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/© 2017 Elsevier B.V. All rights reserved.

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

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 $\pi_k^A = \mathrm{E}(I_k^A) = \mathrm{pr}(k \in S^A)$ and $\pi_k^B = \mathrm{E}(I_k^B) = \mathrm{pr}(k \in S^B)$. The joint inclusion probabilities are $\pi_{k\ell}^A = \mathrm{E}(I_k^A I_\ell^A) = \mathrm{pr}(k, \ell \in S^A)$ and $\pi_{k\ell}^B = \mathrm{E}(I_k^B I_\ell^B) = \mathrm{pr}(k, \ell \in S^B)$, 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 $I_k = I_k^A I_k^B$, $\pi_k = \mathrm{pr}(k \in S) = \pi_k^A \pi_k^B$ and $\pi_{k\ell} = \mathrm{pr}(k, \ell \in S) = \pi_{k\ell}^A \pi_{k\ell}^B$. 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 $\mathrm{pr}(S^B = s^B | S^A) = \mathrm{pr}(S^B = s^B)$, 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 π_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

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

and can be unbiasedly estimated by

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

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

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

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

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

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:

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