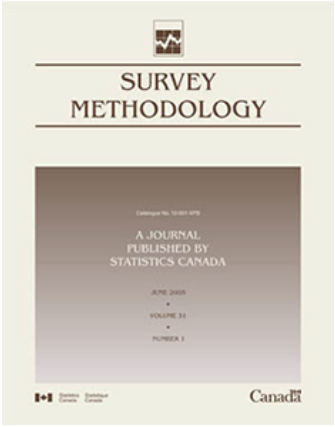
Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

A note on the concept of invariance in two-phase sampling designs



by Jean-François Beaumont and David Haziza

Release date: December 20, 2016

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at STATCAN.infostats-infostats.STATCAN@canada.ca

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-877-287-4369

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard table symbols

The following symbols are used in Statistics Canada publications:

. not available for any reference period
.. not available for a specific reference period
... not applicable
0 true zero or a value rounded to zero
0^s value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p} preliminary
^{r} revised
x suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E} use with caution
F too unreliable to be published
* significantly different from reference category ($p < 0.05$)

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, December 2016
Vol. 42, No. 2, pp. 319-323
Statistics Canada, Catalogue No. 12-001-X

A note on the concept of invariance in two-phase sampling designs

Jean-François Beaumont and David Haziza[1]

Abstract

Two-phase sampling designs are often used in surveys when the sampling frame contains little or no auxiliary information. In this note, we shed some light on the concept of invariance, which is often mentioned in the context of two-phase sampling designs. We define two types of invariant two-phase designs: strongly invariant and weakly invariant two-phase designs. Some examples are given. Finally, we describe the implications of strong and weak invariance from an inference point of view.

Key Words: Double expansion estimator; Horvitz–Thompson estimator; Strong invariance; Two-phase sampling; Weak invariance.

1 Introduction

Two-phase sampling designs are often used in surveys when the sampling frame contains little or no auxiliary information. It consists of first selecting a large sample from the population (typically using a rudimentary sampling design) in order to collect data on variables that are inexpensive to obtain and that are related to the characteristics of interest. The idea behind two-phase sampling is to create a pseudo-sampling frame richer in auxiliary information than the original sampling frame. Then, using the variables observed in the first phase, an efficient sampling procedure can be used to select a (typically small) subsample from the first-phase sample in order to collect the characteristics of interest. Two-phase sampling may also be helpful in a context of nonresponse as the set of respondents is often viewed as a second-phase sample.

We adopt the following notation: consider a population U of size N. A vector $\mathbf{I}_1$ is generated according to the sampling design $F(\mathbf{I}_1)$, where $\mathbf{I}_1 = (I_{11},\ldots,I_{1N})^{\mathrm{T}}$ denotes a vector of indicators such that I_{1i} is either equal to 0 or 1. The first-phase sample, denoted by s_1, is the set of population units for which $I_{1i} = 1$ and $n_1 = \sum_{i\in U} I_{1i}$, is the size of s_1. Then, a vector $\mathbf{I}_2$ is generated according to the sampling design $F(\mathbf{I}_2 \mid \mathbf{I}_1)$, where $\mathbf{I}_2 = (I_{21},\ldots,I_{2N})^{\mathrm{T}}$ denotes the vector of indicators such that I_{2i} is either equal to 0 or 1. The second-phase sample, denoted by s_2 is the set of population units for which both $I_{1i} = 1$ and $I_{2i} = 1$ and $n_2 = \sum_{i\in U} I_{1i}I_{2i}$ is the size of s_2. In practice, note that the indicators I_{2i} are not generated for the population units belonging to the set $U - s_1$. However, at least conceptually, nothing precludes defining these indicators for the units outside the first-phase sample.

Let $\pi_{1i} = P(I_{1i} = 1)$ and $\pi_{1ij} = P(I_{1i} = 1, I_{1j} = 1)$ be the first-order and second-order selection probabilities at the first-phase. Similarly, let $\pi_{2i}(\mathbf{I}_1) = P(I_{2i} = 1 \mid \mathbf{I}_1)$ and $\pi_{2ij}(\mathbf{I}_1) = P(I_{2i} = 1, I_{2j} = 1 \mid \mathbf{I}_1)$ be the first-order and second-order selection probabilities at the second-phase. Note that the (first-order and second-order) selection probabilities at the second-phase may depend on the realized sample s_1.

1. Jean-François Beaumont, International Cooperation and Corporate Statistical Methods Division, Statistics Canada, 100 Tunney's Pasture Driveway, R.H. Coats Building, 25th floor, Ottawa, Canada, K1A 0T6. E-mail: jean-francois.beaumont@canada.ca; David Haziza, Département de mathématiques et statistique, Université de Montréal, Montréal, Canada, H3C 3J7. E-mail: haziza@dms.umontreal.ca.

The paper is organized as follows. In Section 2, we define the concepts of weak and strong invariance and provide some examples. In Section 3, we discuss the implications of weak and strong invariance from an inferential point of view. In particular, we discuss the reverse decomposition of the variance in the case of a strongly invariant two-phase sampling design.

2 The concept of invariance

We distinguish the concept of strong invariance that may also be called distribution invariance from that of weak invariance that may also be called first-two-moment invariance.

Definition 1. *A two-phase sampling design is said to be strongly (or distribution) invariant provided that*

$$F(\mathbf{I}_2 \mid \mathbf{I}_1) = F(\mathbf{I}_2) \tag{2.1}$$

A consequence of Definition 1 is that $F(\mathbf{I}_1, \mathbf{I}_2) = F(\mathbf{I}_1)F(\mathbf{I}_2)$ and therefore, with a strongly invariant two-phase sampling design, the vector $\mathbf{I}_2$ can be generated prior to the vector $\mathbf{I}_1$. In practice, the concept of strong invariance is satisfied for only few two-phase sampling designs. A first example is Poisson sampling at the second phase. This covers the case of nonresponse, which is often viewed as a Poisson sampling design at the second phase. An other example is two-stage sampling. Both are described in greater detail below.

Example 1. *At the first phase, a sample s_1 is selected according to an arbitrary sampling design followed by Poisson sampling at the second phase, where the units selection probability $\pi_{2i}(\mathbf{I}_1)$ are set prior to sampling, which means that $\pi_{2i}(\mathbf{I}_1) = \pi_{2i}$ for $i \in U$. Since Poisson sampling is completely characterized by its first-order selection probabilities, we have $F(\mathbf{I}_2 \mid \mathbf{I}_1) = F(\mathbf{I}_2)$. As a result, this sampling design is strongly invariant. It can be implemented as follows: first, generate the vector $\mathbf{I}_2$ according to the Poisson sampling design $F(\mathbf{I}_2)$ and, independently, generate the vector $\mathbf{I}_1$ according to the design $F(\mathbf{I}_1)$.*

Example 2. *Two-stage cluster sampling can be described as follows: at the first stage, a sample of clusters is selected randomly from the population of clusters. Then, at the second stage, within each cluster selected at the first stage, a sample of elements is randomly selected. Note that, even in this case, the vector $\mathbf{I}_1$ is still defined at the element level, with its size N corresponding to the number of elements in the population. Under this set-up, the selection indicator for an element j within cluster i, I_{1ij}, is equal to 1 for all elements j within a selected cluster i. Therefore, two-stage sampling is a special case of two-phase sampling as described in Section 1. If the selection within clusters is independent of which clusters have been selected in the first phase, then we are in the presence of a strongly invariant two-stage cluster sampling design. This is satisfied if the selection of elements within clusters is independent of the selection of elements in any other cluster. A strongly invariant two-stage cluster sampling designs can be implemented by reversing the actual act of sampling: instead of sampling the clusters first, we begin by selecting the elements in each of the population clusters, and then sampling the clusters.*

Note that our definition of strong invariance for two-stage designs is slightly different from the one given in Särndal, Swensson and Wretman (1992, Chapter 4) because the latter restrict to clusters selected at the

first stage. However, for practical purposes, both definitions are essentially equivalent. We used Definition 1 rather the standard definition of Särndal et al. (1992) because the latter does not extend easily to the case of two-phase sampling.

Definition 2. *A two-phase sampling design is said to be weakly (or first-two-moment) invariant if*

$$\pi_{2i}(\mathbf{I}_1) = \pi_{2i} \quad \text{and} \quad \pi_{2ij}(\mathbf{I}_1) = \pi_{2ij} \quad i \in s_1, j \in s_1.$$

Clearly, a strongly invariant two-phase sampling design is weakly invariant but the opposite is not true. The next example describes a sampling design that is weakly invariant but not strongly invariant.

Example 3. *At the first phase, we select a sample, s_1, of size n_1, according to an arbitrary fixed-size sampling design. From s_1, we select a simple random sample without replacement, s_2, of size n_2, where n_2 is fixed prior to sampling. This two-phase sampling design is weakly invariant since $\pi_{2i} = n_2/n_1$, and $\pi_{2ij} = n_2(n_2-1)/n_1(n_1-1)$, which remain the same from one realization of $\mathbf{I}_1$ to another. However, it is not strongly invariant since it is not possible to generate $\mathbf{I}_2$ prior to $\mathbf{I}_1$ and meet the fixed-size sample size constraint for n_2. In fact, this would also be true for any fixed-size sampling design at the second phase satisfying $\pi_{2i}(\mathbf{I}_1) = \pi_{2i}$ and $\pi_{2ij}(\mathbf{I}_1) = \pi_{2ij}$.*

Finally, we describe a non-invariant two-phase sampling design.

Example 4. *At the first phase, we select a simple random sample without replacement, s_1, of size n_1, according to an arbitrary fixed-size sampling design. For every $i \in s_1$, we record an auxiliary variable x. From s_1, a second-phase sample, s_2, of fixed size n_2, is selected using an inclusion probability proportional-to-size procedure. In this case, we have*

$$\pi_{2i}(\mathbf{I}_1) = \frac{n_2 x_i}{\sum_{i \in U} x_i I_{1i}}.$$

Clearly, the inclusion probability of unit i in s_2 vary from one realization of $\mathbf{I}_1$ to another. Since $\pi_{2i}(\mathbf{I}_1)$ is a function of $\mathbf{I}_1$, it is known only after the first-phase sample s_1 is actually realized.

3 Implications of the invariance property

3.1 Weak invariance

For an arbitrary two-phase sampling design, the inclusion probability of unit i, $\pi_i, i \in s_1$, is generally unknown and is defined as

$$\begin{aligned} \pi_i &= \mathrm{E}(I_{1i} I_{2i}) \\ &= \mathrm{E}\{I_{1i}\mathrm{E}(I_{2i} \mid \mathbf{I_1})\} \\ &= \sum_{\mathbf{i}_1 : i_{1i}=1} \pi_{2i}(\mathbf{I_1}) P(\mathbf{I}_1 = \mathbf{i}_1), \end{aligned} \tag{3.1}$$

where $\mathbf{i}_1$ denotes a realisation of the random vector $\mathbf{I}_1$. Therefore, the π_i's are generally unknown because they require the knowledge of $P(\mathbf{I}_1 = \mathbf{i}_1)$ for every possible $\mathbf{I}_1$ (in many cases, we do) but also of $\pi_{2i}(\mathbf{I}_1)$ for every $\mathbf{I}_1$. The latter are generally unknown because $\pi_{2i}(\mathbf{I}_1)$ may depend on the outcome of phase 1. However, if the sampling design is weakly invariant, then $\pi_{2i}(\mathbf{I}_1) = \pi_{2i}$ and (3.1) reduces to

$$\pi_i = \pi_{2i} \sum_{\mathbf{i}_1:\, i_{1i}=1} P(\mathbf{I}_1 = \mathbf{i}_1) = \pi_{1i}\pi_{2i}. \tag{3.2}$$

Suppose that we are interested in estimating the population total $t_y = \sum_{i \in U} y_i$. Since the π_i's are generally unknown, the Horvitz-Thompson estimator of t_y,

$$\hat{t}_{HT} = \sum_{i \in s_2} \pi_i^{-1} y_i,$$

cannot be used, in general. Instead, it is common practice to use the double expansion estimator

$$\hat{t}_{DE} = \sum_{i \in s_2} \pi_{1i}^{-1} \pi_{2i}(\mathbf{I}_1)^{-1} y_i.$$

In general, both $\hat{t}_{HT}$ and $\hat{t}_{DE}$ differ. However, for weakly invariant two-phase designs, it is clear from (3.2), that both are identical.

3.2 Strong invariance

Let θ be a finite population parameter and $\hat{\theta}$ be an estimator of θ. The total variance of $\hat{\theta}$ can be expressed as

$$V(\hat{\theta}) = VE(\hat{\theta} \mid \mathbf{I}_1) + EV(\hat{\theta} \mid \mathbf{I}_1). \tag{3.3}$$

Decomposition (3.3) is often called the two-phase decomposition of the variance; e.g., Särndal et al. (1992). If the two-phase sampling design is strongly invariant, the total variance of $\hat{\theta}$ can alternatively be decomposed as

$$V(\hat{\theta}) = EV(\hat{\theta} \mid \mathbf{I}_2) + VE(\hat{\theta} \mid \mathbf{I}_2). \tag{3.4}$$

The decomposition (3.4) is often called the reverse decomposition of the variance as the order of sampling is reversed, which can only be justified provided the two-phase design is strongly invariant. The decomposition (3.4) cannot be used in the case of weakly invariant two-phase design as the vector $\mathbf{I}_2$ cannot be generated prior to the vector $\mathbf{I}_1$. The reverse decomposition was studied in the context of nonresponse by Fay (1991), Shao and Steel (1999) and Kim and Rao (2009), among others. In a nonresponse context, assuming that the units respond independently of one another, the set of respondents can be viewed as a second-phase sample selected according to Poisson sampling with unknown inclusion probabilities, called response probabilities. If the latter remain the same from one realization of the sample to another, we are essentially in the presence of a strongly invariant two-phase sampling design. Decomposition (3.4) can be

used to justify simplified variance estimators for two-phase sampling designs; see Beaumont, Béliveau and Haziza (2015).

Acknowledgements

The authors are grateful to an Associate Editor and a reviewer for their comments and suggestions, which improved the quality of this paper. David Haziza's research was funded by a grant from the Natural Sciences and Engineering Research Council of Canada.

References

Beaumont, J.-F., Béliveau, A. and Haziza, D. (2015). Clarifying some aspects of variance estimation in two-phase sampling. *Journal of Survey Statistics and Methodology*, 3, 524-542.

Fay, R.E. (1991). A design-based perspective on missing data variance. *Proceedings of the 1991 Annual Research Conference,* US Bureau of the Census, 429-440.

Kim, J.K., and Rao, J.N.K. (2009). A unified approach to linearization variance estimation from survey data after imputation for item nonresponse. *Biometrika*, 96, 917-932.

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

Shao, J., and Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94, 254-265.