Estimation of a proportion under a certain two-stage sampling design

Danutė Krapavickaitė

Institute of Mathematics and informatics, Lithuania
Statistics Lithuania, Lithuania
e-mail: krapav@ktl.mii.lt

Abstract

The aim of this paper is to demonstrate with examples that the design-based estimator for the proportion of the first-stage sampling elements having associated at least one second-stage element with the attribute of interest using the two-stage sampling design is biased. The situation is encountered in the Adult Education Survey (AES) when estimating the share of individuals in non-formal education involved in job-related learning activities.

1 Introduction

A new problem related to the estimation of a proportion has arisen in the Adult Education Survey (hereinafter referred to as "the AES"). The parameter of interest is the share of individuals in non-formal education involved in job-related learning activities. In the paper, the problem is described in the general framework, and it is shown by the example that the design-based estimator for this parameter is biased. The direction for further research is drawn.

2 Population and parameters

Let us denote by $U_1 = \{u_1, u_2, ..., u_N\}$ the population of the units, to each of which a cluster of subunits of size M_i, i=1,2,…,N, is associated. Thus, the population of all subunits U_2 consists of $M = M_1 + ... + M_N$ elements. Suppose some of the subunits have an attribute of interest, and some of them do not have it. Let us introduce a study variable z in population U_1 with value $z_i = 1$, if there is at least one subunit among M_i subunits associated with unit u_i, and $z_i = 0$ otherwise, i=1,2,…,N.

Then the number of units in the population having associated at least one subunit with the attribute is equal to the total of variable z:

$$t_z = \sum_{i=1}^{N} z_i \,. \tag{0}$$

The share (proportion) of the units in U_1 having associated at least one subunit with the attribute is equal to the mean of variable z: $\mu_z = t_z / N$. Let us consider the estimation of parameters t_z and μ_z from the survey data.

3 Sample and estimators

The sample design of subunits constituting population U_2 can be described by a 2-stage sampling design with some probabilistic sample $\mathbf{s}_I$ of units in U_1 at the first stage and a simple random sample $\mathbf{s}_{IIi}$ of m_i subunits in the cluster associated with unit u_i (or all of them if their number is less than m_i) at the second stage:

$$\mathbf{s} = \bigcup_{i \in \mathbf{s}_I} \mathbf{s}_{IIi} \subset U_2 \,.$$

At the second stage, sample $\mathbf{s}_{IIi}$ size m_i can be any positive number, but for simplicity without loosing the generality let us consider

$$m_i = M_i, \quad \text{for } M_i = 0, 1, 2, \text{ and } m_i = 3 \text{ for } M_i \geq 3 \text{ for } i \in \mathbf{s}_I \,.$$

Let us denote by the $d_i = \dfrac{1}{\pi_i}$ the first stage sampling design weight with

$$\pi_i = P\left(\mathbf{s}_I : i \in \mathbf{s}_I\right), \quad i \in \mathbf{s}_I \,.$$

The design-based estimator suggested for the number of units having associated at least one subunit with the attribute is

$$\hat{t}_z = \sum_{k:k \in \mathbf{s}} d_k \hat{z}_k \tag{1}$$

where $\hat{z}_k$ is the design-based estimator of z_k: $\hat{z}_k = 1$ if at least one subunit with the attribute belongs to $\mathbf{s}_{IIi}$, and $\hat{z}_k = 0$ otherwise. For the share of units having associated at least one subunit with the attribute, the suggested design-based estimator is

$$\hat{\mu}_z = \hat{t}_z / N \,. \tag{2}$$

This estimator is usually used in the pilot AES of statistical offices.

Hypothesis: estimators (1) and (2) are biased, e.g. $E\hat{t}_z \neq t_z$, $E\hat{\mu}_z \neq \mu_z$, the expectation is taken here with respect to the two-stage sampling design.

The following example confirms the hypothesis.

4 Example

Let us study a small population $U_1 = \{u_1, u_2, u_3\}$ consisting of $N = 3$ units. Unit u_1 is associated with one subunit without an attribute, denoted by *nonattr*; unit u_2 is associated with one subunit with the attribute, denoted by *attrib*; unit u_3 is associated with two subunits: one with an attribute (*attrib*) and one without an attribute (*nonattr*). For this population, the number of units with the attribute and their share is equal to

$$t_z = z_1 + z_2 + z_3 = 0 + 1 + 1 = 2, \quad \mu_z = 2/3.$$

Let us draw the first-stage simple random sample $\mathbf{s}_I$ of *n*=2 elements from population U_1. The possible samples according to this sampling design and their sampling design probabilities are:

$$\mathbf{s}_{I1} = (u_1, u_2), \quad \mathbf{s}_{I2} = (u_1, u_3), \quad \mathbf{s}_{I3} = (u_2, u_3),$$

$$P(\mathbf{s}_{I1}) = P(\mathbf{s}_{I2}) = P(\mathbf{s}_{I3}) = \frac{1}{3};$$

Let us simplify the sample design taking for the sample of subunits $m_i = M_i$ for $M_i = 0$, and $m_i = 1$ for $M_i \geq 1$. The second stage sampling design probabilities are as follows:

$$P(nonattr \mid u_1) = 1, \quad P(attrib \mid u_1) = 0,$$
$$P(nonattr \mid u_2) = 0, \quad P(attrib \mid u_2) = 1,$$
$$P(nonattr \mid u_3) = P(attrib \mid u_3) = \frac{1}{2}.$$

Let us estimate t_z in these samples:

$$\mathbf{s}_{I1} = (u_1, u_2): \quad \hat{t}_z^{(1)} = \frac{N}{n}(z_1 + z_2) = \frac{3}{2}(0 + 1) = \frac{3}{2}, \quad \hat{\mu}_z^{(1)} = \frac{1}{2}.$$

For the element u_3, we estimate $\hat{z}_3 = 1$ if the unit with the attribute is selected for the second-stage sample, and $\hat{z}_3 = 0$ otherwise.

$$\mathbf{s}_{I2} = (u_1, u_3): \text{ if } \mathbf{s}_{II3} = \{nonattr\} \text{ then } \quad \hat{t}_z^{(2)} = \frac{N}{n}(z_2 + \hat{z}_3) = \frac{3}{2}(0 + 0) = 0, \quad \hat{\mu}_z^{(2)} = 0,$$

$$\text{if } \mathbf{s}_{II3} = \{attrib\} \text{ then } \hat{t}_z^{(3)} = \frac{N}{n}(z_2 + \hat{z}_3) = \frac{3}{2}(0+1) = \frac{3}{2},\ \hat{\mu}_z^{(3)} = \frac{1}{2},$$

$$\mathbf{s}_{I3} = (u_2, u_3): \text{if } \mathbf{s}_{II3} = \{nonattr\} \text{ then } \hat{t}_z^{(4)} = \frac{N}{n}(z_2 + \hat{z}_3) = \frac{3}{2}(1+0) = \frac{3}{2},\ \hat{\mu}_z^{(4)} = \frac{1}{2},$$

$$\text{if } \mathbf{s}_{II3} = \{attrib\} \text{then } \hat{t}_z^{(5)} = \frac{N}{n}(z_2 + \hat{z}_3) = \frac{3}{2}(1+1) = 3,\ \hat{\mu}_z^{(5)} = 1.$$

Let us calculate the expectation of $\hat{t}_z$ with respect to the sampling design:

$$E\hat{t}_z = \hat{t}_z^{(1)} P(s_{I1}) + \left(\hat{t}_z^{(2)} P(nonattr \mid u_3) + \hat{t}_z^{(3)} P(attrib \mid u_3)\right) P(s_{I2})$$

$$+ \left(\hat{t}_z^{(4)} P(nonattr \mid u_3) + \hat{t}_z^{(5)} P(attrib \mid u_3)\right) P(s_{I3})$$

$$= \frac{3}{2}\frac{1}{3} + \left(0 \cdot \frac{1}{2} + \frac{3}{2}\frac{1}{2}\right)\frac{1}{3} + \left(\frac{3}{2}\frac{1}{2} + 3\frac{1}{2}\right)\frac{1}{3} = \frac{1}{2} + \frac{1}{4} + \frac{3}{4} = \frac{3}{2} \neq t_z = 2.$$

It means that estimator $\hat{t}_z$ is biased. Consequently,

$$E\hat{\mu}_z = \frac{E\hat{t}_z}{N} = \frac{1}{2} \neq \mu_z = \frac{2}{3},$$

and estimator $\hat{\mu}_z$ of the proportion of the units with the attribute is also biased.

It is clear by intuition that estimator (1) is underestimating the true number of the units with the attribute because there are possible cases when the sampled unit based on the sampled subunits is classified as without an attribute when in reality it is a non-sampled subunit with an attribute associated with it, but there are no possible cases when the sampled unit is classified as being associated to the subunit with an attribute when in reality it is not so.

5 Possible direction for the following research

The other kind of estimator for the proportion of the first-stage sampling elements under the two-stage sampling design is as follows. Some auxiliary assumptions about the population of secondary elements have to be stated.

Let us suppose the number M_i of subunits associated with unit u_i is fixed and known, but the number of subunits with attribute X_i is random, $0 \le X_i \le M_i$, $i = 1,2,...,N$. Let us define probabilities

$$p_{M_i}(k) = P(X_i = k \mid M_i), \quad \sum_{k=0}^{M_i} p_{M_i}(k) = 1, \quad i = 1,2,...,N.$$

The number of the sampled subunits with attribute, Y_i, is also random, $0 \le Y_i \le \min(3, X_i)$. We have the following relationship:

$$\hat{z}_i = \begin{cases} 1, \_ \text{if } \_ Y_i > 0 \Leftrightarrow X_i > 0, \\ 0, \_ \text{if } \_ Y_i = 0 \Leftrightarrow \begin{cases} X_i > 0, \\ X_i = 0. \end{cases} \end{cases}$$

Let us investigate a random event $X_i > 0$. Let us denote the random variable

$$J_i = \begin{cases} 1, \_ \text{if } \ X_i > 0, \\ 0, \_ \text{if } \ X_i = 0, \end{cases}$$

$i = 1,2,..., N$. Variable J_i obtains 1 with the probability

$$p_i = P(J_i = 1) = P(X_i > 0) = P(Y_i > 0) + P(X_i > 0 \mid Y_i = 0) P(Y_i = 0) \tag{3}$$

Let us calculate this probability. If $M_i \le 3$ then Y_i coincides with X_i and $P(X_i > 0 \mid Y_i = 0) = 0$, and

$$p_i = P(Y_i > 0) = 1 - P(Y_i = 0) = 1 - p_{M_i}(0).$$

If $M_i > 3$ then

$$P(X_i > 0 \mid Y_i = 0) = 1 - P(X_i = 0 \mid Y_i = 0), \tag{4}$$

$$P(X_i = 0 \mid Y_i = 0) = \frac{P(X_i = 0, Y_i = 0)}{P(Y_i = 0)} = \frac{P(Y_i = 0 \mid X_i = 0) P(X_i = 0)}{P(Y_i = 0)} = \frac{p_{M_i}(0)}{P(Y_i = 0)},$$

$$P(Y_i = 0) = \sum_{k=0}^{M_i} P(Y_i = 0 \mid X_i = k) p_{M_i}(k)$$

$$P(Y_i = 0 \mid X_i = k) = \begin{cases} \dfrac{C_k^0 C_{M_i - k}^3}{C_{M_i}^3} = \left(1 - \dfrac{k}{M_i}\right)\left(1 - \dfrac{k}{M_i - 1}\right)\left(1 - \dfrac{k}{M_i - 2}\right), \_\_ k = 0,1,..., M_i - 3, \\ 0, \_\_ k = M_{i-2}, M_{i-1}, M_i. \end{cases}$$

Hence from (4) we get

$$P(X_i > 0 \mid Y_i = 0) = 1 - \frac{p_{M_i}(0)}{\sum_{k=0}^{M_i-3}\left(1-\frac{k}{M_i}\right)\left(1-\frac{k}{M_i-1}\right)\left(1-\frac{k}{M_i-2}\right)p_{M_i}(k)}. \quad (5)$$

Inserting (5) into (3) we can calculate p_i.

Let us introduce a new estimator for the number of units associated with the subunits with attribute:

$$\hat{\hat{t}}_z = \sum_{k \in s} d_k \mathbf{E} J_k \ . \quad (6)$$

Proposition. *Suppose probabilities* $p_{M_i}(k),\ k = 0,1,...,M_i,\ i = 1,...,N,\ M_i > 0$ *are fixed and known. Then*

- *the expectation of estimator* $\hat{\hat{t}}_z$ (4) *under the sampling design is* $\mathbf{E}\hat{\hat{t}}_z = \sum_{k=1}^{N} p_k$,
- *its variance* $Var\left(\hat{\hat{t}}_z\right) = \sum_{k=1}^{N} \frac{1-\pi_k p_k}{\pi_k} p_k + \sum_{\substack{k,l=1 \\ k \neq l}}^{N} (\pi_{kl} - \pi_k \pi_l) \frac{p_k p_l}{\pi_k \pi_l}$,
- *estimator of variance* $\hat{Var}\left(\hat{\hat{t}}_z\right) = \sum_{k=1}^{N} \frac{1-\pi_k p_k}{\pi_k^2} p_k + \sum_{\substack{k,l=1 \\ k \neq l}}^{N} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{p_k p_l}{\pi_k \pi_l}$ *is unbiased.*

In practice, probabilities $p_{M_i}(k)$ are not known, and they have to be estimated. Then the estimator of the total and its variance becomes more complicated. The example of these approximate probabilities of the Lithuanian AES for the share of job-related learning activities in non-formal education is shown in the figure below.