Statistics and Probability Letters 130 (2017) 71–75

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

# Revisiting variance decomposition when independent samples intersect

Yves Tillé, Audrey-Anne Vallée *

*Institute of Statistics, University of Neuchâtel, 2000 Neuchâtel, Switzerland*

ARTICLE INFO

*Article history:*
Received 10 March 2017
Received in revised form 17 July 2017
Accepted 21 July 2017
Available online 1 August 2017

*MSC 2010:*
62d05

*Keywords:*
Reverse approach
Two-phase sampling
Two-stage sampling
Variance estimation

ABSTRACT

The variance and the estimated variance of the expanded estimator in the intersection of two independent samples can be decomposed into two ways. Due to the inclusion probabilities, it is generally more practical to compute the variance with one decomposition. With the other one, it is more convenient to estimate the variance.


## 1. Introduction

In survey sampling, when the sampling designs include several stages or phases, variance estimation suddenly becomes much more intricate. In two-stage sampling, when secondary units are selected in a sample of primary units, one already obtains a curious result. The variance of the expanded estimator and the estimator of the variance can both be decomposed into two terms. However each term of the estimator does not estimate the corresponding term of the variance (see among others Särndal et al., 1992 pp. 137–139). Beaumont et al. (2015) linked the variance estimator to the reverse approach of the decomposition of the variance. However, these authors did not explain the fact that the variance is obtained from a different conditioning than the variance estimator. Variance decomposition is also crucial in nonresponse because questionnaire nonresponse can be modeled as a second phase of sampling. Several options exist to estimate the variance with nonresponse as the two-phase approach (Särndal, 1992) and the reverse approach (Fay, 1991; Shao and Steel, 1999).

In this paper, we discuss the variance and its estimation in samples that are the intersection of two independent samples. We show that two different decompositions of the variance can be obtained. One of them is more interesting for variance estimation. This is explained by possible simplifications of the joint inclusion probabilities of one of the two samples.

## 2. General case

We consider the case where two independent samples intersect. Define $U = \{1, \dots, N\}$ a population of size $N$ and let $s^A$ and $s^B$ be samples of $U$. Two sampling designs are defined on $U$, say $p^A(s^A)$ and $p^B(s^B)$ such that $p^A(s^A) \geq 0$, $p^B(s^B) \geq 0$,

$$\sum_{s^A \subset U} p^A(s^A) = 1 \quad \text{and} \quad \sum_{s^B \subset U} p^B(s^B) = 1.$$

* Corresponding author.
*E-mail addresses:* yves.tille@unine.ch (Y. Tillé), audrey-anne.vallee@unine.ch (A.A. Vallée).

http://dx.doi.org/10.1016/j.spl.2017.07.012
0167-7152/

Define the two random samples $S^A$ and $S^B$ such that $\text{pr}(S^A = s^A) = p^A(s^A)$ and $\text{pr}(S^B = s^B) = p^B(s^B)$. The two random samples are assumed to be independent in the sense that $\text{pr}(S^A = s^A, S^B = s^b) = p^A(s^A)p^B(s^B)$.

Let $I_k^A$ and $I_k^B$ be respectively the indicator variables of the presence of unit $k$ in samples $S^A$ and $S^B$. The first-order inclusion probabilities are $\pi_k^A = \text{E}(I_k^A) = \text{pr}(k \in S^A)$ and $\pi_k^B = \text{E}(I_k^B) = \text{pr}(k \in S^B)$. The joint inclusion probabilities are $\pi_{k\ell}^A = \text{E}(I_k^A I_\ell^A) = \text{pr}(k, \ell \in S^A)$ and $\pi_{k\ell}^B = \text{E}(I_k^B I_\ell^B) = \text{pr}(k, \ell \in S^B)$, with $\pi_{kk}^A = \pi_k^A$ and $\pi_{kk}^B = \pi_k^B$, for $k, \ell \in U$. Moreover define $\Delta_{k\ell}^A = \pi_{k\ell}^A - \pi_k^A \pi_\ell^A$ and $\Delta_{k\ell}^B = \pi_{k\ell}^B - \pi_k^B \pi_\ell^B$. Consider the sampling design obtained by intersecting two independent samples $S = S^A \cap S^B$. Due to the independence, we have $I_k = I_k^A I_k^B$, $\pi_k = \text{pr}(k \in S) = \pi_k^A \pi_k^B$ and $\pi_{k\ell} = \text{pr}(k, \ell \in S) = \pi_{k\ell}^A \pi_{k\ell}^B$. Next define $\Delta_{k\ell} = \pi_{k\ell} - \pi_k \pi_\ell$, $k, \ell \in U$.

Beaumont and Haziza (2016) define a two-phase sampling design as strongly invariant if $\text{pr}(S^B = s^B | S^A) = \text{pr}(S^B = s^B)$, where $S^A$ is the first phase sample and $S^B$ is the second phase sample. In other words, the selection of the second phase sample does not depend on the selection of the first phase sample, which means that the two samples are independent. This definition does not contain the two-phase sampling design as defined for instance by Särndal and Swensson (1987). Indeed, these authors admit that the second phase of the design can depend on the first phase, that is $\pi_k^B$ and $\pi_{k\ell}^B$ are functions of $S^A$. In this case, the designs are not independent and the theory below does not apply.

The two-stage design can be seen as a specific case of the two-phase design that is strongly invariant. The first stage corresponds to the selection of primary units, e.g. municipalities regrouping households, and the second stage consists in selecting the secondary units, e.g. households. Särndal et al. (1992, pp. 137–139) explain that a two-stage sampling design must satisfy the principles of invariance and independence. For these authors, invariance means that the selection of the secondary units of the second stage does not depend on the first stage. Independence means that the secondary units are selected independently from one primary unit to another one. The definition of strongly invariant samples of Beaumont and Haziza (2016) corresponds to the invariance of Särndal et al. (1992, pp. 137–139). Two-stage sampling can be viewed as the intersection of two independent samples selected by a cluster design and a stratified design. In the cluster sample, a set of clusters, which correspond to the primary units, is selected. All the secondary units in this set are therefore selected. In the stratified sample, one set of secondary units is selected per stratum, where a stratum corresponds to a primary unit. The intersection of this stratified sample and the chosen secondary units of the cluster sample is a two-stage sample. Samples of secondary units are selected in sampled primary units.

Another specific case of independent samples is questionnaire nonresponse. A sample is selected in the population and some units in this sample are respondents, the others are nonrespondents. The sample of units in the population is seen as a first sample, selected according to a sampling design $p^A(.)$. The set of respondents is seen as a second sample which is independent from the first one. Moreover the second sample $p^B(.)$ is in general assumed to be a Poisson design, which means that $\Delta_{k\ell}^B = 0$ when $k \neq \ell$.

## 3. Estimation and variance estimation

Suppose that the variable of interest $y$ takes value $y_k$ on unit $k$ of the population. The variable is observed on units selected in a strongly invariant two-phase sample $S = S^A \cap S^B$. In order to estimate the total $Y = \sum_{k \in U} y_k$, one can use the expanded estimator $\widehat{Y} = \sum_{k \in S} y_k / \pi_k$ (Narain, 1951; Horvitz and Thompson, 1952).

The variance of $\widehat{Y}$ is

$$\text{var}(\widehat{Y}) = \sum_{k \in U} \sum_{\ell \in U} \frac{y_k y_\ell}{\pi_k \pi_\ell} \Delta_{k\ell}$$

and can be unbiasedly estimated by

$$\widehat{\text{v}}(\widehat{Y}) = \sum_{k \in S} \sum_{\ell \in S} \frac{y_k y_\ell}{\pi_k \pi_\ell} \frac{\Delta_{k\ell}}{\pi_{k\ell}}.$$

The delicate elements are the decompositions of $\Delta_{k\ell}$ and $\Delta_{k\ell}/\pi_{k\ell}$. They need to be decomposed in function of $p^A(.)$ and $p^B(.)$ by means of the law of total variance. With this law, the variance of a random variable can be decomposed conditionally to another random variable. Consider for instance two random variables $x_1$ and $x_2$, the variance of $x_1$ is decomposed as

$$\text{var}\,(x_1) = \text{E var}\,(x_1 | x_2) + \text{var E}\,(x_1 | x_2)\,.$$

This can be extended to a covariance. Consider a third random variable $x_3$. The covariance between $x_1$ and $x_2$ can be decomposed as

$$\text{cov}\,(x_1, x_2) = \text{E cov}\,(x_1, x_2 | x_3) + \text{cov}\,\left[\text{E}\,(x_1 | x_3)\,, \text{E}\,(x_2 | x_3)\right].$$

For $\Delta_{k\ell}$, there are two possible decompositions. The usual one consists in using the law of total variance by conditioning with respect to $S^A$:

$$\begin{aligned}\Delta_{k\ell} &= \text{cov}(I_k, I_\ell) = \text{E cov}(I_k, I_\ell \mid S^A) + \text{cov}\left[\text{E}(I_k \mid S^A), \text{E}(I_\ell \mid S^A)\right] \\ &= \Delta_{k\ell}^B \pi_{k\ell}^A + \pi_k^B \pi_\ell^B \Delta_{k\ell}^A. \end{aligned} \tag{1}$$

The reverse decomposition consists in conditioning with respect to $S^B$:

$$\begin{aligned}\Delta_{k\ell} &= \text{cov}(I_k, I_\ell) = \text{E}\,\text{cov}(I_k, I_\ell \mid S^B) + \text{cov}\left[\text{E}(I_k \mid S^B), \text{E}(I_\ell \mid S^B)\right] \\ &= \Delta^A_{k\ell}\pi^B_{k\ell} + \pi^A_k\pi^A_\ell\Delta^B_{k\ell}.\end{aligned}$$

This reverse approach is usable because the sampling design is strongly invariant, we can inverse the order of the two phases.

We can also obtain two decompositions for the ratio $\Delta_{k\ell}/\pi_{k\ell}$. The usual one is

$$\frac{\Delta_{k\ell}}{\pi_{k\ell}} = \frac{\Delta^B_{k\ell}}{\pi^B_{k\ell}} + \frac{\pi^B_k\pi^B_\ell\Delta^A_{k\ell}}{\pi^A_{k\ell}\pi^B_{k\ell}},$$

and the reverse one is

$$\frac{\Delta_{k\ell}}{\pi_{k\ell}} = \frac{\Delta^A_{k\ell}}{\pi^A_{k\ell}} + \frac{\pi^A_k\pi^A_\ell\Delta^B_{k\ell}}{\pi^A_{k\ell}\pi^B_{k\ell}}. \tag{2}$$

The decompositions obtained by conditioning with respect to $S^A$ and with respect to $S^B$ are obviously equal.

## 4. Poisson sampling

A sample with nonresponse is often seen as a two-phase sampling design. The first phase is the sampling design of the survey. The second phase is a Poisson sampling design corresponding to the nonresponse phase. Two-phase sampling with a Poisson sample at the second phase is strongly invariant and there are two decompositions of $\Delta_{k\ell}$. If $S^B$ is selected according to a Poisson design, $\pi^B_{k\ell} = \pi^B_k\pi^B_\ell + I\{k = \ell\}\pi^B_k(1 - \pi^B_k)$ and thus $\Delta^B_{k\ell} = I\{k = \ell\}\pi^B_k(1 - \pi^B_k)$, for all $k, \ell \in U$, where $I\{A\}$ equals 1 if $A$ is true and 0 otherwise. The quantity $\Delta^B_{k\ell}$ vanishes when $k \neq \ell$ and only depends on the first-order inclusion probabilities.

The usual decomposition, called the two-phase approach in nonresponse theory, for $\Delta_{k\ell}$ is

$$\Delta_{k\ell} = I\{k = \ell\}\pi^B_k(1 - \pi^B_k)\pi^A_{k\ell} + \Delta^A_{k\ell}\pi^B_k\pi^B_\ell,$$

and the reverse approach is

$$\begin{aligned}\Delta_{k\ell} &= \Delta^A_{k\ell}\left[\pi^B_k\pi^B_\ell + I\{k = \ell\}\pi^B_k(1 - \pi^B_k)\right] + \pi^A_k\pi^A_\ell I\{k = \ell\}\pi^B_k(1 - \pi^B_k) \\ &= \Delta^A_{k\ell}\pi^B_k\pi^B_\ell + \pi^A_k(1 - \pi^A_k)I\{k = \ell\}\pi^B_k(1 - \pi^B_k) + \pi^A_k\pi^A_\ell I\{k = \ell\}\pi^B_k(1 - \pi^B_k) \\ &= \Delta^A_{k\ell}\pi^B_k\pi^B_\ell + I\{k = \ell\}\pi^B_k(1 - \pi^B_k)\pi^A_k.\end{aligned}$$

The results are equal but the computation of the usual decomposition is simpler and faster than the computation with the reverse one.

The usual decomposition for the ratio $\Delta_{k\ell}/\pi_{k\ell}$ is:

$$\begin{aligned}\frac{\Delta_{k\ell}}{\pi_{k\ell}} &= I\{k = \ell\}(1 - \pi^B_k) + \frac{\pi^B_k\pi^B_\ell\Delta^A_{k\ell}}{\pi^A_{k\ell}\left[\pi^B_k\pi^B_\ell + I\{k = \ell\}\pi^B_k(1 - \pi^B_k)\right]} \\ &= I\{k = \ell\}(1 - \pi^B_k) + \frac{\Delta^A_{k\ell}}{\pi^A_{k\ell}} + I\{k = \ell\}\pi^B_k(1 - \pi^A_k) - I\{k = \ell\}(1 - \pi^A_{k\ell}) \\ &= I\{k = \ell\}\pi^A_k(1 - \pi^B_k) + \frac{\Delta^A_{k\ell}}{\pi^A_{k\ell}}.\end{aligned}$$

The reverse decomposition, is straightforward:

$$\frac{\Delta_{k\ell}}{\pi_{k\ell}} = \frac{\Delta^A_{k\ell}}{\pi^A_{k\ell}} + I\{k = \ell\}\pi^A_k(1 - \pi^B_k).$$

This result is interesting. While the usual decomposition of $\Delta_{k\ell}$ is more direct than the reverse decomposition, the reverse decomposition of $\Delta_{k\ell}/\pi_{k\ell}$ is more straightforward than the usual one.

The reverse decomposition is used in the reverse approach to estimate the variance in the presence of nonresponse proposed by Fay (1991) and Shao and Steel (1999). This approach often simplifies the estimation of variance under nonresponse. The estimator of variance becomes

$$\widehat{\text{v}}(\widehat{Y}) = \sum_{k\in S}\sum_{\ell\in S}\frac{y_k y_\ell}{\pi_k\pi_\ell}\frac{\Delta^A_{k\ell}}{\pi^A_{k\ell}} + \sum_{k\in S}\frac{y_k^2}{(\pi_k)^2}\pi^A_k(1 - \pi^B_k).$$

If $\pi^B_k$ is the probability of response, it must obviously be estimated and plugged into the variance estimator.

## 5. Two-stage sampling

Consider the strongly invariant two-stage sampling design as defined in Beaumont and Haziza (2016). The population is partitioned into $M$ subsets $U_1, \dots, U_i, \dots, U_M$. A sample of primary units $S_1$ is a list of $m$ randomly selected primary units. The sample $S^A$ is the union of the secondary units in the $m$ primary units. All the units of the selected subsets are in sample $S^A = \bigcup_{i \in S_1} U_i$. Define $\pi_i^1$ the probability of selecting the primary unit $i$, $\pi_{ij}^1$ the probability of selecting primary units $i$ and $j$ together with $\pi_{ii}^1 = \pi_i^1$ and $\Delta_{ij}^1 = \pi_{ij}^1 - \pi_i^1 \pi_j^1$ ($i = 1, \dots, M$; $j = 1, \dots, M$). Then $\pi_k^A = \pi_i^1$ if unit $k$ is in subset $U_i$, $\pi_{k\ell}^A = I\{k, \ell \in U_i\}\pi_i^1 + I\{k \in U_i\}I\{\ell \in U_j\}\pi_{ij}^1$, where $i \neq j$ and $\Delta_{k\ell}^A = I\{k, \ell \in U_i\}\pi_i^1(1 - \pi_i^1) + I\{k \in U_i\}I\{\ell \in U_j\}(\pi_{ij}^1 - \pi_i^1\pi_j^1)$. Sample $S^B$ is a stratified sample where a stratum is a subset $U_i$ ($i = 1, \dots, M$). Subsamples are selected in each subset $U_i$ independently from one subset to another one. Then $\pi_{k\ell}^B = I\{k, \ell \in U_i\}\pi_{k\ell}^B + I\{k \in U_i\}I\{\ell \in U_j\}\pi_k^B \pi_\ell^B$ and $\Delta_{k\ell}^B = I\{k, \ell \in U_i\}(\pi_{k\ell}^B - \pi_k^B \pi_\ell^B)$. The two-stage sample is the intersection $S = S^A \cap S^B$.

The expanded estimator can be written

$$\widehat{Y} = \sum_{i \in S} \frac{y_k}{\pi_k} = \sum_{i \in S_1} \frac{\widehat{Y}_i}{\pi_i^1},$$

where $\widehat{Y}_i = \sum_{k \in U_i \cap S^B} y_k / \pi_k^B$ is an unbiased estimator of $Y_i = \sum_{k \in U_i} y_k$ when subset $U_i$ is selected at the first stage.

The usual decomposition is:

$$\Delta_{k\ell} = I\{k, \ell \in U_i\}\pi_i^1 \Delta_{k\ell}^B + \Delta_{k\ell}^A \pi_k^B \pi_\ell^B.$$

The reverse decomposition is:

$$\begin{aligned}
\Delta_{k\ell} &= I\{k, \ell \in U_i\}(\pi_i^1)^2 \Delta_{k\ell}^B + I\{k, \ell \in U_i\}\pi_i^1(1 - \pi_i^1)\pi_{k\ell}^B + I\{k \in U_i\}I\{\ell \in U_j\}\Delta_{ij}^1 \pi_k^B \pi_\ell^B \\
&= I\{k, \ell \in U_i\}\left[\pi_{k\ell}^B \pi_i^1 - \pi_k^B \pi_\ell^B (\pi_i^1)^2 + \pi_k^B \pi_\ell^B \pi_i^1 - \pi_k^B \pi_\ell^B \pi_i^1\right] + I\{k \in U_i\}I\{\ell \in U_j\}\Delta_{ij}^1 \pi_k^B \pi_\ell^B \\
&= I\{k, \ell \in U_i\}\Delta_{k\ell}^B \pi_i^1 + I\{k, \ell \in U_i\}\pi_k^B \pi_\ell^B \pi_i^1(1 - \pi_i^1) + I\{k \in U_i\}I\{\ell \in U_j\}\Delta_{ij}^1 \pi_k^B \pi_\ell^B \\
&= I\{k, \ell \in U_i\}\Delta_{k\ell}^B \pi_i^1 + \Delta_{k\ell}^A \pi_k^B \pi_\ell^B.
\end{aligned}$$

The usual decomposition of $\Delta_{k\ell}/\pi_{k\ell}$ is

$$\begin{aligned}
\frac{\Delta_{k\ell}}{\pi_{k\ell}} &= I\{k, \ell \in U_i\}\frac{\Delta_{k\ell}^B}{\pi_{k\ell}^B} + I\{k, \ell \in U_i\}\frac{\pi_i^1(1 - \pi_i^1)\pi_k^B \pi_\ell^B}{\pi_i^1 \pi_{k\ell}^B} + I\{k \in U_i\}I\{\ell \in U_j\}\frac{\Delta_{ij}^1 \pi_k^B \pi_\ell^B}{\pi_{ij}^1 \pi_k^B \pi_\ell^B} \\
&= I\{k, \ell \in U_i\}\left[\frac{\Delta_{k\ell}^B}{\pi_{k\ell}^B} - (1 - \pi_i^1)\frac{\Delta_{k\ell}^B}{\pi_{k\ell}^B}\right] + \frac{\Delta_{k\ell}^A}{\pi_{k\ell}^A} \\
&= \frac{\Delta_{k\ell}^A}{\pi_{k\ell}^A} + I\{k, \ell \in U_i\}\frac{\pi_i^1 \Delta_{k\ell}^B}{\pi_{k\ell}^B}
\end{aligned}$$

and the reverse one is

$$\frac{\Delta_{k\ell}}{\pi_{k\ell}} = \frac{\Delta_{k\ell}^A}{\pi_{k\ell}^A} + I\{k, \ell \in U_i\}\frac{(\pi_i^1)^2 \Delta_{k\ell}^B}{\pi_i^1 \pi_{k\ell}^B} = \frac{\Delta_{k\ell}^A}{\pi_{k\ell}^A} + I\{k, \ell \in U_i\}\frac{\pi_i^1 \Delta_{k\ell}^B}{\pi_{k\ell}^B}.$$

The usual decomposition leads to the following variance:

$$\begin{aligned}
\operatorname{var}\left(\sum_{i \in S_1} \frac{\widehat{Y}_i}{\pi_i}\right) &= \operatorname{E} \operatorname{var}\left(\sum_{i \in S_1} \frac{\widehat{Y}_i}{\pi_i^1} \,\middle|\, S_1\right) + \operatorname{var} \operatorname{E}\left(\sum_{i \in S_1} \frac{\widehat{Y}_i}{\pi_i^1} \,\middle|\, S_1\right) \\
&= \sum_{i=1}^M \frac{\operatorname{var}\left(\widehat{Y}_i\right)}{\pi_i^1} + \sum_{i=1}^M \sum_{j=1}^M \frac{Y_i Y_j}{\pi_i^1 \pi_j^1} \Delta_{ij}^1. \qquad (3)
\end{aligned}$$

The two terms are in general interpreted as the variance due to the first and the second stage respectively, but this interpretation is probably misleading. Indeed, if we use the reverse approach, we obtain a completely different variance decomposition:

$$\begin{aligned}
\operatorname{var}\left(\sum_{i \in S_1} \frac{\widehat{Y}_i}{\pi_i^1}\right) &= \operatorname{E} \operatorname{var}\left(\sum_{i \in S_1} \frac{\widehat{Y}_i}{\pi_i^1} \,\middle|\, S^B\right) + \operatorname{var} \operatorname{E}\left(\sum_{i \in S_1} \frac{\widehat{Y}_i}{\pi_i^1} \,\middle|\, S^B\right) \\
&= \sum_{i=1}^M \sum_{j=1}^M \frac{\operatorname{E}\left(\widehat{Y}_i \widehat{Y}_j\right)}{\pi_i^1 \pi_j^1} \Delta_{ij}^1 + \sum_{i=1}^M \operatorname{var}\left(\widehat{Y}_i\right) \qquad (4)
\end{aligned}$$

$$= \sum_{i=1}^{M}\sum_{j=1}^{M} \frac{Y_i Y_j}{\pi_i^1 \pi_j^1}\Delta_{ij}^1 + \sum_{i=1}^{M} \frac{\operatorname{var}\left(\widehat{Y}_i\right)}{\pi_i^1}(1-\pi_i^1) + \sum_{i=1}^{M} \operatorname{var}\left(\widehat{Y}_i\right)$$

$$= \sum_{i=1}^{M}\sum_{j=1}^{M} \frac{Y_i Y_j}{\pi_i^1 \pi_j^1}\Delta_{ij}^1 + \sum_{i=1}^{M} \frac{\operatorname{var}\left(\widehat{Y}_i\right)}{\pi_i^1}.$$

Expression (4) can be expanded because $\mathrm{E}\left(\widehat{Y}_i\widehat{Y}_j\right) = Y_iY_j + I(i=j)\operatorname{var}(\widehat{Y}_i)$.

The reverse decomposition can seem more intricate. Nevertheless, the unbiased estimator of the variance is

$$\widehat{\mathrm{v}}(\widehat{Y}) = \sum_{i\in S_1}\sum_{j\in S_1} \frac{\widehat{Y}_i\widehat{Y}_j}{\pi_i^1\pi_j^1}\frac{\Delta_{ij}^1}{\pi_{ij}^1} + \sum_{i\in S_1} \frac{\widehat{\mathrm{v}}\left(\widehat{Y}_i\right)}{\pi_i^1} \tag{5}$$

(see for instance Särndal et al., 1992 pp. 137–139). The first term of the estimator of variance is not an unbiased estimator of the second term of the usually used variance (3). The same happens with the second term of the estimated variance and the first term of variance (3). The reverse decomposition is not usually used to calculate the variance of the two-stage estimator. As noticed by Beaumont et al. (2015), the interesting thing about this decomposition is that clearly, estimator (5) is an unbiased estimator of Expression (4). Each term of estimator (5) estimates each term of variance (4) without bias. The reverse approach is then more appropriate to rapidly find an estimator of the variance.

The two terms of (3) are often interpreted as variances corresponding to the first and second stages. This interpretation is in fact misleading. The two terms of (3) are just the ones obtained by one of the two decompositions given by the law of total variance.

## 6. Detecting the briefest decomposition

For the two-phase sampling with Poisson sampling at the second phase and for the two-stage sampling, the usual variance decomposition is more straightforward than the reverse one. Oppositely, for the variance estimation, the reverse approach is simpler. A common aspect of the two sampling designs is that the element $\Delta_{k\ell}^B$ vanishes in many situations. For the two-phase sampling with Poisson sampling at the second phase, $\Delta_{k\ell}^B = I\{k=\ell\}\pi_k^B(1-\pi_k^B)$ is non-null only when $k=\ell$. For the two-stage sampling, $\Delta_{k\ell}^B = I\{k,\ell \in U_i\}(\pi_{k\ell}^B - \pi_k^B\pi_\ell^B)$ is non-null only when units $k$ and $\ell$ are in the same primary unit.

When most of the $\Delta_{k\ell}^B$ are null, the simplest decomposition occurs when $\Delta_{k\ell}^B$ is multiplied or divided by $\pi_{k\ell}^A$ and not by $\pi_k^A\pi_\ell^A$. For the decomposition of the variance, this is the case for the usual approach given in Expression (1). For the estimation of variance, this is the case for the reverse approach given in Expression (2). A convenient way of computing the variance can thus be misleading when estimating the variance.

## References


Beaumont, J.-F., Béliveau, A., Haziza, D., 2015. Clarifying some aspects of variance estimation in two-phase sampling. J. Surv. Stat. Methodol. 3 (4), 524–542.

Beaumont, J.-F., Haziza, D., 2016. A note on the concept of invariance in two-phase sampling designs. Surv. Methodol. 42 (2), 319–323.

Fay, R.E., 1991. A design-based perspective on missing data variance. In: Proceedings of the 1991 Annual Research Conference. US Census Bureau, pp. 429–440.

Horvitz, D.G., Thompson, D.J., 1952. A generalization of sampling without replacement from a finite universe. J. Amer. Statist. Assoc. 47, 663–685.

Narain, R.D., 1951. On sampling without replacement with varying probabilities. J. Indian Soc. Agricultural Statist. 3, 169–174.

Särndal, C.-E., 1992. Methods for estimating the precision of survey estimates when imputation has been used. Surv. Methodol. 18, 241–252.

Särndal, C.-E., Swensson, B., 1987. A general view of estimation for two phases of selection with applications to two-phase sampling and non-response. Internat. Statist. Rev. 55, 279–294.

Särndal, C.-E., Swensson, B., Wretman, J.H., 1992. Model Assisted Survey Sampling. Springer, New York.

Shao, J., Steel, P., 1999. Variance estimation for survey data with composite imputation and nonneglible sampling fractions. J. Amer. Statist. Assoc. 94, 254–265.