Proceedings of Statistics Canada Symposium 2014
Beyond traditional survey taking: adapting to a changing world

Study of the "product" sampling scheme as illustrated by the ELFE survey

Guillaume Chauvet,[1] Hélène Juillard[2] and Anne Ruiz-Gazen[3]

Abstract

The Étude Longitudinale Française depuis l'Enfance (ELFE) [French longitudinal study from childhood on], which began in 2011, involves over 18,300 infants whose parents agreed to participate when they were in the maternity hospital. This cohort survey, which will track the children from birth to adulthood, covers the many aspects of their lives from the perspective of social science, health and environmental health. In randomly selected maternity hospitals, all infants in the target population, who were born on one of 25 days distributed across the four seasons, were chosen. This sample is the outcome of a non-standard sampling scheme that we call product sampling. In this survey, it takes the form of the cross-tabulation between two independent samples: a sampling of maternity hospitals and a sampling of days. While it is easy to imagine a cluster effect due to the sampling of maternity hospitals, one can also imagine a cluster effect due to the sampling of days. The scheme's time dimension therefore cannot be ignored if the desired estimates are subject to daily or seasonal variation. While this non-standard scheme can be viewed as a particular kind of two-phase design, it needs to be defined within a more specific framework. Following a comparison of the product scheme with a conventional two-stage design, we propose variance estimators specially formulated for this sampling scheme. Our ideas are illustrated with a simulation study.

Keywords: Variance estimation, independence, two-phase design, multi-stage design.

1. Introduction

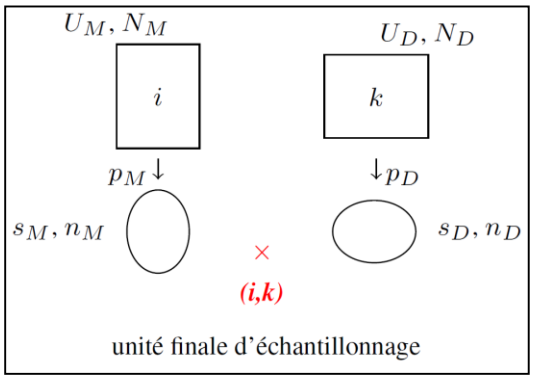
The ELFE is a longitudinal cohort survey, with an initial sample of 18,300 infants. It will track the children from birth to adulthood, covering the many aspects of their lives from the perspective of social science, health and environmental health (Pirus et al., 2010). This study is original because of its multidisciplinary nature, the participation of both parents, and its sampling scheme. The infants in the cohort were selected on the basis of two samples: their date of birth is part of a sample of days in 2011, and their birth location belongs to a sample of maternity hospitals in metropolitan France. Since the sample of days is the same for all selected maternity hospitals (or, conversely, the sample of maternity hospitals is the same for all selected days), this scheme cannot be considered a conventional two-stage design, that is, a design that satisfies the standard assumption that the samples of secondary units selected for the primary units are independent. The final sample is formed by the cross-tabulation of selected locations and dates: it is the result of the product of two samples. In the ELFE, 349 of 544 maternity hospitals were selected at random to take part in the survey, with stratification by hospital size. In addition, four periods in 2011 were selected to represent the seasons: April 1 to 4, June 27 to July 4, September 27 to October 4, and November 28 to December 5. All children born during these periods at the metropolitan maternity hospitals chosen for the ELFE were eligible to participate in the study.

[1] Guillaume Chauvet, Ecole Nationale de la Statistique et de l'Analyse de l'Information (ENSAI), Campus de Ker-Lann, 35710 BRUZ cedex, France, chauvet@ensai.fr
[2] Hélène Juillard, Institut National d'Etudes Démographiques (INED), 133, boulevard Davout - 75020 Paris, France, helene.juillard@ined.fr
[3] Anne Ruiz-Gazen, Toulouse School of Economics (TSE), 21, allée de Brienne - 31015 Toulouse Cedex 6, France, anne.ruiz-gazen@tse-fr.eu

2. Independent product sampling

Consider sampling scheme $p_M(\cdot)$ on population U_M (of maternity hospitals), which yields sample s_M. We will use indices i and j to denote the individuals in this population. Let π_i^M (>0) and π_{ij}^M be the first-order and second-order inclusion probabilities, respectively, and $\Delta_{ij}^M = \pi_{ij}^M - \pi_i^M \pi_j^M$. Consider another sampling scheme, $p_D(\cdot)$, on population U_D (of days), which yields sample s_D. We will use indices k and l to denote the individuals in this population. Let π_k^D (>0) and π_{kl}^D be the first-order and second-order inclusion probabilities, respectively, and $\Delta_{kl}^D = \pi_{kl}^D - \pi_k^D \pi_l^D$.

Figure 1
Sampling in the product population



[text in above figure: final sample unit]

The final sampling unit of interest to us is denoted by the pair (i,k), where $i \in U_M$ and $k \in U_D$ (see Figure 1). We are interested in variable Y, which takes on the value Y_{ik} in maternity hospital i on day k. In the ELFE, we will refer to the pair (i,k) as the "cluster of infants born in the same maternity hospital i on the same day k".

Each final unit belongs to population U, defined as the product of the two source populations:

$$U = U_M \times U_D.$$

We define the **product sample** as

$$s = s_M \times s_D.$$

Within this general framework, product design $p(\cdot)$ can take various forms, as each sampling can be performed conditionally or not on the outcome of the other sampling. The work described below relates to a particular case of the **product design**, in which the two samplings are independent:

$$p(s) = p_M(s_M) \times p_D(s_D).$$

Note that this **independence assumption** of the two sample is analogous to the invariance assumption in conventional two-stage sampling (Särndal, Swensson and Wretman, 1992, p. 134). This sampling scheme is described in Vos (1964) as a special case of space-time designs.[4]

Under these conditions, we can easily calculate the first-order and second-order inclusion probabilities and the

[4] The authors are grateful to symposium participants for pointing out this reference.

covariances for the product design on the basis of the probabilities and covariances of the two source designs. For all units $i, j \in U_M$ and $k, l \in U_D$,

$$\begin{aligned}
\mathbf{E}\left(\mathbf{1}_{\{(i,k)\in s\}}\right) &= \pi_i^M \pi_k^D, \\
\mathbf{E}\left(\mathbf{1}_{\{(i,k)\in s\}}\mathbf{1}_{\{(j,l)\in s\}}\right) &= \pi_{ij}^M \pi_{kl}^D, \\
\Gamma_{ijkl} \equiv \mathbf{Cov}\left(\mathbf{1}_{\{(i,k)\in s\}}, \mathbf{1}_{\{(j,l)\in s\}}\right) &= \pi_{ij}^M \pi_{kl}^D - \pi_i^M \pi_j^M \pi_k^D \pi_l^D \\
&= \Delta_{kl}^D \pi_i^M \pi_j^M + \Delta_{ij}^M \pi_k^D \pi_l^D + \Delta_{kl}^D \Delta_{ij}^M, \qquad (1)
\end{aligned}$$

where $\mathbf{1}_{\{\cdot\}}$ is the indicator function.

We are interested in the total $t_Y = \sum_{i\in U_M}\sum_{k\in U_D} Y_{ik}$, whose unbiased estimator is

$$\hat{t}_Y = \sum_{i\in S_M}\sum_{k\in S_D} \frac{Y_{ik}}{\pi_i^M \pi_k^D} = \sum_{i\in S_M} \frac{\hat{Y}_{i\bullet}}{\pi_i^M} = \sum_{k\in S_D} \frac{\hat{Y}_{\bullet k}}{\pi_k^D}, \qquad (2)$$

where $\hat{Y}_{i\bullet}$ is the Horvitz-Thompson estimator of the total for maternity hospital i and $\hat{Y}_{\bullet k}$ is the Horvitz-Thompson estimator of the total for day k. The variance of estimator $\hat{t}_Y$ can be written as

$$V_{prod}\left(\hat{t}_Y\right) = \sum_{i,j\in U_M}\sum_{k,l\in U_D} \Gamma_{ijkl} \frac{Y_{ik}}{\pi_i^M \pi_k^D} \frac{Y_{jl}}{\pi_j^M \pi_l^D}. \qquad (3)$$

Consider the special case of design SI $\times$ SI where $p_D(\cdot)$ is a simple random sample without replacement (SI) of size n_D in population U_D of size N_D and where $p_M(\cdot)$ is a simple random sample without replacement of size n_M in population U_M of size N_M. Using identity (1), which yields Γ_{ijkl}, we can rewrite the variance given in (3) as

$$\begin{aligned}
V_{prod}\left(\hat{t}_Y\right) &= (N_D)^2\left(\frac{1}{n_D} - \frac{1}{N_D}\right) S^2_{Y_{\bullet\circ}} + (N_M)^2\left(\frac{1}{n_M} - \frac{1}{N_M}\right) S^2_{Y_{\circ\bullet}} \\
&+ (N_D)^2\left(\frac{1}{n_D} - \frac{1}{N_D}\right)(N_M)^2\left(\frac{1}{n_M} - \frac{1}{N_M}\right) S^2 \qquad (4)
\end{aligned}$$

where

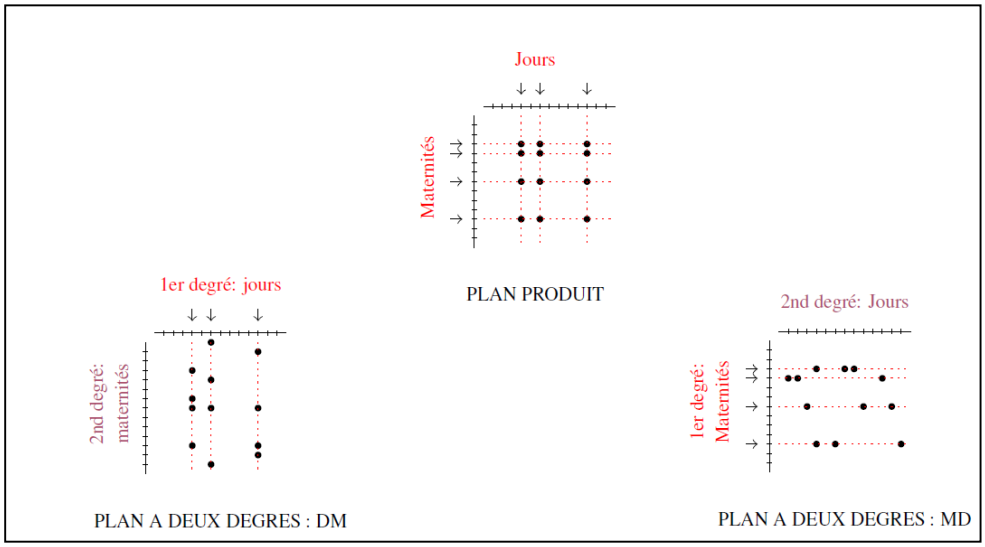
$$\begin{aligned}
S^2_{Y_{\bullet\circ}} &= \frac{1}{N_D - 1}\sum_{k\in U_D}\left(Y_{\bullet k} - \frac{1}{N_D}\sum_{l\in U_D} Y_{\bullet l}\right)^2, \\
S^2_{Y_{\circ\bullet}} &= \frac{1}{N_M - 1}\sum_{i\in U_M}\left(Y_{i\bullet} - \frac{1}{N_M}\sum_{j\in U_M} Y_{j\bullet}\right)^2, \\
S^2 &= \frac{1}{N_D - 1}\frac{1}{N_M - 1}\sum_{k\in U_D}\sum_{i\in U_M}\left(Y_{ik} - \bar{Y}_{i\bullet} - \bar{Y}_{\bullet k} + \bar{\bar{Y}}_{\bullet\bullet}\right)^2
\end{aligned}$$

with $\bar{Y}_{i\bullet} = \frac{1}{N_D}\sum_{l\in U_D} Y_{il}$, $\bar{Y}_{\bullet k} = \frac{1}{N_M}\sum_{j\in U_M} Y_{jk}$ and $\bar{\bar{Y}}_{\bullet\bullet} = \frac{1}{N_D}\frac{1}{N_M}\sum_{l\in U_D}\sum_{j\in U_M} Y_{jl}$.

3. Comparison of product design and two-stage design

A conventional two-stage design requires two assumptions: independence of the various second-stage selections, conditional on the first stage of sampling; and independence of the selections performed at each stage, which, as mentioned previously, is known as the invariance property. For an independent product design, the second assumption is satisfied (independence of the sample of maternity hospitals and the sample of days), but the first is not (the same sample of days is used for all maternity hospitals).

Figure 2
Product sampling in the maternity hospital–day population, two-stage design in which days are selected first, two-stage design in which maternity hospitals are selected first



[text in above figure: TWO-STAGE DESIGN: DM: second stage: maternity hospitals; first stage: days—PRODUCT DESIGN: Maternity hospitals; days—TWO-STAGE DESIGN: MD: first stage: maternity hospitals; second stage: days]

In the case of two-stage sampling design MD (see Figure 2), a sample S_M of size n_M is selected in U_M in the first stage; then in each primary unit i of S_M, a sample S_i of secondary units is selected independently in U_D. All the S_i samples are of the same size, n_D. The variance of this sampling scheme is denoted V_{MD}. In the case of simple random sampling without replacement at each stage, denoted {SI,SI}, we have

$$V_{MD}\left(\hat{t}_Y\right) = N_M^2\left(\frac{1}{n_M}-\frac{1}{N_M}\right)S_{Y_{\bullet\circ}}^2 + \frac{N_M}{n_M}N_D^2\left(\frac{1}{n_D}-\frac{1}{N_D}\right)\sum_{i\in U_M}S_{Y_{i\circ}}^2, \tag{5}$$

where

$$S_{Y_{i\circ}}^2 = \frac{1}{N_D-1}\sum_{k\in U_D}\left(Y_{ik}-\frac{1}{N_D}\sum_{l\in U_D}Y_{il}\right)^2.$$

In the case of two-stage sampling design DM, the process is analogous, using population U_D in the first stage. The variance of this sampling scheme is denoted V_{DM}. In the {SI,SI} case, we have

$$V_{DM}\left(\hat{t}_Y\right) = N_D^2\left(\frac{1}{n_D}-\frac{1}{N_D}\right)S_{Y_{\circ\bullet}}^2 + \frac{N_D}{n_D}N_M^2\left(\frac{1}{n_M}-\frac{1}{N_M}\right)\sum_{i\in U_D}S_{Y_{\circ k}}^2, \tag{6}$$

where

$$S^2_{Y_{\circ k}} = \frac{1}{N_M - 1}\sum_{i \in U_M}\left(Y_{ik} - \frac{1}{N_M}\sum_{j \in U_M} Y_{jk}\right)^2.$$

The difference between the variance of product design SI $\times$ SI given in equation (4) and the variance of design {SI,SI} given in equation (5) or (6) is not necessarily positive. Consider the behaviour model

$$m: Y_{ik} = \mu + \sigma_1 U_i + \sigma_2 V_k + \sigma_3 W_{ik}, \tag{7}$$

where $U_i, V_k, W_{ik} \square N(0,1)$ and $\sigma_1, \sigma_2, \sigma_3 \in \square^+$, and where σ_1 denotes a "maternity hospital" effect, σ_2 a "day" effect, and σ_3 a residual effect. Under model (7), it can be shown that the anticipated variance of the product design is always greater than the anticipated variance of the two-stage design considered. That is,

$$E_m\left[V_{prod}\left(\hat{t}_Y\right) - V_{MD}\left(\hat{t}_Y\right)\right] = N_M^2 N_D^2 \frac{n_M - 1}{n_M}\left(\frac{1}{n_D} - \frac{1}{N_D}\right)\sigma_2^2, \tag{8}$$

$$E_m\left[V_{prod}\left(\hat{t}_Y\right) - V_{DM}\left(\hat{t}_Y\right)\right] = N_M^2 N_D^2 \frac{n_D - 1}{n_D}\left(\frac{1}{n_M} - \frac{1}{N_M}\right)\sigma_1^2, \tag{9}$$

where the expected value under model (7) is denoted E_m. This difference depends on the second stage of sampling: the larger the size of the second-stage samples, the smaller the difference is between the two variances. Conversely, the greater the variability between secondary units σ_2^2, the larger variance V_{prod} is with respect to variance V_{MD}. Analogously, the greater the variability between secondary units σ_1^2, the larger the variance is with respect to variance V_{DM}.

4. Variance estimation

The Horvitz-Thompson estimator of $V_{prod}\left(\hat{t}_Y\right)$ is

$$\hat{V}_{HT}\left(\hat{t}_Y\right) = \sum_{i,j \in S_M}\sum_{k,l \in S_D} \frac{\Gamma_{ijkl}}{\pi_{ij}^M \pi_{kl}^D}\frac{Y_{ik}}{\pi_i^M \pi_k^D}\frac{Y_{jl}}{\pi_j^M \pi_l^D}.$$

This estimator is unbiased if all the π_{ij}^M and all the π_{kl}^D are strictly positive, for all $(i,j) \in U_M{}^2$, $(k,l) \in U_D{}^2$. In the case of product design SI $\times$ SI, this estimator can be written in a symmetrical form for samples S_M and S_D:

$$\begin{aligned}\hat{V}_{HT}\left(\hat{t}_Y\right) &= N_M^2\left(\frac{1}{n_M} - \frac{1}{N_M}\right)s^2_{\hat{Y}_{\circ\bullet}} + N_D^2\left(\frac{1}{n_D} - \frac{1}{N_D}\right)s^2_{\hat{Y}_{\bullet\circ}} \\ &- N_D^2\left(\frac{1}{n_D} - \frac{1}{N_D}\right)N_M^2\left(\frac{1}{n_M} - \frac{1}{N_M}\right)s^2\end{aligned} \tag{10}$$

where

$$s^2_{\hat{Y}_{\bullet\circ}} = \frac{1}{n_D - 1}\sum_{k\in S_D}\left(\hat{Y}_{\bullet k} - \frac{1}{n_D}\sum_{l\in S_D}\hat{Y}_{\bullet l}\right)^2,$$

$$s^2_{\hat{Y}_{\circ\bullet}} = \frac{1}{n_M - 1}\sum_{i\in S_M}\left(\hat{Y}_{i\bullet} - \frac{1}{n_M}\sum_{j\in S_M}\hat{Y}_{j\bullet}\right)^2,$$

$$s^2 = \frac{1}{n_D - 1}\frac{1}{n_M - 1}\sum_{k\in S_D}\sum_{i\in S_M}\left(Y_{ik} - \frac{1}{n_D}\sum_{l\in S_D}Y_{il} - \frac{1}{n_M}\sum_{j\in S_M}Y_{jk} + \frac{1}{n_D n_M}\sum_{l\in S_D}\sum_{j\in S_M}Y_{jl}\right)^2.$$

Note that this estimator is not unbiased on a term-by-term basis for the variance form given in (4). In particular, the third term of (10) is negative. If the absolute value of this term is greater than that of the sum of the first two terms, the variance estimator $\hat{V}_{HT}\left(\hat{t}_Y\right)$ can have negative values. This is possible in particular when n_M and n_D are small.

To our knowledge, this estimator is not available in any software application. We examine some **simplified estimators** so that we can suggest a user-friendly tool for producing variance estimates. The following are three simplified estimators for the SI $\times$ SI design:

$$\hat{V}_{SIMP1}\left(\hat{t}_Y\right) = N_M^2\left(\frac{1}{n_M} - \frac{1}{N_M}\right)s^2_{\hat{Y}_{\circ\bullet}}, \tag{11}$$

$$\hat{V}_{SIMP2}\left(\hat{t}_Y\right) = N_D^2\left(\frac{1}{n_D} - \frac{1}{N_D}\right)s^2_{\hat{Y}_{\bullet\circ}}, \tag{12}$$

$$\hat{V}_{SIMP3}\left(\hat{t}_Y\right) = N_M^2\left(\frac{1}{n_M} - \frac{1}{N_M}\right)s^2_{\hat{Y}_{\circ\bullet}} + N_D^2\left(\frac{1}{n_D} - \frac{1}{N_D}\right)s^2_{\hat{Y}_{\bullet\circ}}. \tag{13}$$

Under standard regularity conditions, $\hat{V}_{SIMP1}\left(\hat{t}_Y\right)$ is approximately unbiased when $n_D \to \infty$ and when n_M is bounded. Analogously, $\hat{V}_{SIMP2}\left(\hat{t}_Y\right)$ is approximately unbiased when $n_M \to \infty$ and when n_D is bounded. Lastly, the simplified estimator $\hat{V}_{SIMP3}\left(\hat{t}_Y\right)$ is approximately unbiased when $n_D \to \infty$ or $n_M \to \infty$. The advantage of these three estimators is that they always have positive values and that they are pre-programmed in applications such as R, SAS and Stata.

5. Simulation study

The results presented here can be illustrated with a simulation study. The model proposed in (7) is used to generate a product population of $N_M = 1{,}000$ maternity hospitals and $N_D = 1{,}000$ days. We use the parameters $\sigma_1 = \sigma_2 = \sigma_3 = 5$ and $\mu = 200$. We repeat $B = 10{,}000$ times the selection of a product sample with design SI $\times$ SI of size $n_M \times n_D$. In the simulations, we varied the sample sizes n_M and n_D between 2 and 300. For each variance estimator $\hat{V}$, we calculate the Monte Carlo relative bias:

$$\%RB_{MC}(\hat{V}) = 100 \times \frac{B^{-1}\sum_{b=1}^{B}\hat{V}^{(b)} - V}{V},$$

where the actual variance V is approximated with a set of 50,000 independent simulations. The number of negative values, # NEGATIVES, taken by the unbiased estimator $\hat{V}_{HT}$ is also calculated.

Table 1
Relative bias of four variance estimators and number of negative values for the unbiased variance estimator

n_M	2	10	300	300
n_D	2	300	10	300
$\%RB_{MC}\left(\hat{V}_{HT}\right)$	-0.71	0.51	0.02	-0.00
# NEGATIVES	1,309	0	0	0
$\%RB_{MC}\left(\hat{V}_{SIMP1}\right)$	-41.10	-98.12	-2.42	-49.64
$\%RB_{MC}\left(\hat{V}_{SIMP2}\right)$	-37.38	-2.23	-97.63	-50.79
$\%RB_{MC}\left(\hat{V}_{SIMP3}\right)$	21.52	-0.54	-0.05	-0.44

The results are shown in Table 1. We note that there are 1,309 negative values in the 10,000 simulations when n_D and n_M are equal to 2. As expected, $\hat{V}_{SIMP1}$ is approximately unbiased when n_M is large and n_D is small (-2.42% bias) and conversely for $\hat{V}_{SIMP2}$. With our set of simulations, the bias of the third simplified estimator, $\hat{V}_{SIMP3}$, is negligible when n_M is large or n_D is large (in this case, equal to 300).

6. Conclusion

Other simplified estimators were studied, as one of the objectives was to help users choose between the procedures available in applications on the basis of their data and the required assumptions.

We covered the case of the SI $\times$ SI design in greater detail, but the framework within which we defined the product design is more general and applies to any $p_D(\cdot)$ and $p_M(\cdot)$ designs. A study is being conducted of various Yates-Grundy estimators, with an application to the case of Poisson sampling conditional on size. We are currently working on taking a possible non-response phase into account in variance estimation for a product design. In addition, we are studying variance estimation by linearization or bootstrapping for parameters more complex than totals.

References

Pirus, C., Bois, C., Dufourg, M.-N., Lanoë, J.-L., Vandentorren, S., Leridon, H., and the ELFE team. (2010), "La construction d'une cohorte: l'expérience du projet français Elfe", *Population* 65(4): p. 637-670.

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

Vos, J.W.E. (1964), "Sampling in space and time", *Review of the International Statistical Institute* 32(3): p. 226-241.