clearly what they are, so people can replicate and adapt (and judge, yes!) our work. For instance, it may not matter much that we are able to give an accurate three-dimensional description of the snow content of the truckload from pure to real dirty snow; a well-chosen "average" composition could certainly do well. This indeed says that a pocket of pure snow is indeed treated as somewhat dirty and vice-versa. But it may not matter because it could be that the error in the end result of assuming an average composition for the snow rather than its exact composition is of a magnitude which is well below our tolerance point. For instance, assume it is in the range of 10-15 minutes when the exact solution will likely call for several hours of traveling time. Assumptions *need* to be made and what

is crucial is to state these clearly and, if at all possible, to what extent they will impact the solution found.

This evaluation is crucial because it is a way for the modeller to act according to Occam's razor. Loosely speaking, Occam's razor states that all other things being equal we should choose the model which is the simplest. In other words, if our model relies on a myriad of assumptions then see if some can be peeled off (hence the "razor" in the principle's name) without changing the end result *much*. In practice, this principle is most likely violated when we do not keep track of our work and create what appears (to others) as a very complex model. The model can be complex because the situation requires it, or simply because it is overkill.

But this does not mean that *any* auxiliary problem (or model) will do. There are indeed two grand classes of auxiliary problems related to any given real-life problem: those that are simplifications of it (and may not be solvable still) and those that correspond to what we know *a priori* how to solve easily. In the latter case, one first surveys the problems that one knows how to solve and then tries to find one that most closely resembles the problem at hand. That class of problems is often too narrow to be of any value and modelling from that perspective rarely is fruitful. Indeed, the solution to the auxiliary problem taken from that restricted class and the one we seek often have little in common, unfortunately. By keeping the focus on the original problem and diluting it gradually to a simpler problem, one is most likely to reach a point where the auxiliary problem may not be *a priori* solvable but will at least be tractable. For example, minimizing some process one may not reach a differentiable (modelled) function as one would have liked but instead one that can be addressed using existing (but unknown to the modeller) optimization algorithms (or one that requires more creativity still: the once-in-a-lifetime opportunity for a mathematical breakthrough!). Of course, it is quite possible that by legitimately simplifying the original problem one always ends up in the class of auxiliary problems that can be solved directly, for all other intermediate problems were intractable. The idea here is that while that class of problems we know how to solve can be the final point of a modeller's trek, it should not be where it *begins*.

There is a classic example of this in sampling and it has to do with the Neyman allocation of a sample size n to a given set of strata. Recall that the Neyman problem is about allocating a sample size n to H given strata of known sizes N_h and "inner-variances" S_{yh}^2 for the variable of interest Y while minimizing the overall variance of the estimator of the mean of Y . There is a natural set of conditions to be met which are not stated explicitly: the sample size of a stratum needs to be an integer between one and the corresponding stratum size. (Actually, the lower boundary could be set by the survey statistician to two, say, or to some larger number, to allow variance estimation to be carried out in practice.)

We can address the Neyman problem analytically *only* if we ignore the side conditions stated (*i.e.*, by supposing instead that the sample sizes by strata are real unbounded variables). By proceeding analytically, we are led to "optimal" sample sizes which are neither integers nor within a range of values which makes practical sense. The entire issue then becomes whether this auxiliary solution is of any use to us or not. As we argued above, if the problem is simplified too much, then solving the auxiliary problem may not teach us much about the original problem. Incidentally, even if the auxiliary solution is not a perfect match for the original problem, it does not necessarily mean that it is useless. Indeed, if we are able to determine that it is close enough for our purposes then while

we will not solve the initial problem, we will get a suitable approximation. The key point here is that we have to monitor closely the assumptions we are making and assess how they will impact the solution identified.

Let us look closely at an example of the Neyman problem in some detail; we will follow [Ardilly \(2000\)](#) (see page 92) which presents a configuration of inner-stratum variances such that the Neyman problem cannot be solved directly but through a series of auxiliary problems instead. Let us summarize the example's key data here:

Stratum h	N_h	\bar{y}_h	s_h^2
1	500	5	1.5
2	300	12	4
3	150	30	8
4	100	150	100
5	10	600	2,500
Total	1,060		

Table E-1: Initial setup of the strata information

The total sample size we seek to allocate to the five strata is $n=300$.

We notice that the strata with large variability are those which are (relatively speaking) small in size; this will make the unrestricted Neyman problem quite different from the original one, as we shall see.

Applying the analytical solution obtained for the unrestricted Neyman problem:

$$n_h = 300 \frac{N_h \sqrt{s_h^2}}{\sum_k N_k \sqrt{s_k^2}} \quad (\text{E-1})$$

we get as a result the following allocation:

Stratum h	N_h	n_h
1	500	58.56966
2	300	57.38631
3	150	40.57825
4	100	95.64385
5	10	47.82193
Total	1,060	300

Table E-2: The first auxiliary problem's solution

There are two problems with this solution: all sample sizes need rounding to make any practical sense and, for stratum 5, the optimal sample size *exceeds* the corresponding population size for the stratum. If we take this as a sign that stratum 5 needs to be a census, then we can take that stratum

out of the problem and re-allocate *à-la*-unrestricted-Neyman the remaining $300-10=290$ units to the remaining strata 1 to 4. If we do, then we get the following solution:

Stratum h	N_h	n_h
1	500	67.354
2	300	65.99317
3	150	46.66422
4	100	109.9886
5	10	10 (Imposed)
Total	1,060	300

Table E-3: The second auxiliary problem's solution

The same issue with the optimal sample size exceeding the population size arises with stratum 4 now. Following the same recipe, we can solve the auxiliary problem of allocating $300-10-100=190$ units to the first three strata, as the last two were taken out by forcing a census upon them. We get:

Stratum h	N_h	n_h
1	500	71.09139
2	300	69.65505
3	150	49.25356
4	100	100 (Imposed)
5	10	10 (Imposed)
Total	1,060	300

Table E-4: The third auxiliary problem's solution

At last, this solution appears to make sense, aside from the fact that the sample sizes for the first three strata are not integers. If we elect to round to the nearest integer, then we get as our final solution:

Stratum h	N_h	n_h
1	500	71 (Rounded)
2	300	70 (Rounded)
3	150	49 (Rounded)
4	100	100 (Imposed)
5	10	10 (Imposed)
Total	1,060	300

Table E-5: The proposed solution to the original problem, pieced up from the auxiliary ones

This may or may not be the solution to the original problem we were facing, which was the *restricted* Neyman allocation problem. Has the series of auxiliary problems we have defined (and solved, with some tinkering) taken us too far away from the solution or are we still on track? We need to address this if we are to have any faith in the solution obtained. It turns out that it is still on track here, but the point is that a careful argument is needed to establish its relevance (something never done in practice: it is always taken for granted that the two solutions are one and the same in this case).

It turns out that there *is* a direct way to solve the restricted Neyman allocation problem, which is provided by computer-based optimization algorithms. Nothing fancy is needed here: Microsoft Excel's own optimization module, the Solver⁴, handles the original problem very nicely with just a few clicks. Indeed, simply define the total variance (*i.e.*, a given cell) as a function of the sample sizes and inner-strata variances (which occupy other cells) and minimize it subject to constraints on the sample sizes (which are yet other cells).

The following snapshot shows the general setup of the problem in Excel and in the Solver:

The screenshot shows a Microsoft Excel spreadsheet titled "Neyman.xls". The spreadsheet contains data for a Neyman allocation problem. The formula bar shows the formula for cell E11: $\text{=SUM}(I3:I7)$.

	A	B	C	D	E	F
1						
2	Stratum h	N _h	Ybar _h	s ² _h	n _h initial	n _h Neyman
3	1	500	5	1.5	130	130
4	2	300	12	4	80	80
5	3	150	30	8	60	60
6	4	100	150	100	25	25
7	5	10	600	2500	5	5
8	Total	1060			300	300
9						
10						
11	Total variance				0.055388586	
12						
13						
14						
15						
16						
17						
18						
19						
20						
21						
22						
23						
24						
25						
26						
27						

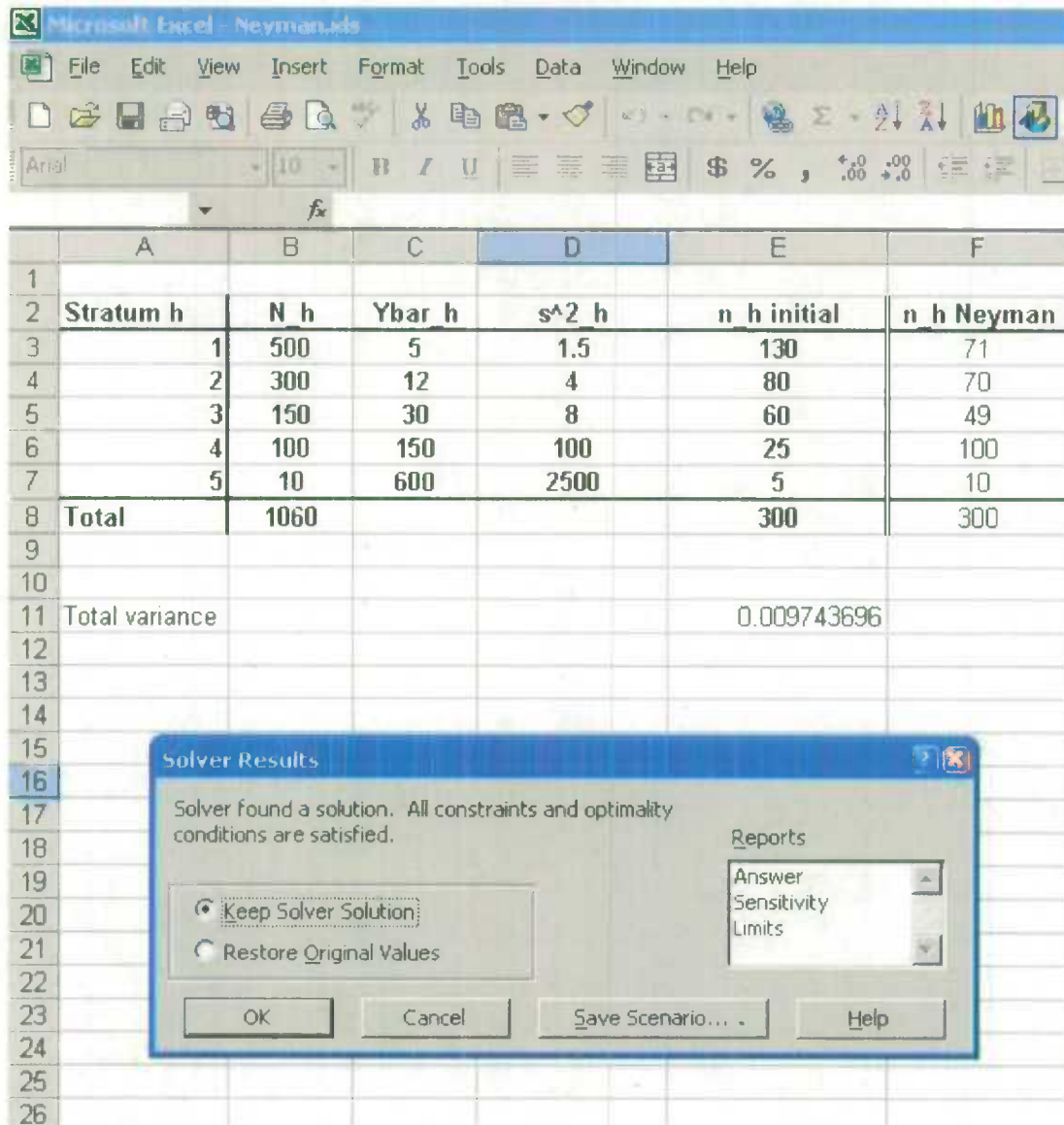
The Solver Parameters dialog box is open, showing the following settings:

- Set Target Cell: E11
- Equal To: ☐ Max ☒ Min ☐ Value of: 0
- By Changing Cells: $\text{\$F\$3:\$F\$7}$
- Subject to the Constraints:
 - $\text{\$F\$3:\$F\$7} \leq \text{\$B\$3:\$B\$7}$
 - $\text{\$F\$3:\$F\$7} = \text{integer}$
 - $\text{\$F\$3:\$F\$7} \geq 1$
 - $\text{\$F\$8} = 300$

⁴ Though not directly accessible from the menus, the Solver does nonetheless come along with the regular Microsoft Office suite: as a specialized tool it's "hidden" from view. (Consult Excel's help on this to see how to make the Solver accessible from the main menus.)

Notice in this snapshot that the Solver, just as any algorithm, needs a starting point; for this we chose the “basic” allocation that [Ardilly \(2000\)](#) provided, but it could have been just about anything else (one of our own making or a standard one like, say, the proportional-to-strata-sizes allocation).

The following shows what we get as solution for the restricted Neyman problem using the Solver:



The screenshot shows a Microsoft Excel window titled "Neyman.xls". The spreadsheet contains a table with 6 columns (A-F) and 26 rows (1-26). The table data is as follows:

	A	B	C	D	E	F
1						
2	Stratum h	N h	Ybar h	s ² h	n h initial	n h Neyman
3	1	500	5	1.5	130	71
4	2	300	12	4	80	70
5	3	150	30	8	60	49
6	4	100	150	100	25	100
7	5	10	600	2500	5	10
8	Total	1060			300	300
9						
10						
11	Total variance				0.009743696	
12						
13						
14						
15						
16						
17						
18						
19						
20						
21						
22						
23						
24						
25						
26						

Overlaid on the spreadsheet is the "Solver Results" dialog box. It contains the following text: "Solver found a solution. All constraints and optimality conditions are satisfied." Below this text are two radio buttons: "Keep Solver Solution" (which is selected) and "Restore Original Values". To the right of these buttons is a "Reports" section with a list box containing "Answer", "Sensitivity", and "Limits". At the bottom of the dialog box are four buttons: "OK", "Cancel", "Save Scenario...", and "Help".

Notice that we do end up with the same solution as previously identified (including the two censuses for the two bottom strata).

In summary, the idea with modelling is to solve a simplified version of the original problem without losing track of the assumptions made (and eventually their toll on our prospect of a solution). The auxiliary problem we choose to solve should be on grounds of relevance to the

original problem rather than solving an easy problem just for the sake of finding some solution. An easily-obtained solution is of no value if it is not close in some sense to the sought-after solution.

Perspective F: On nonresponse bias detection

When trying to evaluate whether or not nonresponse bias exists in a longitudinal survey, the conventional approach is to profile and compare (*e.g.*, via log-odds ratios) the respondents and nonrespondents. This provides a description of who is lost to attrition. Often, there is the finding of middle-class bias: over time, the lower and upper ends of the socio-economic spectrum are lost to attrition. But profiling alone does not indicate that bias is present: we may be disproportionately losing upper and lower income earners, but so long as some are still present and nonresponse is adequately corrected for, it may not result in biased estimates.

So, along with profiling, in order to quantify the presence of bias, practitioners of longitudinal surveys often try to validate their survey's estimates with those of an outside source, which is difficult to do since they are then comparing longitudinal estimates (*i.e.*, referring to some original longitudinal population) with cross-sectional estimates (*i.e.*, referring to some later cross-sectional population). The most widely-cited nonresponse study of this type is [Fitzgerald *et al.* \(1998\)](#).

Along with the finding of middle-class bias, many studies on longitudinal surveys find that bias due to attrition appears to be model specific, so analysts should assess the problem of bias for their particular dataset and analysis.

Detecting the presence of nonresponse bias is a notoriously difficult task since the much needed ingredient – the value for the parameter of interest – is unknown to us. So, failing to work in absolute terms we often fall back to relative measures.

One such relative measure that the NLSCY methodologists use is that of longitudinal consistency. The idea is very simple: for a given cohort and one of its cycles, we look back at the estimates that were released earlier when the cohort was at its first cycle of collection and measure how with current cycle weights the estimates of first cycle characteristics fare. Consider for simplicity the given cycle to be Cycle 2. Unless the Cycle 2 weights were calibrated to yield all Cycle 1 estimates, some differences will be found. (For example, the Cycle 2 estimate of the number of children whose PMK reported living alone at Cycle 1 will differ somewhat from the original estimate of that same characteristic using Cycle 1 weights.) If overall that difference appears large, then there is some indication of bias introduced by the Cycle 2 nonresponse which "explains" our inability to reproduce somewhat accurately earlier-released estimates of the same parameter. The implicit assumption here is that to estimate a parameter of interest the best choice is the first cycle's set of weights: any subsequent set of weights is further plagued by more nonresponse.

Why would it matter that at Cycle 2, say, to pursue our example, we are able to compute an estimate of a Cycle 1 characteristic that matches the one we got back at Cycle 1? After all, if we are still interested in that estimate now that Cycle 2 has been reached, all we need to do is revert to Cycle 1 weights and this will do the trick. Why impose that on the Cycle 2 weights? Indeed, the idea is not to ensure that the latest cycle's set of weights suffices to conduct any past estimates but rather to improve upon the current estimates. We know that the loss of units to nonresponse which brought us to Cycle 2 will require a new set of weights. But it would seem that the best way to tell if the Cycle 2 weights are reasonable is to test their ability to reproduce Cycle 1 estimates. If that were really achieved, then we would have the sense that nothing essential had been lost to

nonresponse since all the "useful" Cycle 1 information needed to create the Cycle 1 estimates had been encapsulated in the Cycle 2 set of weights. The idea is the same one that is behind the use of post-stratification in the NLSCY (or in any other survey for that matter): by ensuring that our weights are "aligned" to Census totals (or projections of these, for inter-censal years), we get an indication that the weights are "on the right track". (The major difference is that census totals are deemed known whereas in our setting the previous cycles' estimates we seek to reinstate are just that – estimates.)

Some people propose the following relative measure to evaluate bias in a longitudinal setting:

$$\frac{\hat{Y}_{non-funnel} - \hat{Y}_{funnel}}{\hat{Y}_{non-funnel}} \quad (F-1)$$

The main flaw, just like with any such measure, is that the main ingredient of a measure of bias is missing: the true value. Actually this is nothing more than a measure of consistency between the various weighting methodologies used at a given cycle. If all estimates compare then it does not mean that bias is not present: it could simply be just an instance of "liars agreeing". On the other hand, if estimates resulting from different methodologies disagree, then these differences should be investigated thoroughly and ultimately explained.

Some people may think of using one form or another of resampling to "hunt" for clues of the presence of nonresponse bias. To give a flavour of this, here is an example using the so-called Poisson bootstrap. (In reality, it is actually nothing more than a Monte Carlo exercise using a Poisson design.) For the sample drawn, propose a nonresponse model and estimate response propensities. Using the set of respondents, r , first compute an estimate of a population parameter of interest. Then draw a large number B of Poisson samples from the set of respondents using the estimated response propensities as the selection probabilities. Finally, for each Poisson sample, compute an estimate using the proper weights (*i.e.*, that of the design *and* the Poisson subsampling) and use the following criterion: if the average of the estimates over the selected B Poisson samples is significantly different than the full sample estimate then this is an indication of bias for the characteristic of interest.

Why a Poisson design? Simply because each respondent gets to be in the sub-sample of random size independently of the others and based on a selection probability which is not (necessarily) the same for all. In other words, the Poisson design is a familiar sampling design which reproduces quite faithfully how the assumed nonresponse model would work at providing us with subsets of respondents, under repeated trials.

The problem with this methodology is that it will never pick up any bias. Indeed, for bias to be present would mean that under the Poisson design the estimator used is design-biased. In the case of a linear estimator such as a total, this will not happen. Any such evidence found will actually be the result of a "bad" series of Poisson samples, and nothing else. In other words, additional Poisson sampling will (approximately) yield the population parameter in simulations because the estimator used is design-unbiased.

This does not mean such a methodology has no use whatsoever. Consider the following simplistic situation. A sample is drawn and among respondents, men and women appear to respond according to different response propensities. Using the estimated propensities by gender we generate a survey estimate. Our client challenges the findings and claims that nonresponse is occurring at random throughout the sample. Despite our best arguments, he/she will not surrender: he/she does not see the difference our assumption will make in the end. We then decide to conduct the Poisson sampling under the missing-at-random assumption. If the client's claim is true (*i.e.*, both working models are equivalent) then the average under the missing at random assumption should be very close to *our* point estimate, which it will not be.

This is really the only use for such a methodology: when there are competing assumptions and the question is whether or not they amount to the same thing. One way to find out is to build the estimate using one methodology and carry out the Poisson sampling with the other, and compare them. In this simple case there really was no need to do the Poisson sampling to be decisive; this was just used for illustrative purposes. But imagine a simple nonresponse model and the competing one is a sophisticated logistic model. We would prefer the simpler one, but only provided it yields about the same estimates. How to tell? By using the methodology just described.

Perspective G: On the estimation of response propensities

There are two competitors; let us recall them:

$$\hat{\theta}_{s_h} = \frac{r_h}{n_h} \quad (\text{G-1})$$

$$\hat{\theta}_{s_h, wgt} = \frac{\sum_{k \in R_h} w_k}{\sum_{k \in s_h} w_k} \quad (\text{G-2})$$

Which one is preferable? The situation may appear ambiguous simply because a natural step was not explicitly described in the construction of the weighted “response rate” estimator which uses (G-2) from the estimator which relies on (G-1). As we shall see, the extra step is omitted because mathematically its presence is artificial since its terms cancel out with other terms so it does not exist permanently. And yet its presence is essential to make sense of the choice of the weighted response rate.

To see this, let us start with the weight obtained after the nonresponse correction (G-2):

$$w_{nr, wgt} = \frac{1}{\pi_D} \times \underbrace{\left(\frac{\sum_{s_h} \frac{1}{\pi_D}}{\sum_{R_h} \frac{1}{\pi_D}} \right)}_{=\hat{\theta}_{s_h, wgt}^{-1}}$$

$$w_{nr, wgt} = \frac{1}{\pi_D} \times \left(\frac{1}{\hat{\pi}_R} \times \hat{\pi}_R \right) \times \frac{\sum_{s \cap RHG} \frac{1}{\pi_D}}{\sum_{R \cap RHG} \frac{1}{\pi_D}} = \left(\frac{1}{\pi_D} \times \frac{1}{\hat{\pi}_R} \right) \times \frac{\sum_{s \cap RHG} \frac{1}{\pi_D}}{\sum_{R \cap RHG} \frac{1}{\pi_D}} \times \hat{\pi}_R$$

$$\text{with } \hat{\pi}_R = \frac{r}{n}$$

$$w_{nr, wgt} = w_{nr, unwt} \times \left(\frac{\sum_{s \cap RHG} \frac{1}{\pi_D} / n}{\sum_{R \cap RHG} \frac{1}{\pi_D} / r} \right) = w_{nr, unwt} \times \left(\frac{\bar{w}_{s \cap RHG}}{\bar{w}_{R \cap RHG}} \right) \quad (\text{G-3})$$

Based on (G-3), it is obvious that the weighted estimator builds from the un-weighted estimator and uses available information (as it should) to improve on it. Indeed, in addition to using the unweighted response rate, it also compares the mean weight of units in the sample to that of responding units and adjusts according to what it finds. So, for example, if the mean of sampled units is larger than that of the responding units, then we need to adjust our weights upwards, and vice-versa. In practice, decomposing the adjustments this way is of no use since mathematically it makes no difference, but conceptually it does, especially when one tries to get a feel for what is going on.

The expression (G-3) is quite interesting in itself because it provides an insight as to when the weighted form provides added-value to the more direct, unweighted form. Indeed, the greater the difference of average weights among sampled units in a RHG compared to that over respondents, the more important the correction. But this is not quite a diagnostic tool in itself to identify RHGs in practice since no consideration is given to ties that may exist between the variable(s) of interest and the nonresponse mechanism.

Incidentally, this decomposition is often useful to make sense of estimators that otherwise seem at first to defy common sense. Another well-suited example is taken from item 3 on p. 391 of [Särndal et al. \(1992\)](#) which concerns the estimation of a domain mean of known population size N_D . Indeed, in that context the suggested estimator is

$$\hat{\bar{y}} = \frac{\sum_{k \in s \cap D} w_k y_k}{\hat{N}_D} \quad (\text{G-4})$$

where

$$\hat{N}_D = \sum_{k \in s \cap D} w_k$$

This is known as the Hajek estimator of the domain mean.

This *seems* to imply that no use is made of the known quantity N_D whatsoever since it does not appear (explicitly) in the expression of the estimator. It is so because one step in its construction is not provided. What makes more methodological sense is to start with the following basic estimator:

$$\hat{\bar{y}}_{alt} = \frac{\sum_{k \in s \cap D} w_k y_k}{N_D} \quad (\text{G-5})$$

This is simply the Horvitz-Thompson estimator.

Now, the Horvitz-Thompson estimator may lead, given a sample s , to an estimate that is too large for either one (or both) of the following reasons:

1) the y values in the sample s *within* the domain are unduly large compared to y -values in the domain not sampled (which of course we do not know);

2) the sample is overly “generous” with domain elements.

For the first point there is nothing we can do: we were unlucky selecting a “bad” sample and that is part of the game which is sampling. But we do not need to suffer through the second point since we have a way to detect before we release the main estimate for the mean that the domain’s share of the sample is larger than expected and thus make the suitable adjustments. How? By comparing the one *auxiliary* estimate \hat{N}_D (which is simply the sum of the weights of the units sampled within the domain) to the known value N_D . If it turns out to be larger, then we will know that the domain is over-represented in the sample, and in that case we would (possibly) gain in revising downward our initial estimate, and vice-versa. In other words, refine the basic Horvitz-Thompson estimator to get:

$$\hat{y}_{alt} = \frac{\sum_{k \in S \cap D} w_k y_k}{N_D} \times \left(\frac{N_D}{\hat{N}_D} \right) \quad (\text{G-6})$$

The extra factor does perform the correction we need. Indeed, if \hat{N}_D is larger than N_D , then we ought to lower our initial estimate, and this is exactly what the factor would do in that case.

Perspective H: On the role of the weights in the creation of the RHGs

In *Perspective G* we addressed the issue of how the weights are used to estimate the response propensities *given* that a set of RHGs has already been agreed upon (somehow). In this *Perspective* we talk briefly about the role weights play in the construction of the RHGs in the first place.

In [Haziza and Beaumont \(2007\)](#), the issue of how to construct the RHGs is addressed in general terms, building their conclusion on the following expression of conditional bias:

$$Bias(\hat{Y}_I | s) \cong \sum_{c=1}^C \frac{1}{\bar{p}_c} \sum_{i \in s_c} w_i (p_i - \bar{p}_c)(y_i - \bar{y}_c) \quad (H-1)$$

From this they advise reducing the bias by creating C groups in which the response is homogeneous and/or the y values are homogeneous. But what about the survey weights which appear in (H-1)? Should the groups be made with the distribution of weights, say, in mind? For instance, in the case of the logistic regression, should the weights be used as part of some generalized least squares procedure?

Although to our knowledge no general consensus has emerged on this issue, accepted wisdom is to identify variables which appear linked to the occurrence of nonresponse. If among these, one finds design variables, then using the weights is redundant since the weights can be seen as a summary of the design information. On the other hand, if no design variable surfaces in the analysis, then using the weights in the creation of the groups will be pretty much useless and will only serve to introduce unwanted noise. So, returning to the example of the logistic-based nonresponse methodology, this means that weights are not used through some generalized least-squares procedure of some kind but simply let the design variables be screened-in as covariates of the logistic model if they were actually found to be significant contributors.

Perspective I: On our use of models and their role in the inferential framework

With the word “model” lurking at every corner, does that mean the survey statistician has to consider model-assisted (as in [Särndal et al. \(1992\)](#)’s title “*Model Assisted Survey Sampling*”) and/or model-based inferential frameworks? In other words, can the survey statistician still operate (tacitly) under the traditional design-based approach (without giving it until then much thought possibly) as he/she is accustomed to despite the growing use of models? And why should the survey statistician even care? And, more generally (and importantly), what is the difference between them all? And what does it imply about the usual bootstrapping activities?

If discussing at length inferential frameworks is not one’s cup of tea, then here is the bottom line: the models the survey statistician refers to, and uses, set up the work under the model-assisted umbrella, which itself falls under the design-based framework (and not under the model-based framework as could have been guessed from the name). And the bootstrap that is used at Statistics Canada, which we present in [Perspective B](#), is a design-based method.

Looking more deeply into this, it ought to be clear reading Section 2.5 of [Särndal et al. \(1992\)](#) that the survey statistician conducts inferences here at Statistics Canada according to a design-based framework. While one might feel that this reminder was not absolutely necessary, it remains that the uneasiness that exists with the model-based/model-assisted/design-based jargon stems directly from not understanding fully what is exactly a design-based framework (and perhaps more importantly what is *not*).

Loosely speaking, the survey statistician has one set of values from which to create the best estimate possible. In order to provide with statistically sound inferences, the survey statistician must explain what stochastic mechanism is responsible for generating that set of values. There are two grand contenders: a sampling selection mechanism and a stochastic model.

Thus, everything begins with valuable information about a characteristic of interest which is expressed in raw form as a set of known values $\{y_1, y_2, \dots, y_n\}$; it is the sample. The two frameworks, design-based and model-based, differ as to how the survey statistician ended up with these values. The work done by the survey statistician falls under the design-based framework if a sampling mechanism from a finite population is what the survey statistician figures is responsible for the obtained values. So, *as samples allowed by the sampling design come and go* (as part of some grand mind exercise since in reality the survey statistician gets to work with just one such sample), the estimator used will take on different values – the estimates; the discrete probability distribution associated with these estimates makes the sampling distribution of the estimator. If the survey statistician evaluates such properties as the mean and variance of the estimator by computing these with respect to that distribution, then the work done indeed falls under the design-based framework. An important nuance is in order: in the design-based view, the y -entities are not random variables, but rather fixed quantities (they are the attributes of elements sampled from the population). Actually, known or not, every unit in the population has one value (and only one) of the y -variable assigned to it: it is its y -attribute (think of a y -attribute just like a person’s current height is that person’s attribute – it is a number not a random entity). What is random is which sample of attributes will be selected from the population; the y -attributes it contains then become

known to us through collection activities. The randomness thus lies with which sample, of all possible ones, will occur

On the other hand, if the survey statistician assumes that the values observed are directly the observations of a random process (to be described using a model) which spits out y -values for a living, then the work done falls under the model-based framework. So, what a design-based survey statistician will call a sample size of n to describe the n observed values, the modeller will speak of a set of realizations from n random variables Y_1, Y_2, \dots, Y_n whose stochastic behaviour is described by the joint statistical distribution underlying the model. In the model-based framework, the mean and variance properties of the estimator, for instance, are determined by the randomness which is described by the postulated model. In other words, to evaluate a statistical property of the estimator like its variance the survey statistician needs to "track" how the variance described in the model gets "transformed" by the estimator into the by-products which are the estimates. For example, by taking the mean of n independent and identically distributed (random) variables with common variance σ^2 the survey statistician realizes that the estimator "shrinks" by a factor of n the "initial" variance on the n random variables stipulated by the model (*i.e.*, the variance of the estimator of the mean under the model is σ^2 / n).

The key feature to tell both frameworks apart is to identify how the statistical properties of the estimator considered are evaluated. The design-based approach will evaluate these properties under a hypothetical replication of the sampling design, while the model-based approach will evaluate its performance with respect to the assumed model.

One reason why it matters to tell the two frameworks apart is that each may suggest a different estimator to be used, based on how we plan on evaluating its performance. In an ideal world, every statistician, from a design-based or model-based school of thought alike, will summarize the information contained in the data at hand the very same way. But in practice, the information contained in the data will not be *perceived* the same way: it may appear to weakly point in one direction to a statistician only appear to point rather strongly in an other direction by another one. In other words, because the collected information is incomplete, some degree of ambiguity will always exist as to how best interpret it. Therefore, in practice, the survey statistician may work with different estimators depending on which framework is used, or even work with the same estimator but assess the performance in different ways. This is not to say that these two frameworks are the only ones possible; some people like to consider blends of these two; in these "hybrid" frameworks, it is possible to seek estimators that perform "equally well" under both design-based and model-based frameworks. We will not go into this any further here.

To illustrate, suppose the following set of five values was observed about a certain variable of interest: $\{y_1 = 8, y_2 = 5, y_3 = 5, y_4 = 1, y_5 = 12\}$. What should the survey statistician make of those two 5s in the sample? In what "direction" are they pointing? Is there a correlation between Y_3 and Y_4 (and thus calls for some correlation analysis to be part of the inferential work), or is it just accidental (and can thus be ignored by using an independence assumption on the Y s)? This is an example of ambiguity. The design-based survey statistician may explain that occurrence by noticing that the sampling design on children he/she used to obtain the sample requires that in addition to a selected child's value, any value of his/her exact twins gets reported as well. The modeller may assume still that the Y s are independent variables and this "oddity" is not enough in

itself to justify that the assumption made will invalidate the results. The modeller may feel that had the evidence of “correlation” been more accentuated, he/she would simply have postulated a different model than the basic one, a more complex one with some covariance structure of some kind to better match the evidence. And here is the key point: if the sampling design does introduce a dependency on the *Y*s, then the modeller can still obtain comparable inferences provided his/her model encapsulates the appropriate information. It is only when there is a significant effect due to the sampling selection which has gone unnoticed (or simply ignored) by the modeller that important differences will appear.

A meatier example is as follows and is based on ideas put forward by [Michael and O'Muircheartaigh \(2008\)](#). Suppose we have a longitudinal sample of children that was obtained from a stratified clustered design of some kind. (Basically, imagine a sample that was obtained by a selection process which depends heavily on relationships among the characteristics of the children.) The research project has two parts: assess from the sample the average height of certain children in the population and determine whether the common wisdom that an adult's height is twice the height the individual was at 2 years old holds true or not.

Using model-based methods to estimate the average height of the children of interest *without* factoring in first the important inner-relationships there exist among children in the sample will undoubtedly lead to questionable inferences. This is because the measure of interest here, the average, is a “blend” of all of the selected children's information, whereas the two-fold factor discussed above is an inner truth: it does not rely on information obtained from other children, it exists *within* each of them. Therefore, estimating the mean, that aggregate of children's information (their height), will be influenced by the structure existing in the sample due to the sampling selection. For instance, if the sample contains disproportionately more males than females, then a direct model-based estimate of the mean will in all likelihood be too high. It does not mean in the case of the mean that the model approach is doomed to failure. It just means that a direct (or overly-simplistic, if one prefers) model which ignores the deep inner-relationships that are “plain to see” to exist will give questionable results. But the same structure within the sample has little or no bearing on the second part of the research project. Indeed, since the two-fold factor is assumed to exist within each child, it does not matter much how the selection of children to form the sample was done: they could all be boys, or they could all have the same ethnic origin, *etc.* That inner-structure does not play itself out in the characteristic, and can therefore be ignored. In practice, though, most research proposals are somewhere in between these two extremes, and evaluating where they stand usually requires a lot of subject-matter and sampling design knowledge.

And what about the model-assisted view in all this? As previously mentioned, despite how it sounds, it is a variant of the design-based framework, not of the model-based one. The idea is the following: in a given survey sampling situation where the choice of a suitable estimator to encapsulate all the known information is not *a priori* obvious, the survey statistician may want to draw inspiration from a similar setting known to exist *outside* survey sampling. The analogy which is *en vogue* features linear regression. If the survey statistician is successful at drawing a parallel between the scenario under investigation and that of linear regression, then he/she may benefit from what is known in that field to best formulate the estimator. The survey statistician may thus formulate a linear regression model which encapsulates the information available and use the estimator this model suggests to formulate a model-inspired estimator, but still go on to conduct

design-based inferences. And the key aspect is this: no matter where the survey statistician found inspiration, no matter what model, if in the end the survey statistician assesses the estimator's properties with respect to the sampling mechanism, then the work falls under the model-assisted variant of the design-based approach. The stipulated model which has assisted the survey statistician is not used in any way to assess the estimator's performance; it is not a by-product of the properties which the survey statistician has assumed to hold when the model was formulated. Whether the elected model was a great choice or a poor one, the performance of the estimator will be assessed through what it yields sample after sample, not from the model assumptions. If the model was a poor choice, then the design-based performance of the estimator *will* suffer, as compared to what it would have been had the model proven to be a great choice. In other words, we cannot cheat and use a model when it is not appropriate and get away with it. The poor value of the choice made will be apparent in the performance of the estimator, even under the design-based approach. The model-assisted approach is a kind of trade-off: when the model is a great choice, it will not perform as well as a pure model-based approach, but when the model is poor, it will perform better than the model-based approach.

Let us consider a trivial example. The survey statistician obtained through simple random sampling without replacement a sample s of n individuals. Assume that they all respond and the interest lies in estimating the total of a variable Y :

$$Y = \sum_{k \in U} y_k \quad (I-1)$$

One approach estimating this quantity, called the prediction approach (see [Valliant et al. \(2000\)](#), section 1.2 or [Samdal et al. \(1992\)](#), section 7.12) is to decompose it into two components, what is known based on the sample and what is not:

$$Y = \sum_{k \in U} y_k = \underbrace{\sum_{k \in s} y_k}_{\text{Known}} + \underbrace{\sum_{k \notin s} y_k}_{\text{Unknown}} \quad (I-2)$$

Suppose that somehow a good guess \hat{y}_k is provided for all of the y_k in the second component, then the survey statistician would be in a position to form an estimate of the total as:

$$\hat{Y} = \underbrace{\sum_{k \in s} y_k}_{\text{Known}} + \underbrace{\sum_{k \notin s} \hat{y}_k}_{\text{Modeled}} \quad (I-3)$$

But modelled how? What does the survey statistician know about those non-selected y_k values? Suppose "Nothing" is an accurate assessment of the situation. Using a model, one way of saying just that is:

$$\begin{aligned} E_{\text{model}}(y_k) &= \beta \\ V_{\text{model}}(y_k) &= \sigma^2 \end{aligned} \quad (I-4)$$

In words: *if* a model were responsible for generating the y -values of the population, then to the survey statistician all values would look *a priori* (i.e., prior to being realized) the same: they are some unknown β . (In reality they are not all just one value; the differences between these values and β make the errors, and the variance statement contains what is assumed known about them.) It is possible to show that in such a situation the best (“best” in some sense that would need to be made precise) assessment of β the survey statistician can make from the information available on the y_k values, which stems entirely from the sample, is the sample-mean. Consequently, the survey statistician’s choice of estimator would be in this case:

$$\hat{Y} = \underbrace{\sum_{k \in s} y_k}_{\text{Known}} + \underbrace{\sum_{k \notin s} \sum_s y_j / n}_{\text{Modeled}} = \frac{N}{n} \sum_s y_j \quad (\text{I-5})$$

Two important remarks are in order. First, this is a naïve example. It is only because the model did not encapsulate any useful “external-to-the-sample” information that the estimator proposed in this case is the basic Horvitz-Thompson estimator. There could be other situations where the survey statistician would like to integrate into the estimator additional information than what is contained in the sample but does not know how to proceed. This is where the model-assisted idea comes into play: it is there to suggest a way to integrate the information encapsulated by a model. Second, strictly speaking, the prediction approach does not lead straight to the GREG estimator, which is the flag-bearer for the model-assisted point of view. We took that line of argument because of its appeal for pedagogical purposes: while it does not explain how the GREG estimator came to be *per se*, the angle it takes is a valuable entry point into the whole model-assisted business. The prediction approach does not usually lead to an asymptotically design-unbiased (ADU) estimator which GREG is. And even if the prediction approach is made to build a prediction-based estimator which is ADU, then the estimator will not be GREG. Indeed, see Example 7.12.1 of [Särndal et al. \(1992\)](#) which shows that under the model at hand, the ADU prediction-based estimator identified is not GREG but an estimator called Brewer’s estimator. But again, the prediction approach offers an intuitive and simple introduction to the model-assisted point of view, and this is why we referred to it.

Perspective J: On the choice of a calibration distance

With calibration the methodologist has to be committed to one distance which measures how far the calibrated weights are from the survey weights. There are several candidates, among which those described on page 378 of [Deville and Särndal \(1992\)](#). Given that the distance plays such a central role in calibration, it would be comforting to have good reasons to favour, in a given context, one distance over another but that is not the case. Indeed, while [Deville and Särndal \(1992\)](#) do argue that the listed distances have been well tried out in other statistical contexts, and that in the end the choice of which one to use often has little bearing, not much is known about the heuristics of any of them. One example is the following distance which is quoted in [Deville and Särndal \(1992\)](#):

$$d_{\chi^2} = \sum_k p_k = \sum_k \frac{(w_{k,cal} - w_k)^2}{w_k} \quad (J-1)$$

for which they simply remark that it resembles the statistic used in the Chi-square test of goodness-of-fit. The p_k 's are the individual penalties incurred by replacing the survey weights by the calibration weights.

In the course of our investigations into calibration we identified one good reason to use the "Chi-square" distance, or one motivation for it; we believe this observation, described in the next Claim, has not been made before in a survey sampling setting.

Claim

Consider individual penalties p_k of the following form:

$$p_k = (w_{k,cal} - w_k)^2 q_k \quad (J-2)$$

with the factor q_k yet to be determined. Of all choices of q_k , the one which results in minimized penalties satisfying the constraint that

$$\sum_k w_{k,cal} = T \quad (J-3)$$

for some total T , and such that the ratio p_k / w_k is constant over all k is

$$q_k = \frac{1}{w_k} \quad (J-4)$$

(up to a multiplicative constant).

In other words, that choice of q_k leads to optimized penalties which, relative to the weight of the units, is the same for everyone.

Before tackling the proof of this claim, it is worthwhile to explain heuristically why that one form is a reasonable choice to start with.

First, it is not appropriate for the penalties to be of the form

$$p_k = w_{k,cal} - w_k \quad (J-5)$$

since summing them over all units would allow for some cancelling among penalties to take place. To address this we could take the absolute difference of these penalties as our revised penalties but given that optimization will take place, squaring the differences is preferable to taking their absolute values. For instance, it is known that the function $x \mapsto |x|$ is not differentiable at zero while $x \mapsto x^2$ is differentiable everywhere.

Proof.

We form the lagrangian function

$$L = \sum_k (w_{k,cal} - w_k)^2 q_k - \lambda (\sum_k w_{k,cal} - T) \quad (J-6)$$

where λ is the Lagrange multiplier associated to our constraint.

The $n+1$ differential equations to be considered (one for each k and an extra one for λ) are:

$$\begin{aligned} (1 \text{ to } n) : \frac{\partial L}{\partial w_k^{cal}} &= 2(w_{k,cal} - w_k)q_k - \lambda = 0 \\ (n+1) : \frac{\partial L}{\partial \lambda} &= \sum_k w_{k,cal} - T = 0 \Rightarrow \sum_k w_{k,cal} = T \end{aligned} \quad (J-7)$$

For each of the first n equations we get:

$$w_{k,cal} = \frac{\lambda}{2q_k} + w_k \quad (J-8)$$

Summing them all and using our constraint we get the following relationship:

$$\underbrace{\sum_k w_{k,cal}}_{=T} = \frac{\lambda}{2} \sum_k \frac{1}{q_k} + \sum_k w_k \quad (J-9)$$

Isolating λ leads to:

$$\lambda = \frac{2 \left(T - \sum_j w_j \right)}{\sum_j \frac{1}{q_j}} \quad (\text{J-10})$$

Substituting this back into our earlier expressions for the $w_{k,cal}$ as functions of λ :

$$w_{k,cal} = \frac{\lambda}{2q_k} + w_k = 2 \frac{\left(T - \sum_j w_j \right)}{\sum_j 1/q_j} \frac{1}{2q_k} + w_k \quad (\text{J-11})$$

Now, for the following ratio

$$\frac{p_k}{w_k} = C, \forall k \quad (\text{J-12})$$

to hold for the optimized penalties just found we need

$$\frac{p_k}{w_k} = \frac{\left[\frac{\left(T - \sum_j w_j \right)}{\sum_j 1/q_j} \frac{1}{q_k} \right]^2 q_k}{w_k} = \frac{\left(\frac{T - \sum_j w_j}{\sum_j 1/q_j} \right)^2}{q_k w_k} = C, \forall k \quad (\text{J-13})$$

Observe that no matter what is the choice of the q_k the numerator of the ratio is the same for all units k . Consequently, up to a multiplicative factor, the choice $q_k = 1/w_k$ will then render the whole ratio independent of the units k as claimed.

Perspective K: On post-stratification

It is known that post-stratification leads to legitimate variance estimates of zero for population counts of domains which are estimated post-strata totals. But what about the variance for domain totals when the domain does not coincide perfectly with a post-stratum? Is it possible for such domains to claim an arbitrary low (and thus illegitimate) variance estimate because an excessive number of post-strata were created? In other words, can the survey statistician abuse of post-stratification to the extent where variance estimates of totals for tiny domains become so small as to be of dubious value?

The answer is “no”: the survey statistician cannot abuse post-stratification in that way, and the same rationale applies to stratification as well. On the other hand, it does not mean that the survey statistician should *always* create as many post-strata as possible.

To make sense of all this we first need to recall that the sampling error measurement is almost always restricted, in practice, to the computation of the variance component: the bias component is assumed to be nil. In that context, a variance estimate of zero associated with an estimate does not mean that the estimate is free of error. It simply says that the estimator will yield that one and same value, sample in and sample out: it is free of *sampling variability* error. Whether that one value is the true value or not is another issue altogether: it has to do with the bias of the estimator. In the case of post-stratification, if the post-strata totals turn out to be inaccurate, then bias will be present if a post-stratified estimator is used. In such a situation, the total measurement of the sampling error, which is the estimated variance of zero *plus* the square of the bias introduced, has to be used instead. For the remainder of this discussion on post-stratification we assume that the variance represents all of sampling error (*i.e.*, there is no bias: the post-strata totals provided to the survey statistician are accurate).

We need to see why there is generally a gain in using a post-stratified estimator to estimate a domain size, no matter how tiny the domain is and the number of post-strata used. For that purpose, let us consider a simple case, namely SRSWOR and a dichotomous variable for which we are interested in estimating the total. In other words, let us consider a domain, D_g , of (unknown) size, N_{Dg} , within a post-stratum, g , of known size, N_g . We need to examine closely the algebraic expression of the variances of the Horvitz-Thompson and the post-stratified estimators. Alternatively, for our purposes, we could just as well investigate all this from a simulation perspective rather than deal with formulae. Let us denote by P the proportion of units of the population that are in D_g . In this context we wonder if there is any gain in variance estimating N_{Dg} using a post-stratified estimator rather than the usual Horvitz-Thompson estimator. After all, the domain is not the whole post-stratum, but just some part of it: will the control on the estimates provided by post-stratifying to known totals extend beyond the post-strata?

Adapting the variance expression (7.6.4) from [Särndal et al. \(1992\)](#) of the approximate variance of the post-stratified estimator to this scenario, thus setting:

$$y_k = \begin{cases} 1 & \text{if } k \in D_g \\ 0 & \text{Otherwise} \end{cases} \quad (\text{K-1})$$

we get:

$$AV_{SRSWOR_post}(\hat{N}_{D_g}) = N^2 \frac{1-f}{n} \left(\frac{N_{D_g}}{N-1} \right) P(1-P) \quad (K-2)$$

The variance of the Horvitz-Thompson estimator under SRSWOR is:

$$V_{SRSWOR}(\hat{N}_{D_g}) = N^2 \frac{1-f}{n} \left(\frac{N}{N-1} \right) \left(P \frac{N_{D_g}}{N} \right) \left(1 - P \frac{N_{D_g}}{N} \right) \quad (K-3)$$

Therefore, the ratio of the variance for the Horvitz-Thompson estimator to that of the post-stratified estimator is:

$$\frac{V_{SRSWOR}(\hat{N}_{D_g})}{AV_{SRSWOR_post}(\hat{N}_{D_g})} = \frac{\left(\frac{N}{N-1} \right) \left(P \frac{N_{D_g}}{N} \right) \left(1 - P \frac{N_{D_g}}{N} \right)}{\left(\frac{N_{D_g}}{N-1} \right) P(1-P)} = \frac{\left(1 - P \frac{N_{D_g}}{N} \right)}{1-P} \quad (K-4)$$

The reader can work out the variance expressions involved in the case of stratification and confirm that the ratio of the variances simplifies to the same factors as in (K-4).

A few things are worth noticing about (K-4). First, as P approaches 1, the gain gets more and more substantial (*i.e.*, the ratio is much greater than 1) because the domain we are considering is approaching the entire post-stratum. Indeed, the denominator approaches 0 as P approaches 1. At the other end of the spectrum, as P approaches 0, the two variances become equivalent. But what happens in between? It is easy to see that the ratio in (K-4) is always greater than 1, from which we conclude that post-stratification always yields an advantage over the Horvitz-Thompson estimator in the estimation of a domain size within a post-stratum.

What is surprising to many about this result on the “unilateral gain” of post-stratifying in such a context is that we are told that stratification is particularly beneficial only if the y -values are homogenous within strata and heterogeneous across them. In other words, we are often left to think that stratification is beneficial mainly if the y -values have a certain configuration with respect to the strata. The reason why we used a dichotomous variable here is to strip down the y -values to their simplest expression, in a way. In other words, we tried to get the y -values out of the picture in order to focus on the other, and often neglected, factor at work: the weights. There are essentially two reasons for an estimate to “locally” go awry: either the y -values involved are extreme or the weights are ill-matched for the domain (or both). The first aspect is what we call a “bad sample”. The latter aspect needs further explanation. The weights are devised under SRSWOR, say, to yield the one and same value when summed up over the whole sample: the sum has to be the (known) population size. See that as a “check” being performed on the weights. But that is a lax check since it is only performed if all units of the sample participate in the domain. If the domain happens to be a proper subset of the sample, then the ensuing sum of weights over the domain cannot be just about anything but no tight control is exerted. This “lack of control” will contribute to the variance

of the estimator. On the other hand, when post-stratification is used, see the post-strata as “periodic checks” being performed on the weights. Indeed, the weights are controlled every time a post-stratum is encountered, as they have to yield the post-stratum total. So, instead of performing the check on the weights just once with the Horvitz-Thompson (that is when all units are involved in the analysis), the post-stratified estimator uses more frequent checks to ensure that the weights are on track. This contributes in giving the post-stratified estimator an edge variance-wise over the Horvitz-Thompson estimator.

To see heuristically how the result above works consider the following: imagine the Horvitz-Thompson estimator of the post-stratum population size yielded an estimate for a given sample which is 10% short of the known count. In other words, assume that with the current methodology we have an under-estimate of the known post-stratum size. With post-stratification, we take this into account and boost all of the concerned weights up by 10% so that the revised (or post-stratified) estimate now matches the known count. This is a quite reasonable measure to take but will the benefit solely be for that one particular estimate (*i.e.*, that of the post-stratum size) or will it prevail also, albeit to a lesser extent, for any estimate of a sub-domain? Said differently, is the gain in variance achieved by post-stratifying to the known post-stratum total just “local” or is it “global” in reach? After all, if the gain was solely local, then there would be little interest in post-stratification because we are not in this business to exclusively produce post-strata population size estimates. Indeed, if that were the case, then our auxiliary source would do just great. But the result above shows that the gain is global. Consider now a proper domain of the post-stratum (*i.e.*, our D_g above). Chances are the sum of the weights of the units of the sample within the domain of the post-stratum, the Horvitz-Thompson estimate really, will *also* fall short of the (unknown) total. Probably not by exactly 10% as with the whole post-stratum, but maybe something like 7% short of the true domain size or something closer like 9.5%; but chances are the estimate for the domain size will fall *short* for the domain just like it fell *short* for the whole post-stratum. It would be very surprising to get all of a sudden an over-estimate for a domain when for the whole post-stratum it is a part of we had an under-estimate. It is reasonable to expect that a given domain in practice (aside from those maliciously constituted to prove this rule wrong) will follow the same trend as the whole post-stratum to which it belongs.

We have shown that the preliminary Horvitz-Thompson estimate of a domain size will gain from the correction introduced in the weights through post-stratification. Assuredly one can (maliciously) *devise* a domain *and* a sample in such a way that the trend observed with a post-stratum will not be that of the domain devised. These cases aside, though, *chances are* domains will behave like the post-strata in a given sample: their population counts are either both under-estimated or both over-estimated prior to post-stratification.

And the same goes for bootstrap estimates for the domain at hand: each gets corrected away from its slight downward trend, and, as a result, a smaller variance estimate can be claimed *provided* the bootstrap is not compromised by the post-strata being too small in size. Indeed, while the comparison of “exact” variance expressions above makes it clear that post-stratification is beneficial, and therefore the more post-strata the better variance-wise, the gain in practice may not be achievable. Indeed, how are we to conduct variance estimation if the sample is not large enough to transmit the information we know to be present? With the bootstrap as the variance estimation method, it is clear that some replicates will have no selected units for a given post-stratum if it is

too small. And if post-stratification is over-used, then chances are that many (if not all!) of the post-strata will turn up empty in some or many of the bootstrap replicates, compromising the whole variance estimation process.

Put differently, the possible gains in variance reduction may tempt the survey practitioner to create numerous post-strata, but the variance estimate obtained in practice may be so unreliable that all potential gains by (over-) post-stratifying may simply vanish. In other words, for a *design-based* variance estimation method like the bootstrap, adequate sample size is a must if the information available for post-stratification is to be exploited. After all, the bootstrap can only track the benefits of post-stratification through what it “finds” in the bootstrap estimates: we do not provide it with any “external” guidance or tip about the anticipated effect of post-stratification. In the end, the survey statistician must strike a good balance between the use of many post-strata to reduce the sampling error, and the use of only well-populated post-strata to ensure a stable variance estimation process.

By focusing here on a dichotomized variable of interest, we have somewhat isolated the contribution of the weights of the sampled units to the variance under post-stratification, by keeping the other major component, the variable of interest, to a “bare minimum”. In that setting we saw that we always gain (with the exception of ill-fated samples) by controlling the weights “here and there” (through the post-strata totals) and thus making sure that for a given sample, they do not go astray.

In summary, post-stratification makes good use of auxiliary information that not only serves to reduce to zero the variance for post-strata totals but also propagates itself to other domains, though the further the domain is from a post-stratum the more tenuous the beneficial effects of post-stratification. While post-stratification is beneficial in theory, in practice, stability of the variance estimator must be taken into account which may actually limit its use (or over-use).

[illegible]

STATISTICS CANADA LIBRARY
BIBLIOTHÈQUE STATISTIQUE CANADA



1010460920

c.3

2008