Estimation of Some Proportion in a Clustered Population

D. Krapavickaitė

Institute of Mathematics and Informatics
Akademijos str. 4, LT-08663 Vilnius, Lithuania
krapav@ktl.mii.lt

Received: 2008-12-29 **Revised:** 2009-07-01 **Published online:** 2009-11-11

Abstract. The aim of this paper is to show that the traditional design-based estimator for the proportion of population units, associated with at least one subunit having an attribute of interest using the two-stage sampling design, is biased. We face such a situation in the Adult Education Survey of official statistics of the European countries when estimating the share of individuals in non-formal education, involved in job-related learning activities. The alternative *design and model*-based estimators are proposed.

Keywords: finite population, proportion, two-stage sampling design, bias.

1 Introduction

A new problem related to the estimation of a proportion has arisen in the Adult Education Survey of the European countries ([1–3], hereinafter referred to as “AES”). The parameter of interest is the share of individuals in non-formal education involved in job-related learning activities. According to the sampling design, the individuals included into the first stage AES sample, present the second stage simple random sample of size $m \leq 3$ of the learning activities of non-formal education in which they have been involved during a year. Some of them are job-related, but some of them are not job-related. Even if there are no job-related learning activities in the sample, they can occur among non-sampled ones, and have to be taken into account.

The problem has arisen in practical work. The author has not met any similar problem solved, or at least touched, in the literature. In the paper, the problem is described in the general framework, and it is shown by an example that the design-based estimator of this parameter is biased, and the size of bias is demonstrated by an example.

In order to take into account possible non-sampled job-related learning activities for sampled individuals assumption on the distribution of the number of such learning activities for each individual is made. Alternative design and model-based estimators are proposed. Their application is shown by Examples 5, 6.

2 Population and parameters

Let us denote by $U_1 = \{u_1, u_2, \ldots, u_N\}$, (or $U_1 = \{1, 2, \ldots, N\}$ without restriction of generality) population of units, with each of which a cluster of subunits U_{2i} of size M_i, $i = 1, 2, \ldots, N$, is associated. Thus, the population of all subunits U_2 consists of $M = M_1 + \ldots + M_N$ elements: $U_2 = \cup_{i=1}^N U_{2i}$. Suppose that some of the subunits have an attribute of interest, some of them do not have it. Let us introduce an attribute indicator – a study variable z – in population U_1 with the value $z_i = 1$, if there is at least one subunit with the attribute among M_i subunits associated with the unit u_i, and $z_i = 0$, otherwise, $i = 1, 2, \ldots, N$. Then the number of units in the population associated with at least one subunit having an attribute is equal to the total of the variable z:

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

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

3 Sample and the usual estimator

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

$$\mathbf{s} = \bigcup_{i \in \mathbf{s}_I} \mathbf{s}_{IIi} \subset U_2, \quad \mathbf{s}_{IIi} \subset U_{2i}.$$

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

$$m_i = \begin{cases} M_i, & \text{if } M_i = 0, 1, 2, \\ 3, & \text{if } M_i \geq 3, \end{cases}$$

for $i \in \mathbf{s}_I$. This is the case in the Lithuanian AES. Denote by $d_i = 1/\pi_i$ the first stage sampling design weight with the first and second order inclusion probabilities

$$\pi_i = P(\mathbf{s}_I\colon i \in \mathbf{s}_I) > 0,$$
$$\pi_{ii} = \pi_i, \ \ \pi_{ij} = P(\mathbf{s}_I\colon i \in \mathbf{s}_I \ \& \ j \in \mathbf{s}_I) > 0, \quad i, j, \in U_1, \ \ i \neq j.$$

The Horvitz-Thompson estimator of the population total t_z of the variable z

$$\sum_{i \in \mathbf{s}_I} \frac{z_i}{\pi_i}$$

cannot be used to estimate the number of units, associated at least one subunit with the attribute in the population U_1 because for $m_i < M_i$ the values z_i may be not observable.

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

$$\hat{t}_z = \sum_{i:\ i \in \mathbf{s}_I} d_i \widehat{z}_i, \tag{2}$$

where $\widehat{z}_i$ is the design-based estimator of z_i:

$$\widehat{z}_i = \begin{cases} 1, & \text{if at least one subunit with the attribute belongs to } \mathbf{s}_{IIi}, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

For the share of units associated with at least one subunit having an attribute, the suggested design-based estimator is

$$\widehat{\mu}_z = \hat{t}_z / N. \tag{4}$$

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

Hypothesis: estimators (2) and (4) are biased, e. g., $\mathbf{E}\hat{t}_z \neq t_z$, $\mathbf{E}\widehat{\mu}_z \neq \mu_z$, the expectation is taken here with respect to the two-stage sampling design.

The following examples confirm the hypothesis.

4 Bias

We show by example 1 that using design-based approach to the problem we unavoidably obtain the biased estimator of a parameter.

Example 1 (Existence of a bias of estimator $\hat{t}_z$)**.** Let us study a Small Population $U_1 = \{u_1, u_2, u_3\}$ consisting of $N = 3$ units. The unit u_1 is associated with one subunit without an attribute, denoted as *nonattr*; the unit u_2 is associated with one subunit with an attribute, denoted as *attrib*; the 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 an 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 realizations of the sample according to this sampling design and their sampling 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

$$m_i = \begin{cases} M_i, & \text{for } M_i = 0, \\ 1, & \text{for } M_i \geq 1. \end{cases}$$

The second stage sampling design probabilities are as follows:

$$\begin{aligned} &P(nonattr|u_1) = 1, \quad P(attrib|u_1) = 0, \\ &P(nonattr|u_2) = 0, \quad P(attrib|u_2) = 1, \\ &P(nonattr|u_3) = P(attrib|u_3) = \frac{1}{2}. \end{aligned}$$

Let us estimate t_z using estimator (2) and the data of these samples:

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

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

$$\begin{aligned} &\qquad \mathbf{s}_{I2} = (u_1, u_3): \\ &\text{if } \ \mathbf{s}_{II3} = \{nonattr\}, \quad \text{then } \ \hat{t}_z^{(2)} = \frac{N}{n}(z_1 + \widehat{z}_3) = \frac{3}{2}(0+0) = 0,\ \ \widehat{\mu}_z^{(2)} = 0, \\ &\text{if } \ \mathbf{s}_{II3} = \{attrib\}, \qquad \text{then } \ \hat{t}_z^{(3)} = \frac{N}{n}(z_1 + \widehat{z}_3) = \frac{3}{2}(0+1) = \frac{3}{2},\ \widehat{\mu}_z^{(3)} = \frac{1}{2}, \\ &\qquad \mathbf{s}_{I3} = (u_2, u_3): \\ &\text{if } \ \mathbf{s}_{II3} = \{nonattr\}, \quad \text{then } \ \hat{t}_z^{(4)} = \frac{N}{n}(z_2 + \widehat{z}_3) = \frac{3}{2}(1+0) = \frac{3}{2},\ \widehat{\mu}_z^{(4)} = \frac{1}{2}, \\ &\text{if } \ \mathbf{s}_{II3} = \{attrib\}, \qquad \text{then } \ \hat{t}_z^{(5)} = \frac{N}{n}(z_2 + \widehat{z}_3) = \frac{3}{2}(1+1) = 3,\ \ \widehat{\mu}_z^{(5)} = 1. \end{aligned}$$

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

$$\begin{aligned} \mathbf{E}\hat{t}_z &= \hat{t}_z^{(1)} P(\mathbf{s}_{I1}) + \big(\hat{t}_z^{(2)} P(nonattr|u_3) + \hat{t}_z^{(3)} P(attrib|u_3)\big) P(\mathbf{s}_{I2}) \\ &\quad + \big(\hat{t}_z^{(4)} P(nonattr|u_3) + \hat{t}_z^{(5)} P(attrib|u_3)\big) P(\mathbf{s}_{I3}) \\ &= \frac{3}{2}\frac{1}{3} + \left(0 \cdot \frac{1}{2} + \frac{3}{2}\frac{1}{2}\right)\frac{1}{3} + \left(\frac{3}{2}\frac{1}{2} + 3\frac{1}{2}\right)\frac{1}{3} \\ &= \frac{1}{2} + \frac{1}{4} + \frac{3}{4} = \frac{3}{2} \neq t_z = 2. \end{aligned}$$

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

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

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

It is clear by intuition that estimator (2) underestimates the true number of the units with an attribute, because the cases are possible, where a sampled unit based on the sampled subunits is classified as without an attribute (cases $\mathbf{s}_{I2}$, $\mathbf{s}_{I3}$ with $\mathbf{s}_{II3} = \{nonattr\}$), while in reality there exists a non-sampled subunit with an attribute associated with it. On the other hand, there are no possible cases where a sampled unit is classified as being associated with the subunit with an attribute, as in reality it is not so.

The situation is visualized in Fig. 1. Big circles mean units, all the small circles mean subunits; the small black circles mean subunits with an attribute. A subunit joined with the unit means a sampled subunit. "+" means $\widehat{z}_i = 1$, "–" means $\widehat{z}_i = 0$, "?" means $\widehat{z}_i = 0$ and a source of bias. Fig. 1 show that estimator (2) underestimates the number of units with an attribute.

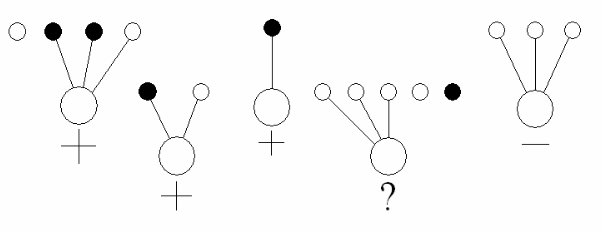

Fig. 1. Sample of subunits.

Example 2 (Size of a bias)**.** Let us write an expression for the bias in some special case. Denote by X_i the number of subunits with the attribute associated with the ith unit, and by Y_i the number of sampled subunits with the attribute. Consider $M_i = M$, $m_i = m$, $X_i = k > 0$ are fixed numbers for all $i = 1, \dots, N$. It means that each population unit is associated with exactly k subunits having attributes. The numbers of sampled subunits with the attribute Y_i are independent identically distributed random variables. Suppose they have the same distribution as the random variable Y.

For self-weighting 1st stage sampling design, the population size can be expressed in such a way:

$$N = t_z = NP(Y = 0) + NP(Y > 0).$$

For the estimator $\hat{t}_z$ in our example we have:

$$\mathbf{E}\hat{t}_z = \mathbf{E} \sum_{i \in \mathbf{s}_I} d_i \widehat{z}_i = NP(Y > 0).$$

From the last two expressions we obtain

$$Bias(\hat{t}_z) = \mathbf{E}\hat{t}_z - t_z = -NP(Y = 0).$$

For our sampling design we calculate:

$$P(Y=0)=P(Y=0|X=k)=\frac{C_k^0C_{M-k}^m}{C_M^m}$$
$$=\left(1-\frac{k}{M}\right)\left(1-\frac{k}{M-1}\right)\dots\left(1-\frac{k}{M-m+1}\right).$$

Then

$$Bias(\hat{t}_z)=-N\left(1-\frac{k}{M}\right)\left(1-\frac{k}{M-1}\right)\dots\left(1-\frac{k}{M-m+1}\right).$$

Some numerical values of the bias in the case the parameters close to the Lithuanian AES ones are given in Table 1.

Table 1. Values of the $Bias(\hat{t}_z)$ for the case $N = 2\,000\,000$, $M = 10$, $m = 3$, $k = 1,2,\dots,M$

k	$Bias(\hat{t}_z)$	$\big(Bias(\hat{t}_z)/N\big)100$ (%)
1	1 400 000	70
2	933 333	47
3	583 333	29
4	333 333	17
5	166 667	8
6	66 667	3
7	16 667	1
8, 9, 10	0	0

We see that the higher the number of subunits with an attribute in the population, the lower the bias of estimator (2) is for the number of units associated with the subunits with an attribute. Bias is unavoidable when $M-k\geq 3$.

In order to adjust estimator to the bias, we introduce a superpopulation model for the distribution of the number of subunits with the attribute.

5 Alternative estimators

We propose some *design and model*-based estimator for the proportion of the first-stage sampling elements associated with the subunits with an attribute under the two-stage sampling design. Some auxiliary assumptions on the superpopulation of subunits have to be stated.

1. Suppose that the number M_i of subunits associated with the unit u_i is fixed and known, but the number of subunits X_i with an attribute is random, $0\leq X_i\leq M_i$, $i=1,\dots,N$. Let us define the probabilities

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

We consider these probabilities (distribution of the variable X_i) to be known.

2. The values of the study variable z become random because they depend on the values of the random variables X_i, $i = 1, \dots, N$. The population total t_z is also random.

3. The number of sampled subunits with an attribute, Y_i, is random, $0 \le Y_i \le \min(3, X_i)$. The values Y_i, Y_j are independent for $i \ne j$.

For estimator (3) of the value z_i of the attribute indicator (study variable) z, the following relationship is valid:

$$\hat{z}_i = \begin{cases} 1, & \text{if } Y_i > 0 \;\Leftrightarrow\; X_i > 0,\ z_i = 1, \\ 0, & \text{if } Y_i = 0 \;\Leftrightarrow\; \begin{cases} X_i > 0, & z_i = 1, \\ X_i = 0, & z_i = 0. \end{cases} \end{cases} \tag{5}$$

Conditional distribution of Y_i under the condition that the value of X_i is known, is also known due to the known simple random sampling design of subunits.

Taking into account (5), we obtain an expression for the probability, denoted by p_i, that the variable z_i obtains value 1:

$$\begin{aligned} &p_i = P(z_i = 1) = P(X_i > 0), \quad \text{or, equivalently,} \\ &p_i = P(Y_i > 0) + P(X_i > 0 | Y_i = 0) P(Y_i = 0). \end{aligned} \tag{6}$$

In order to obtain the new estimator of the total t_z, the value of $\hat{z}_i$ in (2) is changed by p_i from (6).

The expectation of t_z with respect to the distribution of X_i, $i = 1, \dots, N$, is

$$\mathbf{E}_X t_z = \sum_{i=1}^{N} \mathbf{E}_X z_i = \sum_{i=1}^{N} P(X_i > 0) = \sum_{i=1}^{N} p_i.$$

We are going to estimate this expectation.

Estimator A. Let us estimate t_z by

$$\hat{\hat{t}}_z^{(A)} = \sum_{i \in \mathbf{s}_I} d_i p_i. \tag{7}$$

This is a Horvitz-Thompson type estimator of the total of the study variable with the values p_i, $i = 1, 2, \dots, N$, and we use further the well known result [4, p. 43], for this estimator.

Proposition A. *Suppose the probabilities $p_{M_i}(k)$, $k = 1, 2, \dots, M_i$, $i = 1, 2, \dots, N$, $M_i > 0$, are fixed and known. Then*

(i) *the estimator $\hat{\hat{t}}_z^{(A)}$ given in* (7) *is unbiased for $\mathbf{E}_X t_z$ under the sampling design described in Section* 3*:*

$$\mathbf{E} \hat{\hat{t}}_z^{(A)} = \sum_{i=1}^{N} p_i,$$

(ii) *its variance*

$$Var\big(\hat{\hat{t}}_z^{(A)}\big) = \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} p_i^2 + \sum_{\substack{i,j=1\\ i\neq j}}^{N} (\pi_{ij} - \pi_i\pi_j)\frac{p_i p_j}{\pi_i\pi_j},$$

(iii) *the estimator of variance*

$$\widehat{Var}\big(\hat{\hat{t}}_z^{(A)}\big) = \sum_{i\in\mathbf{s}_I} \frac{1-\pi_i}{\pi_i^2} p_i^2 + \sum_{\substack{i,j\in\mathbf{s}_I\\ i\neq j}} \frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}} \frac{p_i}{\pi_i}\frac{p_j}{\pi_j}$$

is unbiased for $Var\big(\hat{\hat{t}}_z^{(A)}\big)$.

Estimator B. Let us introduce a *design and model*-based estimator of a value z_i

$$\widehat{\widehat{z}}_i = \begin{cases} 1, & \text{if } Y_i > 0,\\ P(X_i > 0|Y_i = 0), & \text{if } Y_i = 0. \end{cases}$$

and define a new estimator of the total t_z:

$$\hat{\hat{t}}_z^{(B)} = \sum_{i\in\mathbf{s}_I} d_i \widehat{\widehat{z}}_i. \tag{8}$$

The probability $P(X_i > 0|Y_i = 0)$ in the case $Y_i = 0$ is included here in the definition of $\widehat{\widehat{z}}_i$ in comparison to $\widehat{z}_i$ in (3).

Proposition B. *Assume the distribution of random variables X_i, $i = 1, 2, \ldots, N$, to be known, and M_i, $M_i > 0$ to be fixed and known. Then*

(i) *the estimator $\hat{\hat{t}}_z^{(B)}$, given in* (8)*, is unbiased for $\mathbf{E}_X t_z$ under the sampling design described in Section* 3 *and model:*

$$\mathbf{E}\hat{\hat{t}}_z^{(B)} = \sum_{i=1}^{N} p_i, \tag{9}$$

(ii) *its variance*

$$Var\big(\hat{\hat{t}}_z^{(B)}\big) = Var\big(\hat{\hat{t}}_z^{(A)}\big) + \sum_{i=1}^{N} d_i\big(p_i - p_i^2 - P(X_i > 0|Y_i = 0)p_{M_i}(0)\big),$$

(iii) *the suggested estimator of variance is*

$$\widehat{Var}\big(\hat{\hat{t}}_z^{(B)}\big) = \widehat{Var}\big(\hat{\hat{t}}_z^{(A)}\big) + \sum_{i\in\mathbf{s}_I} d_i^2\big(p_i - p_i^2 - P(X_i > 0|Y_i = 0)p_{M_i}(0)\big).$$

The first term in the expression of $Var(\hat{\hat{t}}_z^{(B)})$ is due to the sampling design, and second term is due to the distribution of X_i, $i = 1, 2, \ldots, N$.

Remark 1. If X_i is non-random, then

$$p_i = \begin{cases} 1, & \text{if } X_i > 0, \\ 0, & \text{if } X_i = 0, \end{cases}$$

and $\mathbf{E}\hat{\hat{t}}_z^{(A)} = \mathbf{E}\hat{\hat{t}}_z^{(B)} = t_z$ is the number of population units associated with at least one subunit with an attribute.

Remark 2. Calculation of the probability $P(X_i > 0|Y_i = 0)$ used for $Var(\hat{\hat{t}}_z^{(B)})$ for $m = 3$:

$$\begin{aligned} &P(X_i > 0|Y_i = 0)P(Y_i = 0) \\ &\quad = \sum_{k=1}^{M_i-3} P(X_i = k|Y_i = 0)P(Y_i = 0) \\ &\quad = \sum_{k=1}^{M_i-3} P(Y_i = 0|X_i = k)P(X_i = k) \\ &\quad = \sum_{k=1}^{M_i-3} \left(1 - \frac{k}{M_i}\right)\left(1 - \frac{k}{M_i - 1}\right)\left(1 - \frac{k}{M_i - 2}\right)p_{M_i}(k). \end{aligned} \tag{10}$$

Hence,

$$\begin{aligned} &P(X_i > 0|Y_i = 0) \\ &\quad = \frac{1}{P(Y_i = 0)} \sum_{k=1}^{M_i-3} \left(1 - \frac{k}{M_i}\right)\left(1 - \frac{k}{M_i - 1}\right)\left(1 - \frac{k}{M_i - 2}\right)p_{M_i}(k). \end{aligned}$$

Estimator C. In practice, distribution of X_i is not known and it is estimated. Suppose that estimators $\widehat{p}_{M_i}(k)$ are used for the probabilities $p_{M_i}(k)$. Then we define

$$\begin{aligned} &\widehat{P}(X_i > 0|Y_i = 0) \\ &\quad = \frac{1}{\widehat{P}(Y_i = 0)} \sum_{k=1}^{M_i-3} \left(1 - \frac{k}{M_i}\right)\left(1 - \frac{k}{M_i - 1}\right)\left(1 - \frac{k}{M_i - 2}\right)\widehat{p}_{M_i}(k), \\ &\widehat{p}_i = \widehat{P}(Y_i > 0) + \widehat{P}(X_i > 0|Y_i = 0). \end{aligned}$$

Denote $\mathbf{E}_{\widehat{p}}(\cdot)$, $Var_{\widehat{p}}(\cdot)$ expectation and variance with respect to the distribution of the estimators $\widehat{p}_{M_i}(k)$, $k = 1, \ldots, M_i$, $i = 1, \ldots, N$,

$$\widehat{\widehat{\widehat{z}}}_i = \begin{cases} 1, & \text{if } Y_i > 0, \\ \widehat{P}(X_i > 0|Y_i = 0), & \text{if } Y_i = 0, \end{cases}$$

as well as the estimator of the total t_z

$$\hat{\hat{t}}_z^{(C)} = \sum_{i \in \mathbf{s}_I} d_i \widehat{\widehat{\widehat{z}}}_i. \tag{11}$$

Proposition C. *Assume that the estimators* $\widehat{p}_{M_i}(k)$ *are defined for the probabilities* $p_{M_i}(k)$*,* $i = 1, 2, \dots, N$*,* $k = 1, \dots, M_i$*. Then*

(i) *the expectation* $\big($*under model, design and distribution of* $\widehat{p}_{M_i}(k)\big)$ *of estimator* $\hat{\hat{t}}_z^{(C)}$*, given in* (11)*, is*

$$\mathbf{E}\hat{\hat{t}}_z^{(C)} = \sum_{i=1}^{N} \mathbf{E}_{\widehat{p}}\widehat{p}_i,$$

(ii) *its variance is expressed by*

$$\begin{aligned} Var\big(\hat{\hat{t}}_z^{(C)}\big) =& Var\Bigg(\sum_{i\in\mathbf{s}_I} d_i \mathbf{E}_{\widehat{p}}\widehat{p}_i\Bigg) + \sum_{i=1}^{N} d_i Var_{\widehat{p}}(\widehat{p}_i) \\ & + \sum_{i=1}^{N} d_i \mathbf{E}_{\widehat{p}}\big(\widehat{p}_i - \widehat{p}_i^2 + \widehat{P}(Y_i = 0)\widehat{P}(X_i > 0|Y_i = 0) \\ & \times \big(\widehat{P}(X_i > 0|Y_i = 0) - 1\big)\big). \end{aligned}$$

The first term in the expression of variance is due to the sampling design, the third term is is due to the distribution of X_i, $i = 1, 2, \dots, N$, and the second term is due to estimation of the superpopulation distribution probabilities.

Remark 3. The situation can occur that the clusters of subunits are associated only with some, but not all the elements of the population. Then the number of units n' in the sample associated with some subunits may be random, and $n' \leq n$. This invokes one more source of randomness in the estimators of the number of population units, associated with the subunits having attributes, which is not considered here.

Estimation of the proportion. From the equalities

$$\begin{aligned} &\widehat{\mu}_z = \hat{t}_z/N, \\ &Var(\widehat{\mu}_z) = Var(\hat{t}_z)/N^2, \\ &\widehat{Var}(\widehat{\mu}_z) = \widehat{Var}(\hat{t}_z)/N^2 \end{aligned}$$

we can obtain the estimator needed for a proportion, using any estimator of the total presented above.

6 Possible distributions of X_i

Example 3. Let any subunit attached to some unit have an equal probability $p \in (0, 1)$ of bearing the attribute, and subunits have attributes independently of one another. Then the number of subunits X_i is distributed according to the binomial distribution with the parameter p, and

$$p_{M_i}(k) = P(X_i = k|M_i) = C_{M_i}^{k} p^k (1-p)^{M_i - k}, \quad k = 0, 1, \dots, M_i.$$

Then $p_i = P(X_i > 0) = 1 - P(X_i = 0) = 1 - (1-p)^{M_i}$.

For some case of the binomial distribution of X_i, the probabilities $p_{M_i}(k)$ are given in Table 2. They have a peak for $k = k_0 \in (0, M_i)$ with fixed M_i.

Table 2. Probabilities $p_{M_i}(k)$ for the binomial distribution of X_i and $p = 0.6$

	k				
M_i	0	1	2	3	4
1	0.4	0.6			
2	0.16	0.48	0.36		
3	0.064	0.288	0.432	0.216	
4	0.026	0.154	0.346	0.346	0.130

Example 4. Each subunit associated with the *i*th population unit has an attribute with its own probability $p^{(i,j)}$, $j = 1, 2, \ldots, M_i$. Attributes are obtained by the subunits independently of one another. Then the probabilities needed are as follows:

$$\begin{aligned}
&P(X_i = 0) = \big(1 - p^{(1,i)}\big)\big(1 - p^{(2,i)}\big) \ldots \big(1 - p^{(M_i,i)}\big),\\
&p_i = 1 - P(X_i = 0) = 1 - \big(1 - p^{(1,i)}\big)\big(1 - p^{(2,i)}\big) \ldots \big(1 - p^{(M_i,i)}\big),\\
&p_{M_i}(u) = \sum_{\substack{\omega \subseteq U_{2i} \\ |\omega| = u}} \prod_{k \in \omega} p^{(k,i)} \prod_{k \in U_{2i} \setminus \omega} (1 - p^{(k,i)}).
\end{aligned}$$

Example 5. Let us suppose the superpopulation distribution of the number of subunits having attribute (variables X_i, $i = 1, 2, \ldots, N$) to be known, e. g. probabilities $p_i(k) = p_{M_i}(k) = P(X_i = k|M_i)$, $k = 1, 2, \ldots, M_i$, $i = 1, 2, \ldots, N$, to be known. We apply estimator *B* for the Small Population described in Example 1.

Probabilities defining distribution of $X_1,\ X_2,\ X_3$ – Model 1 – are

$$\begin{aligned}
&p_1(0) = 1, \quad p_1(1) = 0, \quad p_2(0) = 0, \quad p_2(1) = 1,\\
&p_3(0) = 0, \quad p_3(1) = 1, \quad p_3(2) = 0.
\end{aligned}$$

Then we calculate according to (10)

$$P(X_3 > 0|Y_3 = 0)P(Y_3 = 0) = \left(1 - \frac{1}{2}\right)p_3(1) = \frac{1}{2}. \tag{12}$$

We can find easily $P(Y_3 > 0) = 1/2$. Hence, $P(Y_3 = 0) = 1 - P(Y_3 > 0) = 1/2$. From (12) we obtain $P(X_3 > 0|Y_3 = 0) = 1$. Then we can calculate all the estimates:

$$\hat{t}_z^{(B1)} = \frac{3}{2}, \quad \hat{t}_z^{(B2)} = 1, \quad \hat{t}_z^{(B3)} = \frac{3}{2}, \quad \hat{t}_z^{(B4)} = 3, \quad \hat{t}_z^{(B5)} = 3.$$

The average of the estimator (8) with respect to the design and model is

$$\begin{aligned}\mathbf{E}\hat{t}_z^B =& \frac{1}{3}\big(\hat{t}_z^{(B1)} + \hat{t}_z^{(B2)}P(Y_3 = 0) + \hat{t}_z^{(B3)}P(Y_3 > 0)\\ &+ \hat{t}_z^{(B4)}P(Y_3 = 0) + \hat{t}_z^{(B5)}P(Y_3 > 0)\end{aligned} \tag{13}$$

and we obtain $\mathbf{E}\hat{t}_z^{(B)} = 2$. It means $\mathbf{E}\hat{t}_z^{(B)} = t_z$, and unbiasedness of the estimator *B*.

On the other hand, we have that

$$\begin{aligned}&p_1 + p_2 + p_3\\ &\quad = P(X_1 > 0) + P(X_2 > 0) + \big(P(Y_3 > 0) + P(X_3 > 0|Y_3 = 0)P(Y_3 = 0)\big)\\ &\quad = 0 + 1 + 1/2 + 1/2 = 2\end{aligned}$$

coincides with $\mathbf{E}\hat{t}_z^{(B)}$, as it is said in Proposition B.

Example 6. Let us suppose other models for superpopulation distribution of X_i in Small Population of Example 1 for $0 < \varepsilon < 1$. The results of estimation are presented in Table 3.

Table 3. Results of the total estimation in the case of Small Population and various superpopulation models

	Model 2	Model 3
	$p_1(0) = 1,\ p_1(0) = 0,$ $p_2(0) = 0,\ p_2(1) = 1,\ p_3(0) = \varepsilon,$ $p_3(1) = 1 - 2\varepsilon,\ p_3(2) = \varepsilon$	$p_1(0) = 1,\ p_1(0) = 0,$ $p_2(0) = 0,\ p_2(1) = 1,\ p_3(0) = \varepsilon^2,$ $p_3(1) = 2\varepsilon,\ p_3(2) = 1 - 2\varepsilon - \varepsilon^2$
$P(Y_3 > 0)$	$1/2$	$1 - \varepsilon - \varepsilon^2$
$P(Y_3 = 0)$	$1/2$	$\varepsilon + \varepsilon^2$
$P(X_3 > 0\|Y_3 = 0)$	$1 - 2\varepsilon$	$1/(1 + \varepsilon)$
$\hat{t}_z^{(B1)}$	$3/2$	$3/2$
$\hat{t}_z^{(B2)}$	$3(1 - 2\varepsilon)/2$	$3/\big(2(1 + \varepsilon)\big)$
$\hat{t}_z^{(B3)}$	$3/2$	$3/2$
$\hat{t}_z^{(B4)}$	$3(1 - \varepsilon)$	$3\big(1 + 1/(1 + \varepsilon)\big)/2$
$\hat{t}_z^{(B5)}$	3	3
$\mathbf{E}\hat{t}_z^{(B)}$	$2 - \varepsilon$	$2 - \varepsilon^2$

Average of the estimator $\hat{t}_z^{(B)}$ with respect to the design and model is calculated according to the formula (13). We see how average of the estimator $\hat{t}_z^{(B)}$ depends on the probability of the unit to have at least one subunit with the attribute. We see also the expectation of the estimator $\hat{t}_z^{(B)}$ for changed model assumptions (distribution of X_3).

Model 3 shows distribution of X_3 for small ε of the type similar to the distribution in Lithuanian AES (compare Table 4), and this is, of course, non-linear function.

Example 7. Let us try to find an approximation of the distribution of the number of the subunits with an attribute attached with some unit, which is met in the Lithuanian AES. $n = 1\,128$ individuals participated in non-formal education in the sample of the year 2007 of Lithuanian AES. The estimated probabilities $\widehat{p}_{M_i}(k) = \widehat{P}(X_i = k|M_i)$ are given in Table 4. They are increasing with an increase of k for fixed M_i and are far from those given in Table 2.

Table 4. Relative frequencies $\widehat{p}_{M_i}(k)$ of the number of job-related learning activities in non-formal education in the Lithuanian AES

	k			
M_i	0	1	2	3
1	0.18	0.82		
2	0.05	0.10	0.85	
3	0.04	0.04	0.11	0.81
≥ 4	0.00	0.04	0.09	0.87

The analytical expression of the function can be used for approximating the probabilities $p_{M_i}(k)$ for real data:

$$f(x) = (1 + cx)^{\alpha}, \quad \alpha > 0,\ \ c > 0,\ \ x \geq 0.$$

Choosing the proper parameters α, c we derive

$$\widehat{p}_{M_i}(k) = \widehat{P}(X_i = k|M_i) = \frac{(1+ck)^{\alpha}}{\sum_{j=0}^{M_i}(1+cj)^{\alpha}}, \quad k = 0, 1, \ldots, M_i. \tag{14}$$

The probabilities $\widehat{p}_{M_i}(k) = \widehat{P}(X_i = k|M_i)$ estimated, using this function with $\alpha = 6$ and $c = 2$, are given in Table 5. They seem to be quite close to the values of the real survey in Table 4.

Table 5. Estimated probabilities $\widehat{p}_{M_i}(k) = \widehat{P}(X_i = k|M_i)$ using the function proposed in (14)

	k				
M_i	0	1	2	3	4
1	0.05	0.95			
2	0.01	0.21	0.78		
3	0.00	0.07	0.27	0.65	
4	0.00	0.03	0.12	0.29	0.56

The empirical results show that probability approximations of the type (14) can be used in estimator *C* (11) for AES.

We were successful in the Lithuanian AES of the year 2007: there are no cases in the sample with $M_i > 3$ and $Y_i = 0$. Anyway, there can be the case in subsequent survey and the estimators proposed may be needed.

7 Conclusion

The alternative estimators have been proposed for the proportion of the population units associated with at least one subunit with the attribute of interest, using the two-stage sampling design and assumptions on the superpopulation distribution of the number of subunits having the attribute. The Examples 1, 5 and 6 show that estimator *B* allows to obtain unbiased estimates to the problem. The success of usage of the estimators proposed depends on knowledge of the distribution of the number of subunits with the attribute associated with the population units.

Acknowledgements

The author is thankful to Statistics Lithuania for a possibility to analyze the survey data and to anonymous referees for useful comments.

Appendix

Proof of the Proposition B. The expectation of the estimator under the design and model is as follows:

$$\begin{aligned}\mathbf{E}\hat{\hat{t}}_z^{(B)} &= \mathbf{E}\sum_{i\in\mathbf{s}_I} d_i\mathbf{E}(\widehat{\hat{z}}_i|\mathbf{s}_I)\\ &= \mathbf{E}\sum_{i\in\mathbf{s}_I} d_i\big(P(Y_i>0)+P(X_i>0|Y_i=0)P(Y_i=0)\big)\\ &= \mathbf{E}\sum_{i\in\mathbf{s}_I} d_ip_i = \sum_{i=1}^N p_i.\end{aligned}$$

The variance is calculated taking into account that sampling designs at both stages are independent, and Y_i, $i\in\mathbf{s}_I$, are independent random variables:

$$\begin{aligned}Var\big(\hat{\hat{t}}_z^{(B)}\big) &= Var\big(\mathbf{E}(\hat{\hat{t}}_z^{(B)}|\mathbf{s}_I)\big)+\mathbf{E}\big(Var(\hat{\hat{t}}_z^{(B)}|\mathbf{s}_I)\big)\\ &= Var\Bigg(\sum_{i\in\mathbf{s}_I} d_i\mathbf{E}(\widehat{\hat{z}}_i|\mathbf{s}_I)\Bigg)+\mathbf{E}\Bigg(\sum_{i\in\mathbf{s}_I} d_i^2Var(\widehat{\hat{z}}_i|\mathbf{s}_I)\Bigg).\end{aligned}$$

Hence,

$$Var\big(\hat{\hat{t}}_z^{(B)}\big) = Var\big(\hat{\hat{t}}_z^{(A)}\big)+\sum_{i=1}^N d_iVar(\widehat{\hat{z}}_i). \tag{15}$$

For $Var(\widehat{\hat{z}}_i)=\mathbf{E}\widehat{\hat{z}}_i^2-(\mathbf{E}\widehat{\hat{z}}_i)^2$, we find:

$$E\widehat{\hat{z}}_i = P(Y_i>0)+P(X_i>0|Y_i=0)P(Y_i=0)=p_i,$$

$$\begin{aligned}
E\widehat{\widehat{z}}_i^2 &= P(Y_i > 0) + P(X_i > 0|Y_i = 0)^2 P(Y_i = 0)\\
&= p_i + P(X_i > 0|Y_i = 0)P(Y_i = 0)\big(P(X_i > 0|Y_i = 0) - 1\big)\\
&= p_i - P(X_i > 0|Y_i = 0)P(Y_i = 0)P(X_i = 0|Y_i = 0)\\
&= p_i - P(X_i > 0|Y_i = 0)P(X_i = 0)\\
&= p_i - P(X_i > 0|Y_i = 0)p_{M_i}(0).
\end{aligned}$$

Hence it follows that

$$Var(\widehat{\widehat{z}}_i) = p_i - p_i^2 - P(X_i > 0|Y_i = 0)p_{M_i}(0).$$

By substituting $Var(\widehat{\widehat{z}}_i)$ in (15), we obtain the expression of variance.

The estimator of variance is obtained using the expression of (iii) Proposition A for the first term and the unbiased Horvitz-Thompson estimator of the total for the second term of the variance. □

References

1. Eurostat, *Adult Education survey – AES*, `http://epp.eurostat.ec.europa.eu/cache/ITY_SDDS/EN/ trng_aes_base.htm♯access/`.
2. Eurostat, *CIRCA – a collaborative workspace with partners of the European Institutions*, `http://circa.europa.eu/`.
3. Eurostat, Database, `http://epp.eurostat.ec.europa.eu/portal/page?_pageid=1996,45323734&_dad=portal&_schema=PORTAL&screen=welcomeref&open=/edtr/trng/trng_aes&language=en&product=EU_MASTER_education_training&root=EU_MASTER_education_training&scrollto=0`.
4. C.-E. Särndal, B. Swensson, J. Wretman, *Model Assisted Survey Sampling*, New York, Springer, 1992.