STATISTICS IN TRANSITION new series, December 2020
Vol. 21, No. 5, pp. 1–16, DOI 10.21307/stattrans-2020-052
Received – 17.03.2020; accepted – 07.08.2020

Estimating the population mean using a continuous sampling design dependent on an auxiliary variable

Janusz L. Wywiał [1]

ABSTRACT

Continuous distribution of variables under study and auxiliary variables are considered. The purpose of the paper is to estimate the mean of the variable under study using a sampling design which is dependent on the observation of a continuous auxiliary variable in the whole population. Auxiliary variable values observed in this population allow to estimate the inclusion density function of the sampling design. The variance of the continuous version of the Horvitz-Thompson estimator under the proposed sampling design is compared with the variance of the mean of a simple random sample. The accuracy of the estimation strategies is analysed by means of simulation experiments.

Key words: continuous sampling design, Horvits-Thompson estimator, inclusion density, sampling scheme, bivariate gamma distribution, ratio estimator.

1. Introduction

Survey sampling theory is well developed for inference based on a finite and fixed population, where the variable under study as well as auxiliary variables are non-random (see, e.g. Särndal, Swenson, Wretman (1992) and Tillé (2006)). The estimation of population parameters is based on a sampling design defined as functions of auxiliary variable values observed in the whole population.

In this paper, the auxiliary variable is also treated as random. We assume that the continuous distribution function of the variable under study and the auxiliary variable (denoted by X and Y respectively) is known, or can be estimated. Values of X and Y are observed on the whole population of size N and in the sample respectively. For instance, the joint distribution of these two variables can be suggested by economic theory. Tax registers are an example of auxiliary variable observation in the whole population.

Another example deals with application of statistics in auditing. Book values of accounting documents are inspected (audited) in order to assess the true values of the documents. Calculating the mean of the true values is one of the purposes of auditing. We can consider joint continuous distribution of the book values and the true values of the documents. The book values can be treated as observations of X throughout the population of the documents, while values of Y are observations of the variable under study. Our aim is to estimate the mean of Y based on a sample selected according to a sampling design dependent on X. For example, Frost and Tamura (1986) and Wywiał (2018) considered gamma distribution for modelling book values in statistical auditing.

[1] University of Economics in Katowice, Poland. E-mail: janusz.wywial@ue.katowice.pl. ORCID: https://orcid.org/0000-0002-3392-1688.

Benhenni and Cambanis (1992) and Thompson (1997) considered continuous sampling for Monte Carlo integration. Some continuous sampling designs were studied in Bąk (2014, 2018), Wilhelm, Tillé and Qualité (2017), and Wywiał (2016). The efficiency of estimation of parameters based on stratified and systematic samples was studied by, e.g. Cressie (1993) and Zubrzycki (1958). A sampling design dependent on the positively valued continuous auxiliary variable proposed by Cox and Snell (1979) was applied to financial auditing. The continuous sampling designs and inclusion density functions were defined by Cordy (1993), who also adapted the well-known Horvitz-Thompson (1952) estimator to estimate parameters. This paper draws on these two sources. In Section 2.1, the properties of the Horvitz-Thompson statistic for the continuous sampling design are presented. Next, in Section 2.2, these properties are generalized to the joint distribution of Y and X. A continuous sampling design with inclusion function proportional to the density function of the auxiliary variable is considered in the third chapter. In the fourth chapter, the main results of the paper are used to construct the estimation strategies under the assumption that the sample was drawn from the continuous population defined by bivariate gamma distribution. The accuracy of these strategies is studied using simulation analysis. In the last chapter, the main conclusions are formulated.

2. Horvitz-Thompson statistic from sample selected according to continuous sampling design

2.1. Basic results

This section has been prepared according to Cordy (1993) results. Let the population $U \subset R^q$, $q = 1, 2, \ldots$. To simplify our analysis we assume that $q = 1$. The sample space, denoted by $S_n = U^n$, is the set of ordered samples denoted by $\mathbf{y} = (y_1, \ldots, y_n)$, $y_k \in U$, $k = 1, \ldots, n$, where y_i is the outcome of the variable observed in the first draw. Let $\mathbf{y}$ be a value of the n-dimensional random variable $\mathbf{Y} = (Y_1, \ldots, Y_n)$ with density function $f(\mathbf{y}) = f(y_1, \ldots, y_n)$. Let $f_i(y)$ and $f_{i,j}(y, y')$, $y \in U$, $y' \in U$, be marginal density functions of Y_i and (Y_i, Y_j) respectively, $j > i = 1, \ldots, n$. The inclusion functions of the first order and the second order are defined respectively as follows:

$$\pi(y) = \sum_{i=1}^{n} f_i(y), \qquad \pi(y, y') = \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} f_{i,j}(y, y'), \quad y \in U, y' \in U \tag{1}$$

and $\int_U \pi(y)dy = n$, $\int_U \int_U \pi(y, y')dydy' = n(n-1)$.

Let $f(y_i|y_{i-1}, y_{i-2}, \ldots, y_1)$, $i = 1, \ldots, n-1$ be the conditional density function of the randomly selected y_i value in the i-th draw (provided that the values $(y_{i-1}, y_{i-2}, \ldots, y_1)$ were drawn earlier). Therefore, the density function of the sampling design can be written as follows:

$$f(y_n, \ldots, y_i, y_{i-1}, \ldots, y_1) = f(y_1) \prod_{i=2}^{n} f(y_i|y_{i-1}, y_{i-2}, \ldots, y_1) \tag{2}$$

Let $g(y)$ be an integrable function $g: U \to R$. We estimate the following parameter:

$$\theta = \int_U g(y)dy. \tag{3}$$

The continuous version of the well-known Horvitz and Thompson (1952) estimator is:

$$T_{\mathbf{Y}} = \sum_{i=1}^{n} \frac{g(Y_i)}{\pi(Y_i)} \tag{4}$$

Theorem 2.1. [Cordy (1993)] The statistic $T_{\mathbf{Y}}$ is an unbiased estimator for θ, if the function $g(y)$ is either bounded or non-negative, and $\pi(y) > 0$ for each $y \in U$.

Theorem 2.2 [Cordy (1993)] If the function $g(y)$ is bounded, $\pi(y) > 0$ for each $y \in U$, and $\int_U (1/\pi(y))dy < \infty$, then

$$\begin{aligned} V(T_{\mathbf{Y}}) = \int_U \frac{g^2(y)}{\pi(y)}dy + \int_U \int_U g(y)g(y')\frac{\pi(y,y') - \pi(y)\pi(y')}{\pi(y)\pi(y')}dydy' = \\ = \int_U \frac{g^2(y)}{\pi(y)}dy + \int_U \int_U g(y)g(y')\frac{\pi(y,y')}{\pi(y)\pi(y')}dydy' - \theta^2. \end{aligned} \tag{5}$$

When, in addition, $\pi(y_i, y_j) > 0$ for all $y_i, y_j \in U$, $i \neq j = 1, ..., n$, an unbiased estimator of the variance in (5) is:

$$\hat{V}(T_{\mathbf{Y}}) = \sum_{i=1}^{n} \frac{g^2(Y_i)}{\pi^2(Y_i)} + \sum_{i=1}^{n} \sum_{j=1, i\neq j}^{n} g(Y_i)g(Y_j)\frac{\pi(Y_i,Y_j) - \pi(Y_i)\pi(Y_j)}{\pi(Y_i,Y_j)\pi(Y_i)\pi(Y_j)}$$

In particular, when $h(y)$ is a density function and $g(y) = \eta(y)h(y)$, then $\theta = E(\eta(Y))$. Of course if $\eta(y) = y$, then $\theta = E(Y)$.

When $Y_1, \ldots, Y_n$ is a random sample from a distribution with density $f(y)$, then the density function of the sampling design defined by (2) and its inclusion functions become as follows:

$$f(y_1, ..., y_n) = \prod_{i=1}^{n} f(y_i), \quad \pi(y) = nf(y), \quad \pi(y,y') = n(n-1)f(y)f(y'). \tag{6}$$

This allows us to transform expressions (4) and (5) as follows:

$$T_{\mathbf{Y}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\eta(Y_i)h(Y_i)}{f(Y_i)}, \quad E(T_{\mathbf{Y}}) = \theta, \tag{7}$$

$$\begin{aligned} V(T_{\mathbf{Y}}) = \frac{1}{n}\left(\int_U \frac{\eta^2(y)h^2(y)}{f(y)}dy - \theta^2\right) = \\ = \frac{1}{n}\left(E\left(\frac{\eta^2(Y)h^2(Y)}{f^2(Y)}\right) - E^2\left(\frac{\eta(Y)h(Y)}{f(Y)}\right)\right) = \frac{1}{n}V\left(\frac{\eta(Y)h(Y)}{f(Y)}\right). \end{aligned} \tag{8}$$

Sampling design $f(y_n,...,y_1)$, given by (6) provides what is known as the *importance sample* considered, e.g. by Bucklew (2004) and Ripley (1987). When the importance sample is drawn from density $h(y)$, then it becomes the well-known simple random sample defined as the sequence of independent and identically distributed random variable (see e.g. Wilks (1962)) and $\theta = E(Y) = \mu_y$ is estimated by means of the following statistic:

$$T_Y = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \quad V(T_Y) = V(\bar{Y}) = \frac{1}{n}V(Y) \tag{9}$$

where $V(Y) = \int_{-\infty}^{\infty}(y - E(Y))^2 f(y)dy$.

2.2. Estimation using auxiliary variable

Let $h(x,y)$, $(x,y) \in U \subseteq R^2$, be the density function. The marginal densities are: $h_1(x)$ and $h_2(y)$. $h(y|x) = h(x,y)/h_1(x)$ is the conditional density. Moreover, $\mu_y = E(Y) = \int_{-\infty}^{-\infty} yh_2(y)dy$, $\mu_x = E(X) = \int_{-\infty}^{-\infty} xh_1(x)dx$, $E(Y|x) = \int_{-\infty}^{-\infty} yh(y|x)dy$, $V(Y|x) = \int_{-\infty}^{-\infty}(y - E(Y|x))^2 h(y|x)dy$. Our purpose is estimation of parameter θ, given by (3) where

$$g(x) = E(\eta(Y)|x)h_1(x) = h_1(x)\int_{-\infty}^{\infty}\eta(y)h(y|x)dy.$$

We set $\eta(y) = y$. Therefore:

$$g(x) = E(Y|x)h_1(x) = h_1(x)\int_{-\infty}^{\infty} yh(y|x)dy. \tag{10}$$

In this case:

$$\theta = \mu_y = \int_{-\infty}^{\infty} E(Y|x)h_1(x)dx = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yh(y|x)h_1(x)dxdy. \tag{11}$$

Parameter μ_y is estimated by means of the following statistic:

$$T_{\mathbf{X},\mathbf{Y}} = \sum_{i=1}^{n}\frac{Y_i h_1(X_i)}{\pi(X_i)} \tag{12}$$

where $\{X_i, i = 1,..,n\}$ is the sample drawn according to sampling design defined by expression (2) and y_i should be replaced by x_i. Let us assume that:

$$h(y|x) = h(y_1,...,y_n|x_1,...,x_n) = \prod_{i=1}^{n} h(y_i|x_i) \tag{13}$$

Theorem 2.3 If $E(Y) < \infty$ and $\pi(x) > 0$ for all $(x,y) \in U$ and assumption (13) holds, then $E_{\mathbf{f(X)}}E_{\mathbf{h(Y/X)}}(T_{\mathbf{X},\mathbf{Y}}) = \mu_y$.

Proof: When in (4) we replace $g(Y_i)$ with $g(X_i)$, given by (10), then Theorem 2.1 let us write

$$E_{\mathbf{f(X)}}\left(\sum_{i=1}^{n}\frac{g(X_i)}{\pi(X_i)}\right) \quad = \quad E_{\mathbf{f(X)}}\left(\sum_{i=1}^{n}\frac{E_{\mathbf{h(Y/X)}}(Y_i)h_1(X_i)}{\pi(X_i)}\right) \quad = \quad E_{\mathbf{f(X)}}E_{\mathbf{h(Y/X)}}\left(T_{\mathbf{X,Y}}\right).$$

This derivation shows that Theorem 2.3 is a special case of Theorem 2.1.

Theorem 2.4 If the function $E(Y)$ is bounded, $\pi(y)>0$ for each $(x,y)\in U$, and $\int_U(1/\pi(y))dy<\infty$, then

$$V(T_{\mathbf{X,Y}})=\int_U\frac{V(Y|x)h_1^2(x)}{\pi(x)}dx+\int_U\frac{E^2(Y|x)h_1^2(x)}{\pi(x)}dx+A \tag{14}$$

where

$$A=\int_U\int_U E(Y|x)h_1(x)E(Y|x')h_1(x')\frac{\pi(x,x')-\pi(x)\pi(x')}{\pi(x)\pi(x')}dxdx'$$

or

$$A=\int_U\int_U E(Y|x)h_1(x)E(Y|x')h_1(x')\frac{\pi(x,x')}{\pi(x)\pi(x')}dxdx'-E^2(Y).$$

Proof: Adding $E_{\mathbf{h(Y/X)}}(T_{\mathbf{X,Y}})$ to $E_{f(\mathbf{X})}E_{\mathbf{h(Y/X)}}(T_{\mathbf{X,Y}}-\mu_y)^2$ we have:

$$\begin{aligned}V(T_{\mathbf{X,Y}}) &= E_{f(\mathbf{X})}E_{\mathbf{h(Y/X)}}((T_{\mathbf{X,Y}}-E_{\mathbf{h(Y/X)}}(T_{\mathbf{X,Y}}))+(E_{\mathbf{h(Y/X)}}(T_{\mathbf{X,Y}})-E(Y)))^2=\\ &=E_{f(\mathbf{X})}E_{\mathbf{h(Y/X)}}\left(\left(\sum_{i=1}^{n}\frac{Y_i-E_{\mathbf{h(Y/X)}}(Y_i)h_1(X_i)}{\pi(X_i)}\right)+(E_{\mathbf{h(Y/X)}}(T_{\mathbf{X,Y}})-\mu_y)\right)^2=\\ &=E_{f(\mathbf{X})}\left(\sum_{i=1}^{n}\frac{V_{\mathbf{h(Y/X)}}(Y_i)h_1^2(X_i)}{\pi^2(X_i)}\right)+E_{f(\mathbf{X})}\left(\sum_{i=1}^{n}\frac{E_{\mathbf{h(Y/X)}}(Y_i)h_1(X_i)}{\pi(X_i)}-\mu_y\right)^2,\end{aligned}$$

because $E_{\mathbf{h(Y/X)}}(Y_i-E_{\mathbf{h(Y/X)}}(Y_i))=0$ and $E_{\mathbf{h(Y/X)}}(Y_i-E_{\mathbf{h(Y/X)}}(Y_i))^2=V_{\mathbf{Y/X}}(Y_i)$. Continuing the derivation we have:

$$V(T_{\mathbf{X,Y}})= \\ =E_{f(\mathbf{X})}\left(\sum_{i=1}^{n}\frac{V(Y_i|X_i)h_1^2(X_i)}{\pi^2(X_i)}\right)+E_{f(\mathbf{X})}\left(\sum_{i=1}^{n}\frac{E(Y_i|X_i)h_1(X_i)}{\pi(X_i)}-\mu_y\right)^2. \tag{15}$$

By setting $\frac{V(Y_i|X_i)h_1^2(X_i)}{\pi(X_i)}=g(X_i)$ Theorem 2.1 allows us to write the following:

$$E_{f(\mathbf{X})}\left(\sum_{i=1}^{n}\frac{V(Y_i|X_i)h_1^2(X_i)}{\pi^2(X_i)}\right)=\int_U\frac{V(Y|x)h_1^2(x)}{\pi(x)}dx. \tag{16}$$

Similarly, by setting $E(Y_i|X_i)h_1(X_i) = g(X_i)$, the second term in (15) becomes:

$$E_{f(\mathbf{X})}\left(\sum_{i=1}^{n}\frac{g(X_i)}{\pi(X_i)} - \mu_y\right)^2 =$$

$$= E_{f(\mathbf{X})}\left(\sum_{i=1}^{n}\frac{g(X_i)}{\pi(X_i)} - E_{f(\mathbf{X})}\left(\sum_{i=1}^{n}\frac{g(X_i)}{\pi(X_i)}\right)\right)^2 = V_{f(\mathbf{X})}\left(\sum_{i=1}^{n}\frac{g(X_i)}{\pi(X_i)}\right). \quad (17)$$

This, expression (16) and Theorem 2.2 lead straightforward to the conclusion of Theorem 2.4.

Similarly to expression (6) let us assume that

$$f(x_1,...,x_n) = \prod_{i=1}^{n} f(x_i), \quad \pi(x) = nf(x), \quad \pi(x,x') = n(n-1)f(x)f(x'). \quad (18)$$

This, expression (17) and Theorem 2.4 lead to the following:

$$V(T_{\mathbf{X},\mathbf{Y}}) = \frac{1}{n}\left(\int_U \frac{V(Y|x)h_1^2(x)}{f(x)}dx + \int_U \frac{E^2(Y|x)h_1^2(x)}{f(x)}dx - E^2(Y)\right) =$$

$$= \frac{1}{n}\left(E_{f(X)}\left(\frac{V(Y|X)h_1^2(X)}{f^2(X)}\right) + V_{f(X)}\left(\frac{E(Y|X)h_1(X)}{f(X)}\right)\right) \quad (19)$$

We estimate μ_y with the following sampling design:

$$f(x_1,...,x_n) = \prod_{i=1}^{n} h_1(x_i). \quad (20)$$

Under additional assumption that $E(Y|x) = ax$ where $a = \rho\sqrt{\frac{V(Y)}{V(X)}}$ and ρ is the correlation coefficient between X and Y then expressions (12) and (19) lead to the following:

$$T_{\mathbf{X},\mathbf{Y}} = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \qquad E(\bar{Y}) = \mu_y, \qquad V(\bar{Y}) = \frac{V(Y)}{n}(1+\rho^2) \quad (21)$$

Hence, when $\rho \neq 0$, estimator $T_{\mathbf{X},\mathbf{Y}}$ of the mean based on sampling design, given by (20) is less accurate than the simple random sample mean.

3. Inclusion function of sampling design proportional to values of auxiliary variable

3.1. Density function of the auxiliary variable is known

After Cox and Snell (1979), let us consider the following sampling design:

$$f(x_1,...,x_n) = \prod_{i=1}^{n} f(x_i), \qquad f(x_i) = \frac{x_i h_1(x_i)}{\mu_x}. \tag{22}$$

where $\mu_x = E(X) = E(X_i)$ for all $i = 1,...,n$. In this case, according to (18) the inclusion function is proportional to the value of the auxiliary variable because $\pi(x) = \frac{nxh_1(x)}{\mu_x}$. Expression (12), (19), Theorems 2.3 and Theorem 2.4 lead to the following:

$$T_{\mathbf{X},\mathbf{Y}} = \hat{Y}_R = \frac{\mu_x}{n}\sum_{i=1}^{n}\frac{Y_i}{X_i}, \qquad E(\hat{Y}_R) = \mu_y, \tag{23}$$

$$V(T_{\mathbf{X},\mathbf{Y}}) = \frac{1}{n}\left(\mu_x \int_U \frac{V(Y|x)h_1(x)}{x}dx + \mu_x \int_U \frac{E^2(Y|x)h_1(x)}{x}dx - \mu_y^2\right) =$$
$$= \frac{\mu_x}{n}\int_U \frac{V(Y|x)h_1(x)}{x}dx + \frac{\mu_x}{n}V\left(\frac{E(Y|x)}{x}\right). \tag{24}$$

Statistic $\hat{Y}_R$ is an unbiased ratio-type estimator of μ_y.

When parameter μ_x and other parameters of the auxiliary variable density function are known, the sample can be select. The following sections address selection when these parameters are estimated.

3.2. Estimated parameters of the auxiliary variable density function

The values $x_1,...,x_N$ of the auxiliary variable observed in whole population are regarded as a random sample from a distribution with density $h_1(x,\theta_1,...,\theta_r)$. Let $\hat{\theta}_1...\hat{\theta}_r$ and $\hat{\mu}_x$ be consistent estimators of parameters $\theta_1,...,\theta_r$ and μ_x respectively. According to expression (22) we have the following density function of sampling design:

$$\hat{f}(x_1,...,x_n) = f(x_1,...,x_n,\hat{\theta}_1...\hat{\theta}_r) = \prod_{i=1}^{n}\hat{f}(x_i), \quad \hat{f}(x) = \frac{xh_1(\hat{\theta}_1...\hat{\theta}_r)}{\hat{\mu}_x}. \tag{25}$$

Estimation of parameters could be based on data observed, e.g. in the previous round of a regularly conducted survey.

By replacing μ_x in eq. (23) with $\hat{\mu}_x$ we obtain the following estimator:

$$T_{\mathbf{X},\mathbf{Y}} = \tilde{Y}_R = \frac{\hat{\mu}_x}{n}\sum_{i=1}^{n}\frac{Y_i}{X_i}. \tag{26}$$

where $X_1,...,X_n$ is the sample drawn according to sampling design based on density $f(x_1,...,x_n, \hat{\theta}_1...\hat{\theta}_r)$. The variance of $\tilde{Y}_R$ could be estimated by means of the well-known parametric or non-parametric method of bootstrap.

3.3. Kernel estimator of the auxiliary variable density function

Density function $h_1(x)$ can be estimated by means of the following well-known kernel-type estimator on the basis of all observations of auxiliary variable in the population:

$$\tilde{h}_1(x) = \frac{1}{N}\sum_{i=1}^{N} k(x,x_i,\Delta), \quad \int_{-\infty}^{\infty} k(x,x_i,\Delta)dx = 1 \tag{27}$$

where $\Delta > 0$ is the bandwidth parameter. This leads to the following estimator of $f(x)$:

$$\tilde{f}(x) = \frac{x\tilde{h}_1(x)}{\tilde{\mu}_x} = \frac{\sum_{i=1}^{N} xk(x,x_i,\Delta)}{N\tilde{\mu}_x} \tag{28}$$

where:

$$\tilde{\mu}_x = \int_{-\infty}^{\infty} x\tilde{h}_1(x)dx \tag{29}$$

is the estimator of μ_x.

Let us consider the following simple kernel function based on the uniform distribution:

$$k(x,x_i,\Delta) = \begin{cases} \frac{1}{2\Delta}, & x \in [x_i-\Delta;x_i+\Delta], \\ 0, & x \notin [x_i-\Delta;x_i+\Delta]. \end{cases} \tag{30}$$

For this kernel function we have:

$$\int_{-\infty}^{\infty} xk(x,x_i,\Delta)dx = x_i \quad \textbf{for} \quad i = 1,...N, \quad \textbf{and} \quad \tilde{\mu}_x = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{31}$$

Expression (28) leads to the following:

$$\tilde{f}(x) = \frac{1}{N\bar{x}}\sum_{i=1}^{N} xk(x,x_i,\Delta) = \frac{1}{N\bar{x}}\sum_{i=1}^{N} x_i\tilde{f}_i(x,x_i,\Delta) = \sum_{i=1}^{N} w_i\tilde{f}_i(x,x_i,\Delta) \tag{32}$$

where: $w_i = \frac{x_i}{N\bar{x}}$, for $i = 1,...,N$ and

$$\tilde{f}_i(x,x_i,\Delta) = \begin{cases} \frac{x}{2x_i\Delta}, & x \in [x_i-\Delta;x_i+\Delta], \\ 0, & x \notin [x_i-\Delta;x_i+\Delta] \end{cases} \tag{33}$$

where $\tilde{f}_i(x,x_i,\Delta)$ is the trapezoid density function of the probability distribution on interval

$[x_i - \Delta; x_i + \Delta]$. After simplifications we have:

$$\tilde{f}(x) = \frac{1}{2\Delta N\bar{x}} \sum_{i=1}^{N} xI(x, x_i, \Delta) \tag{34}$$

where:

$$I(x, x_i, \Delta) = \begin{cases} 1, & x \in [x_i - \Delta; x_i + \Delta], \\ 0, & x \notin [x_i - \Delta; x_i + \Delta]. \end{cases} \tag{35}$$

Expressions (32) and (33) allows us to derive the following distribution function estimator:

$$\tilde{F}(x) = \int_{-\infty}^{x} \tilde{f}(t)dt = \sum_{i=1}^{N} w_i \tilde{F}_i(x, x_i, \Delta) \tag{36}$$

where: $w_i = \frac{x_i}{N\bar{x}}$, for $i = 1, ..., N$ and

$$\tilde{F}_i(x, x_i, \Delta) = \begin{cases} 0, & x \in (-\infty; x_i - \Delta], \\ \frac{x^2 - (x_i - \Delta)^2}{4x_i\Delta}, & x \in (x_i - \Delta; x_i + \Delta], \\ 1, & x \in [x_i + \Delta; \infty). \end{cases} \tag{37}$$

The inverse function to $\tilde{F}_i(x)$ (the quantile function), $i = 1, .., N$, is as follows:

$$x = \tilde{F}_i^{-1}(u) = \sqrt{4x_i\Delta u + (x_i - \Delta)^2}, \qquad z \in [0; 1] \tag{38}$$

where u has uniform distribution on interval $[0; 1]$. This allows us to easily generate the pseudovalues of the trapezoid distribution on interval $[x_i - \Delta; x_i + \Delta]$.

3.4. Sampling schemes

Let us assume that observations of $\mathbf{x} = [x_1, ..., x_k, ..., x_N]$ are known book values or they are gathered from a census or surveys made on a previous occasion. Function $h_1(x)$ is also known. Our purpose is to select sample $\mathbf{x}_s = [x_1, ..., x_k, ..., x_n]$ as the sub-vector of $\mathbf{x}$ according to the sampling design defined by expression (22). In order to do this, values of vector $\mathbf{x}'_s = [x'_1, ..., x'_n]$ are generated by means of the quantile functions $x' = F^{-1}(u)$, where u is the value of the uniformly distributed variable on interval $[0; 1]$, $F(x) = \int_{-\infty}^{x} f(t)dt$ and $f(t)$ are given by (22). Elements of $\mathbf{x}_s$ are selected from $\mathbf{x}$ according to

$$x_k = arg \min_{j=1,...,N} |x_j - x'_k|. \tag{39}$$

This algorithm could lead to a repetition of the elements in $\mathbf{x}_s$. If the algorithm yields a sample with duplicate elements, the sample is rejected and the algorithm repeated until a sample with no duplicates is obtained.

The next algorithm, which leads to drawing $\mathbf{x}_s$ without repetition, is explained by expression:

$$x_s = arg \min_{\mathbf{x}_s \in \mathbf{X}_s} (\mathbf{x}_s - \mathbf{x}'_s)(\mathbf{x}_s - \mathbf{x}'_s)^T \tag{40}$$

where $\mathbf{X}_s$ consists of all n-element combinations selected without replacement from $\mathbf{x}$. The complete data $\mathbf{d} = [(x_1, y_1), ..., (x_n, y_n)]$ are evaluated after observation values y_j, $j = 1, ..., n$ (observations of the variable under study) are attached to the appropriate elements of vector $\mathbf{x}_s$. This algorithm becomes simpler when elements of $\mathbf{x}'_s$ and $\mathbf{x}$ are ordered from the lowest to highest.

The next variant of the sampling design is as follows. Let us note that the kernel density function $\tilde{f}(x)$, defined by expression (32), could be treated as a mixture of density functions $\tilde{f}_i(x)$, $i = 1, ..., N$ given by (33). Therefore, the k-th element of vector $\mathbf{x}'_s$ could be generated as follows. Firstly, the value of index i is randomly (with probability w_i) selected from the sequence $1, ..., N$. Next, the values x'_k (k=1,...,n) are generated by means of the quantile function, given by (36)-(38). Finally, the elements of vector $\mathbf{x}_s$ could be selected according to expression (39) or (40).

The complete data $\mathbf{d} = [(x_1, y_1)...(x_n, y_n)]$ are evaluated after observation values y_j, $j = 1, ..., n$ are attached to appropriate elements of vector $\mathbf{x}$.

4. Estimation in the case of McKay's bivariate gamma distribution

Suppose the random variables U_i have distributions with gamma densities

$$l_i(u_i) = l_i(u_i, \theta_i, c) = \frac{c^{\theta_i}}{\Gamma(\theta_i)} u_i^{\theta_i - 1} e^{-cu_i} \tag{41}$$

where: $u_i > 0$, $c > 0$, $\theta_i > 0$, $E(U_i) = \frac{\theta_i}{c}$, $V(U_i) = \frac{\theta_i}{c^2}$, $i = 0, 1, 01$, $\theta_{01} = \theta_0 + \theta_1$ and $U_{01} = U_0 + U_1$ provided U_0 and U_1 are independent. θ and c are called the shape parameter and the scale parameter respectively.

The McKay's (1934) density function of joint probability distribution of $X = U_{01}$ and $Y = U_0$ takes the following form (see also Ghirtis (1967) and Kotz et al. (2000)):

$$l(x, y) = \frac{c^{\theta_{01}}}{\Gamma(\theta_0)\Gamma(\theta_1)} y^{\theta_0 - 1} (x - y)^{\theta_1 - 1} e^{-cx}, \quad x > y > 0. \tag{42}$$

This could be useful with valuation of damage supported by declared observed data as values of X. In this case μ_y is mean of the true valuation of damage.

According to expression (22), the sampling design density function is defined as follows:

$$f(x) = \frac{x}{\mu_x} l_{01}(x) \tag{43}$$

where $f(x)$ is also density function of gamma distribution with shape and scale parameters equal to $\theta_{01} + 1$ and c respectively.

The conditional density function is:

$$l(y|x) = \frac{\Gamma(\theta_{01})}{\Gamma(\theta_0)\Gamma(\theta_1)} x^{-\theta_0} y^{\theta_0 - 1} \left(1 - \frac{y}{x}\right)^{\theta_1 - 1}, \qquad x > y.$$

Its first two moments are:

$$\begin{cases} E(Y|x) = xE(U) = \frac{\theta_0 x}{\theta_{01}}, \\ V(Y|x) = E(Y^2|x) - E^2(Y|x) = x^2V(U) = \frac{\theta_0\theta_1 x^2}{(\theta_{01})^2(\theta_{01}+1)} \end{cases} \tag{44}$$

where U has the beta probability distribution with parameters θ_0 and θ_1.

Expressions (24) and (44) lead to the following:

$$V(\hat{Y}_R) = \frac{\theta_0}{nc}\left(\left(\frac{\theta_1}{\theta_{01}(\theta_{01}+1)} + \frac{\theta_0}{\theta_{01}}\right)E(X) - \frac{\theta_0}{c}\right).$$

By substituting the expression $\frac{\theta_{01}}{c}$ for $E(X)$ we obtain:

$$V(\hat{Y}_R) = \frac{\theta_0\theta_1}{nc^2(\theta_{01}+1)} = \frac{1}{n}\mu_y(\mu_x - \mu_y)\frac{\gamma_x^2}{1+\gamma_x^2}, \quad \gamma_x = \frac{\sigma_x}{\mu_x}.$$

Finally, we have:

$$V(\hat{Y}_R) = \frac{\theta_1}{\theta_{01}+1}V(\bar{Y}) < V(\bar{Y}) = \frac{\theta_0}{nc^2}. \tag{45}$$

The variation coefficient of the estimator is as follows:

$$\gamma(\hat{Y}_R) = 100\%\frac{\sqrt{V(\hat{Y}_R)}}{\mu_y}. \tag{46}$$

The relative efficiency coefficient takes the following form:

$$deff(\hat{Y}_R) = 100\%\frac{V(\hat{Y}_R)}{V(\bar{Y})} = \frac{100\%\theta_1}{\theta_{01}+1} < 100\%. \tag{47}$$

Hence, the estimator $\hat{Y}_R$ is more precise than $\bar{Y}$.

Parameters θ_0 and c of the auxiliary variable can be estimated based on the observed data $\mathbf{x} = [x_1, ..., x_N]$. The method of moments yields the following estimates of the parameters:

$$\hat{\theta}_{01} = \frac{\bar{x}^2}{\hat{v}_x} = \hat{\gamma}_x^{-2}, \quad \hat{\theta}_0 = \tilde{Y}_R\frac{\bar{x}}{\hat{v}_x}, \quad \hat{\theta}_1 = \frac{(\bar{x}-\tilde{Y}_R)\bar{x}}{\hat{v}_x}, \quad \hat{c} = \frac{\bar{x}}{\hat{v}_x} \tag{48}$$

where

$$\hat{v}_x = \frac{1}{N-1}\sum_{k=1}^{N}(x_k - \bar{x})^2, \quad \bar{x} = \frac{1}{N}\sum_{k=1}^{N}x_k, \quad \hat{\gamma}_x = \frac{\hat{v}_x}{\bar{x}^2}.$$

We estimate the density $f(x)$ by

$$\hat{f}(x) = \frac{x}{\bar{x}}\hat{l}_{01}(x, \hat{\theta}_{01}, \hat{c}) \tag{49}$$

which is the gamma density with parameters $\hat{\theta}_{01} + 1 = \bar{x}\hat{c} + 1$ and $\hat{c}$. The expectation μ_y can be estimated using the statistic $\tilde{Y}_R$, given by (26).

Owing to (45), the variance $V(\hat{Y}_R)$ can be estimated by means of the following statistic:

$$\tilde{V}(\tilde{Y}_R, \hat{f}(x)) = \frac{1}{n}\tilde{Y}_R(\bar{x} - \tilde{Y}_R)\frac{\hat{\gamma}_x^2}{1 + \hat{\gamma}_x^2}. \tag{50}$$

The variance could be estimated by means of the following non-parametric bootstrap method. Firstly, the value of the estimator $\hat{Y}_R$ is evaluated based on the data observed in the original sample $\mathbf{D} = [(Y_j, X_j), j = 1, ..., n]$. Bootstrap samples will be denoted by $\mathbf{D}^{(k)} = \left[\left(Y_j^{(k)}, X_j^{(k)}\right), j = 1, ..., n\right]$, $k = 1, ..., B$ which are independently drawn with replacement from sample $\mathbf{D}$. This leads to the following bootstrap-type estimators of variance:

$$\hat{V}\left(\tilde{Y}_R\right) = \frac{1}{B-1}\sum_{k=1}^{B}\left(\tilde{Y}_R^{(k)} - \tilde{Y}_R\right)^2, \quad \tilde{Y}_R^{(k)} = \frac{\bar{x}}{n}\sum_{k=1}^{n}\frac{Y_i^{(k)}}{X_i^{(k)}} \tag{51}$$

or

$$\hat{V}'\left(\tilde{Y}_R\right) = \frac{1}{B-1}\sum_{k=1}^{B}\left(\tilde{Y}_R^{(k)} - \bar{\tilde{Y}}_R\right)^2, \qquad \bar{\tilde{Y}}_R = \frac{1}{B}\sum_{k=1}^{B}\tilde{Y}_R^{(k)}. \tag{52}$$

We set that $B = 1000$.

Example

Let us suppose that the population data are generated according to bivariate gamma distribution defined by density $l(x,y)$, given by (42). We estimate μ_y by two methods denoted by $(\tilde{Y}_R, \tilde{f}(x))$ and $(\tilde{Y}_R, \hat{f}(x))$, explained by expressions (32) and (49) respectively. They are implemented in ”R” language.

First, the program draws random samples $\mathbf{D}_i = \left[(Y_j, X_j)_i, j = 1, ..., 3000\right]$, $i = 1, ..., T$ from McKay distribution. Next, the parameters of the inclusion density function are estimated. This allows us to draw the samples $\mathbf{D}_{1i} = \left[(Y_j, X_j)_i, j = 1, ..., n\right]$ from $\mathbf{D}_i$ and evaluate the values of $\tilde{Y}_R^{(i)}$ of μ_y, $i = 1, ..., T$. This is replicated $T = 1000$-times. Results for some alternative sample sizes and the gamma density function parameters are in columns 1-6 of Table 1. Under the assumed parameters of gamma distribution, the true values of the variation coefficient and *deff* coefficient (given by expression (46) and (47) respectively) have been calculated. They are presented in columns 7 and 8 respectively. In columns 10 and 12 there are values of the relative bias coefficients of the variance estimation, given by the following expressions:

$$b_2 = 100\frac{\tilde{V}(\tilde{Y}_R, \hat{f}(x))}{\check{V}(\tilde{Y}_R, \hat{f}(x))}, \qquad b_2' = 100\frac{\bar{\hat{V}}(\tilde{Y}_R, \tilde{f}(x))}{\check{V}(\tilde{Y}_R, \tilde{f}(x))}, \quad \bar{\hat{V}}(\tilde{Y}_R, \tilde{f}(x)) = \frac{1}{T}\sum_{i=1}^{T}\hat{V}_i(\tilde{Y}_R, \tilde{f}(x)) \tag{53}$$

where $\hat{V}_i(\tilde{Y}_R, \tilde{f}(x))$ explains the right side of equation (51) for the bootstrap samples: $\mathbf{D}_{1i}^{(k)} = \left[\left(Y_j^{(k)}, X_j^{(k)}\right)_i, j = 1, ..., n\right]$, $k = 1, ..., B$ drawn from $\mathbf{D}_{1i}$, $i = 1, ...T$. In columns 9, 11 and

13, there are the following estimated relative efficiency coefficients:

$$d = 100\frac{\check{V}(\tilde{Y}_R,\hat{f}(x))}{\check{V}(\bar{Y})}, \qquad d' = 100\frac{\check{V}(\tilde{Y}_R,\tilde{f}(x))}{\check{V}(\bar{Y})}, \qquad e = 100\frac{\check{V}(\tilde{Y}_R,\tilde{f}(x))}{\check{V}(\tilde{Y}_R,\hat{f}(x))} \tag{54}$$

where $\tilde{y}_R$ is given by (26) and:

$$\check{V}(\tilde{Y}_R,.) = \frac{1}{T-1}\sum_{i=1}^{T}\left(\tilde{Y}_R^{(i)} - \bar{\tilde{Y}}_R\right)^2, \qquad \bar{\tilde{Y}}_R = \frac{1}{T-1}\sum_{i=1}^{T}\tilde{Y}_R^{(i)} \tag{55}$$

are evaluated based on samples $\mathbf{D}_{1i}$, $i = 1,...T$.

Table 1. Relative efficiency and bias of the estimation methods.

								$(\tilde{Y}_R,\hat{f}(x))$		$(\tilde{Y}_R,\tilde{f}(x))$		
n	θ_1	θ_0	c	μ_y	μ_x	$\gamma(\hat{Y}_R)$	$deff$	d	b_2	d'	b'_2	e
1	2	3	4	5	6	7	8	9	10	11	12	13
30	1	10	1	10	11	1.7	8.3	8.3	93.7	9.0	82.8	97.1
60	1	10	1	10	11	1.2	8.3	9.4	89.0	10.5	72.5	102.3
150	1	10	1	10	11	0.8	8.3	11.5	65.9	13.4	59.2	109.6
60	1	10	0.01	1000	1100	1.2	8.3	8.3	90.1	10.8	74.7	110.1
60	3	10	0.01	1000	1300	1.9	21.4	24.0	99.9	22.1	94.0	85.9
60	10	3	0.01	300	1300	6.3	71.4	65.2	105.6	75.1	91.7	99.9

Source: Own calculations.

Statistic $\check{V}(\bar{Y})$ is evaluated by replacing $\tilde{Y}_R$ with the sample mean in equation (55). The relative efficiency coefficients in columns 9 and 10 deal with the case when the sample is selected according to the inclusion density function defined by expression (49). The coefficients from columns 11-12 are calculated based on the data from the sample drawn according to the inclusion density function defined by expressions (32) and (33), where we assumed that the bandwidth parameter $\Delta = \sqrt{\hat{v}_x}$. Moreover, in this case variance of $\tilde{Y}_R$ is estimated by means of the bootstrap method based on expression (51). In column 13, there are values of the relative efficiency coefficient of the estimation methods $(\tilde{Y}_R,\hat{f}(x))$ and $(\tilde{Y}_R,\tilde{f}(x))$ denoted by e. This is evaluated based on expressions (54).

The simulation analysis allows us to calculate values of the relative bias coefficient of the mean estimation defined by $b_1 = 100\bar{\tilde{y}}_R/\mu_y$. Its values for both considered estimation methods oscillate between 98% and 101%. This confirms that both methods give unbiased estimates of the expected value of the variable under study. Therefore, the values of the coefficient b_1 are not presented in Table 1.

Column 7 shows that in the case when $\theta_1 > \theta_0$, a value of the variation coefficient of $\hat{Y}_R$ is larger then its value for $\theta_1 < \theta_0$. Column 8 allows us to conclude that the variance of the estimator under the continuous sampling design equal to the modified density function of the auxiliary variable has a lower value than the variance of the simple random sample mean. Column 9 gives the relative efficiency coefficient value evaluated under the assumption that the parameters of the inclusion density function are estimated. Values of this coefficient differ from appropriate values of *deff* by no more than 4.2%. This is the effect of variability of the parameter estimators. Similarly (see column 11), the kernel-type estimator of the inclusion density function leads to the higher (but not by more than 4.3%) values of d' than the appropriate values of *deff*.

The proposed estimators of the variances are quite significantly biased. Usually, they underestimate the variances (see columns 10 and 12). The bias depends on the parameter values of gamma distribution, and its level is not more than 11% of the true variance.

Efficiency of the two estimators is compared in the last column of Table 1. The relative efficiency coefficient, given in expression (54), oscillates between 85.9% and 110.1%. The estimators of the expected value have comparable accuracy. Both estimation methods are unbiased. Their variances differ from each other by not more than 14.1%. However, the method based on a kernel-type estimator of the inclusion density function is preferable because it does not entail the assumption of bivariate gamma distribution.

5. Conclusion

This paper contributes to research on estimating of the mean value of the variable under study using continuous sampling designs. The well-known properties of the conditional distribution of the variable under study under an assumed value of the auxiliary variable and results from Cordy (1993) allow us to construct the estimator of the mean of the variable under study. It has been shown that this estimator is unbiased. The theorems presented in this paper also deal with estimating parameters other than the mean. These results allow us to consider a particular (inspired by Cox and Snell (1979)) sampling design with inclusion function dependent on the auxiliary variable. This provides a ratio-type estimator of the mean value. Estimation of the inclusion density function by means of a kernel-type estimator is also proposed. It does not need additional assumptions about density functions. From the results of a simulation study, we conclude that the expected value can be estimated more efficiently than by the sample mean.

Perhaps, additional studies could show, if the considered estimation method can be useful in statistical applications like auditing, insurance problems, and analysis of joint distributions of income and expenditures. There are many possibilities for modifying the sampling designs represented by continuous inclusion functions and their estimators. For instance, other kernels can be applied. We could apply classical statistical inference procedures for large sample sizes. All the considered estimators could be shown as sums of independent identically distributed random variables. Therefore, the well-known asymptotic methods of statistical inference could be used to constructions of confidence intervals and statistical tests. Moreover, there are possibilities for applying well-known bootstrap techniques to test statistical hypotheses or confidence interval estimation.

Acknowledgement

This paper is a result of a grant supported by the *National Science Centre, Poland*, no. 2016/21/B/HS4/00666.

References

BĄK, T., (2014). Triangular method of spatial sampling. *Statistics in Transition*, Vol. 15, No. 1, pp. 9–22. http://stat.gov.pl/en/sit-en/issues-and-articles-sit.

BĄK, T., (2018). An extension of Horvitz-Thompson estimator used in adaptive cluster sampling to continuous universe. *Communications in Statistics – Theory and Methods*, vol. 46, Issue 19, pp. 9777–9786, DOI: 10.1080/03610926.2016.1218028.

BENHENNI, K., CAMBANIS, S., (1992). Sampling Designs for Estimating Integrals of Stochastic Processes. *The Annals of Statistics*, Vol. 20, No. 1, pp. 161–194.

BUCKLEW, J. A., (2004). *Introduction to Rare Event Simulation*. Springer, New York, Berlin, Heidelberg, Hong Kong, London, Milan, Paris, Tokyo.

CORDY, C. B., (1993). An extension of the Horvitz-Thompson theorem to point sampling from a continuous universe. *Statistics and Probability Letters*, Vol. 18, pp. 353–362.

COX, D. R., SNELL, E. J., (1979). On sampling and the estimation of rare errors. *Biometrika*, Vol. 66, 1, pp. 125–32.

CHERIYAN, K. C., (1941). A bivariate correlated gamma-type distribution function. Journal of the Indian Mathematical Society, Vol. 5, pp. 133–144.

CRESSIE, N. A. C., (1993). *Statistics for Spatial Data*. Wiley, New York.

FROST, P. A., TAMURA, H., (1986). Accuracy of auxiliary information interval estimation in statistical auditing. *Journal of Accounting Research* 24, pp. 57–75.

GHIRTIS G. C., (1967). Some problems of statistical inference relating to double-gamma distribution. *Trabajos de Estatistica*, Vol. 18, pp. 67–87.

HORVITZ, D. G., THOMPSON, D. J., (1952). A generalization of the sampling without replacement from finite universe. *Journal of the American Statistical Association*, Vol. 47, pp. 663–685.

KOTZ, S., BALAKRISHNAN, JOHNSON, N. L., (2000). *Continuous Multivariate Distributions, Vol. 1: Models and Applications*. John Wiley & Sons, Inc., New York, Chichester, Wenheim, Brisbane, Sigapore, Toronto.

MCKAY, A. T., (1934). Sampling from batches. *Journal of the Royal Statistical Society* 2, pp. 207–216.

RIPLEY, B. D., (1987). *Stochastic Simulation*. Wiley, 1987, New York. Sarndal Särndal C. E., Swenson B., Wretman J. (1992). *Model Assisted Survey Sampling*. Springer Verlag, New York-Berlin-Heidelberg-London-Paris-Tokyo-Hong Kong-Barcelona-Budapest.

TILLÉ, Y., (2006). *Sampling Algorithms*. Springer.

THOMPSON, M. E., (1997). *Theory of Sample Survey.* Chapman & Hall, London, Weinheim, New York, Tokyo, Melbourne, Madras.

WILHELM M., TILLÉ, Y., QUALITÉ, M., (2017). Quasi-systematic sampling from a continuous population. *Computational Statistics & Data Analysis*, 105, pp. 11–23.

WILKS, S. S., (1962). *Mathematical Statistics*. John Wiley & Sons, New York, London.

WYWIAŁ, J. L., (2016). *Contributions to Testing Statistical Hypotheses in Auditing*. PWN, Warsaw.

WYWIAŁ, J. L., (2018). Application of two gamma distribution mixture to financial auditing. Sankhya B, Vol. 80, issue 1, pp. 1–18.

ZUBRZYCKI, S., (1958). Remarks on random, stratified and systematic sampling in a plane. *Colloquium Mathematicum*, Vol. 6, pp. 251–262. DOI: 10.4064/cm-6-1-251-264, http://matwbn.icm.edu.pl/ksiazki/cm/cm6/cm6135.pdf.