STATISTICS IN TRANSITION new series, Summer 2013
Vol.14, No. 2, pp. 231-248

SAMPLING DESIGNS PROPORTIONATE TO SUM OF TWO ORDER STATISTICS OF AUXILIARY VARIABLE

Janusz L. Wywiał[1]

ABSTRACT

In this paper the case of a conditional sampling design proportional to the sum of two order statistics is considered. Several strategies including the Horvitz-Thompson estimator and ratio-type estimators are discussed. The accuracy of these estimators is analyzed on the basis of computer simulation experiments.

Key words: sampling design, order statistic, sample quantile, auxiliary variable, Horvitz-Thompson statistic, inclusion probabilities, sampling scheme, ratio estimator.

1. Introduction

Sampling designs are usually constructed on the basis of an auxiliary variable observations in order to improve the accuracy of estimation of population parameters. For instance, the sampling design of Lahiri (1951), Midzuno (1952) and Sen (1953) proportional to the sample mean of the positive valued auxiliary variable leads to unbiasednees of the well known ordinary ratio estimator. Wywial (2008, 2009) proposed sampling designs dependent on order statistics of the auxiliary variable. Here that approach is continued because the sampling design proportional to the sum of two auxiliary variable order statistics is proposed. The conditional version of this sampling design is considered in order to improve the estimation effects. The review of the sampling designs or schemes dependent on auxiliary variables and their conditional versions is considered by Tillé (2006).

[1]Katowice University of Economics, Poland, janusz.wywial@ue.katowice.pl.

2. Sampling design

Let U be a fixed population of the size N. The observation of a variable under study and of an auxiliary variable are denoted by y_i and $x_i, i = 1, \ldots, N$, respectively. Moreover, let $0 < x_i \leq x_{i+1}, i = 1, \ldots, N-1$. Our problem is the estimation of the population average $\bar{y} = \frac{1}{N}\sum_{k\in U} y_k$. Let us consider the sample space **S** of the samples s of the fixed effective size $1 < n < N$. The sampling design is denoted by $P(s)$, so $P(s) \geq 0$ for all $s \in \mathbf{S}$ and $\sum_{s\in\mathbf{S}} P(s) = 1$.

Let $(X_{(j)})$ be the sequence of the order statistics of observations of auxiliary variable in the sample s. It is well known that the sample quantile of the order $\alpha \in (0;1)$ is defined as follows: $Q_{s,\alpha} = X_{(r)}$ where $r = [n\alpha] + 1$, the function $[n\alpha]$ means the integer part of the value $n\alpha$, $r = 1, 2, ..., n$. Let us note that $X_{(r)} = Q_{s,\alpha}$ for $\frac{r-1}{n} \leq \alpha < \frac{r}{n}$. In this paper it will be more convenient to consider the order statistic rather than the quantile. Let $G(r,u,i,j) = \{s : X_{(r)} = x_i, X_{(u)} = x_j\}$, $r = 1, .., n-1$; $u = 2, ..., n$, $r < u$ be the set of all samples whose r -th and u -th order statistics of the auxiliary variable are equal to x_i and x_j, respectively, where $r \leq i < j \leq N - n + u$.

$$\bigcup_{i=r}^{N-n+r} \bigcup_{j=i+u-r}^{N-n+u} G(r,u,i,j) = \mathbf{S}. \tag{1}$$

The size of the set $G(r,u,i,j)$ is denoted by $g(r,u,i,j) = Card(G(r,u,i,j))$ and

$$g(r,u,i,j) = \binom{i-1}{r-1}\binom{j-i-1}{u-r-1}\binom{N-j}{n-u}, \tag{2}$$

$$\binom{N}{n} = Card(\mathbf{S}) = Card\left(\bigcup_{i=r}^{N-n+r} \bigcup_{j=i+u-r}^{N-n+u} G(r,u,i,j)\right) =$$
$$= \bigcup_{i=r}^{N-n+r} \bigcup_{j=i+u-r}^{N-n+u} Card(G(r,u,i,j)) = \sum_{i=r}^{N-n+r} \sum_{j=i+u-r}^{N-n+u} g(r,u,i,j),$$

$$P\left(X_{(r)} = x_i, X_{(u)} = x_j\right) = \frac{g(r,u,i,j)}{\binom{N}{n}}. \tag{3}$$

Let $h(x_j, x_i)$ be a non-negative function of values x_j, x_i of the order statistics $X_{(u)}$ and $X_{(r)}$, respectively. Moreover let

$$f(x_j, x_i, c) = \begin{cases} h(x_j, x_i) & \text{if} \quad h(x_j, x_i) \geq c, \\ 0 \quad for & h(x_j, x_i) < c. \end{cases} \tag{4}$$

and

$$z(r,u,c)=\sum_{i=r}^{N-n+r}\sum_{j=i+u-r}^{N-n+u} f(x_j,x_i,c)g(r,u,i,j). \quad (5)$$

The straightfoward generalization of the Wywiał's (2009) sampling design is as follows.

Definition 1.1. The conditional sampling design proportional to the non-negative functions of the order statistics $X_{(u)}$, $X_{(r)}$ is as follows:

$$P_{r,u}(s|c)=\frac{f(x_j,x_i,c)}{z(r,u,c)} \qquad \text{for} \quad s\in G(r,u,i,j) \quad (6)$$

where $i<j, \quad r\le i\le N-n+r \quad and \quad r<u\le j\le N-n+u$, $0\le c\le c_0$.

As it is well known, the inclusion probability of the first order is determined by the equations: $\pi_k(r,u,c)=P_{r,u}(s:k\in s|c)=\sum_{\{s:k\in s\}}P_{r,u}(s|c), \quad k=1,...,N$.

Now, the upper possible value c_0 of the constant c should be stated in such a way that $\pi_k(r,u,c)>0$ for all $k=1,...,N$. It is important because under the just defined condition the well known Horvitz-Thompson estimator (considered in the sections 3) is unbiased for the a population mean.

The above defined sampling design is treated as conditional (unconditional) when $c>0$ $(c=0)$, see the definition of the conditional sampling design considered by Tillé (1999, 2006).

Particularly, let $h(x_j,x_i)=x_j+x_i$. Thus

$$f(x_j,x_i,c)=\begin{cases} x_j+x_i & for \quad x_j+x_i\ge c,\\ 0 \quad for & x_j+x_i<c.\end{cases} \quad (7)$$

We have to assume that $0\le c\le c_0=x_1+x_N$. Thus, under this assumption the inclusion probabilities $\pi_N(r,u,c)>0$ for all $i=1,...,N$. When $c>c_0$, $x_1+x_i<c$ for all $i=2,...,N$ and $\pi_1(r,u,c)=0$ and $\pi_k(r,u,c)\ge 0$ for $k=2,...,N$. In this case, as it is well known, the Horvitz-Thompson's statistic is a biased estimator of the population mean.

Let $\delta(x)$ be such a function that if $x\le 0$ then $\delta(x)=0$ otherwise $\delta(x)=1$. Moreover, $\delta(x)\delta(x-1)=\delta(x-1)$. The following two theorems are the straightforward generalizations of those ones proved by Wywiał(2009).

Theorem 1. *The inclusion probabilities of the first order for the conditional sampling design $P_{r,u}(s|c)$ are as follows:*

$$\pi_k(r,u,c)=$$
$$=\frac{1}{z(r,u,c)}\Bigg(\delta(r-1)\sum_{i=r\delta(r-k)+(k-1)\delta(k+1-r)\delta(N-n+r-k)}^{N-n+r}\ \sum_{j=i+u-r}^{N-n+u}\binom{i-2}{r-2}$$
$$\binom{j-i-1}{u-r-1}\binom{N-j}{n-u}f(x_j,x_i,c)+\delta(k-r)\delta(N-n+u-k)\delta(u-r-1)$$
$$\sum_{i=r}^{min(k-1,N-n+r)}\ \sum_{j=max(i+u-r,k+1)}^{N-n+u}\binom{i-1}{r-1}\binom{j-i-2}{u-r-2}\binom{N-j}{n-u}f(x_j,x_i,c)+$$
$$+\delta(k-u)\delta(n-u)\delta(N-n+u-k+1)$$
$$\sum_{i=r}^{k-u+r-1}\ \sum_{j=i+u-r}^{k-1}\binom{i-1}{r-1}\binom{j-i-1}{u-r-1}\binom{N-j-1}{n-u-1}f(x_j,x_i,c)+\delta(n-u)$$
$$\delta(k-N+n-u)\sum_{i=r}^{N-n+r}\ \sum_{j=i+u-r}^{N-n+u}\binom{i-1}{r-1}\binom{j-i-1}{u-r-1}\binom{N-j-1}{n-u-1}f(x_j,x_i,c)+$$
$$+\delta(k+1-r)\delta(N-n+r-k+1)\binom{k-1}{r-1}$$
$$\sum_{j=k+u-r}^{N-n+u}\binom{j-k-1}{u-r-1}\binom{N-j}{n-u}f(x_j,x_k,c)+\delta(k-u+1)\delta(N-n+u-k+1)$$
$$\binom{N-k}{n-u}\sum_{i=r}^{k-u+r}\binom{i-1}{r-1}\binom{k-i-1}{u-r-1}f(x_k,x_i,c)\Bigg)\quad(8)$$

Theorem 2. *The inclusion probabilities of the second order for the conditional sampling design $P_{r,u}(s|c)$ are as follows:*

$$\pi_{k,t}(r,u,c)=P\left(k,t\in s_1\right)+P\left(k\in s_1,X_{(r)}=x_t\right)+P\left(k\in s_1,t\in s_2\right)+$$
$$+P\left(k\in s_1,X_{(u)}=x_t\right)+P\left(k\in s_1,t\in s_3\right)+P\left(X_{(r)}=x_k,t\in s_2\right)+$$
$$+P\left(X_{(r)}=x_k,X_{(u)}=x_t\right)+P\left(X_{(r)}=x_k,t\in s_3\right)+P\left(k,t\in s_2\right)+$$
$$+P\left(k\in s_2,X_{(u)}=x_t\right)+P\left(k\in s_2,t\in s_3\right)+P\left(X_{(u)}=x_k,t\in s_3\right)+$$
$$+P\left(k,t\in s_3\right)\quad(9)$$

where

$$P\left(k,t\in s_1\right)=\frac{\delta(r-2)\delta(N-n+r-t)}{z(r,u)}\cdot$$
$$\cdot\sum_{i=max(r,t+1)}^{N-n+r}\sum_{j=i+u-r}^{N-n+u}\binom{i-3}{r-3}\binom{j-i-1}{u-r-1}\binom{N-j}{n-u}f(x_j,x_i,c).$$

$$P\left(k\in s_1,X_{(r)}=x_t\right)=$$
$$=\frac{\delta(r-1)\delta(N-n+r-k)\delta(N-n+r+1-t)\delta(t+1-r)}{z(r,u)}\cdot$$
$$\cdot\binom{t-2}{r-2}\sum_{j=t+u-r}^{N-n+u}\binom{j-t-1}{u-r-1}\binom{N-j}{n-u}f(x_j,x_t,c),$$

$$P\left(k\in s_1,t\in s_2\right)=\delta(N-n+r-k)\cdot$$
$$\cdot\frac{\delta(t-r)\delta(r-1)\delta(u-r-1)\delta(N-n+u-t)\delta(t-k-1)}{z(r,u)}\cdot$$
$$\cdot\sum_{i=max(r,k+1)}^{min(t-1,N-n+r)}\sum_{j=max(t+1,i+u-r)}^{N-n+u}\binom{i-2}{r-2}\binom{j-i-2}{u-r-2}\binom{N-j}{n-u}f(x_j,x_i,c),$$

$$P\left(k\in s_1,X_{(u)}=x_t\right)=\delta(N-n+r-k)\cdot$$
$$\cdot\frac{\delta(t+1-u)\delta(N-n+u+1-t)\delta(t-k-u+r)\delta(r-1)}{z(r,u)}\cdot$$
$$\cdot\binom{N-t}{n-u}\sum_{i=max(r,k+1)}^{t-u+r}\binom{i-2}{r-2}\binom{t-i-1}{u-r-1}f(x_t,x_i,c),$$

$$P\left(k\in s_1,t\in s_3\right)=$$
$$=\frac{\delta(N-n+r-k)\delta(t-u)\delta(r-1)\delta(n-u)\delta(t-k-u+r-1)}{z(r,u)}\cdot$$
$$\cdot\sum_{i=max(r,k+1)}^{min(t-u+r-1,N-n+r)}\sum_{j=i+u-r}^{min(t-1,N-n+u)}\binom{i-2}{r-2}\binom{j-i-1}{u-r-1}\binom{N-j-1}{n-u-1}\cdot$$
$$\cdot f(x_j,x_i,c),$$

$$P\left(X_{(r)}=x_k, t\in s_2\right)=$$
$$=\frac{\delta(k+1-r)\delta(N-n+r+1-k)\delta(t-r)\delta(N-n+u-t)\delta(u-r-1)}{z(r,u)}\cdot$$
$$\cdot\binom{k-1}{r-1}\sum_{j=max(t+1,k+u-r)}^{N-n+u}\binom{j-k-2}{u-r-2}\binom{N-j}{n-u}f(x_j,x_k,c),$$

$$P\left(X_{(r)}=x_k, X_{(u)}=x_t\right)=$$
$$=\frac{\delta(k+1-r)\delta(N-n+r+1-k)\delta(t+1-u)\delta(N-n+u-t+1)}{z(r,u)}\cdot$$
$$\cdot\,\delta(t+1-k-u+r)\binom{k-1}{r-1}\binom{t-k-1}{u-r-1}\binom{N-t}{n-u}f(x_t,x_k,c),$$

$$P\left(X_{(r)}=x_k, t\in s_3\right)=$$
$$=\frac{\delta(k+1-n)\delta(N-n+r+1-k)\delta(t-u)\delta(n-u)\delta(t-k-u+r)}{z(r,u)}\cdot$$
$$\cdot\binom{k-1}{r-1}\sum_{j=k+u-r}^{min(t-1,N-n+u)}\binom{j-k-1}{u-r-1}\binom{N-j-1}{n-u-1}f(x_j,x_k,c),$$

$$P\left(k,t\in s_2\right)=\frac{\delta(k-r)\delta(N-n+u-t)\delta(u-r-2)}{z(r,u)}\cdot$$
$$\cdot\sum_{i=r}^{min(k-1,N-n+r)}\sum_{j=max(t+1,i+u-r)}^{N-n+u}\binom{i-1}{r-1}\binom{j-i-3}{u-r-3}\binom{N-j}{n-u}f(x_j,x_i,c),$$

$$P\left(k\in s_2, X_{(u)}=x_t\right)=$$
$$=\frac{\delta(k-r)\delta(N-n+u-k)\delta(t+1-u)\delta(N-n+u+1-t)\delta(u-r-1)}{z(r,u)}\cdot$$
$$\cdot\binom{N-t}{n-u}\sum_{i=r}^{min(k-1,t-u+r)}\binom{i-1}{r-1}\binom{t-i-2}{u-r-2}f(x_t,x_i,c),$$

$$P\left(k\in s_2,t\in s_3\right)=$$
$$=\frac{\delta(k-r)\delta(N-n+u-k)\delta(t-u)\delta(t-k-1)\delta(u-r-1)\delta(n-u)}{z(r,u)}\cdot$$
$$\cdot\sum_{i=r}^{min(k-1,t-u+r-1,N-n+r)}\sum_{j=max(i+u-r,k+1)}^{min(t-1,N-n+u)}\binom{i-1}{r-1}\binom{j-i-2}{u-r-2}\binom{N-j-1}{n-u-1}\cdot$$
$$\cdot f(x_j,x_i,c),$$

$$P\left(X_{(u)}=x_k,t\in s_3\right)=$$
$$=\frac{\delta(k+1-u)\delta(N-n+u+1-k)\delta(t-u)\delta(n-u)}{z(r,u)}\cdot$$
$$\cdot\binom{N-k-1}{n-u-1}\sum_{i=r}^{k-n+r}\binom{i-1}{r-1}\binom{k-i-1}{u-r-1}f(x_k,x_i,c),$$

$$P\left(k,t\in s_3\right)=\frac{\delta(k-u)\delta(n-u-1)}{z(r,u)}\cdot$$
$$\cdot\sum_{i=r}^{min(N-n+r,k-1-u+r)}\sum_{j=i+u-r}^{min(k-1,N-n+u)}\binom{i-1}{r-1}\binom{j-i-1}{u-r-1}\binom{N-j-2}{n-u-2}\cdot$$
$$\cdot f(x_j,x_i,c).$$

3. Sampling scheme

The sampling scheme implementing the sampling design $P_{r,u}(s|c)$ is as follows. Firstly, population elements are ordered according to increasing values of the auxiliary variable. Let $s=s_1\cup\{i\}\cup s_3\cup\{j\}\cup s_3$ where $s_1=\{k:k\in U,x_k<x_i\}$ is the simple random sample of the size $r-1$ drawn without replacement from the subpopulation $U(1,i-1)=(1,...,i-1)$, $s_2=\{k:k\in U,x_j>x_k>x_i\}$ is the simple random sample of the size $u-r-1$ drawn without replacement from $U(i+1,j-1)=(i+1,...,j-1)$ and $s_3=\{k:k\in U,x_k>x_j\}$ s the simple random sample of the size $n-u$ drawn without replacement from $U(j+1,N)=(j+1,...,N)$. Let us note that $U=U(1,i-1)\cup\{i\}\cup U(i+1,j-1)\cup\{j\}\cup U(j+1,N)$. Let $\mathbf{S}\left(U(1,i-1);s\right)$ be sample space of the sample s_1, let $\mathbf{S}\left(U(i+1,j-1);s\right)$ be sample space of the sample s_2 and let $\mathbf{S}\left(U(j+1,N);s\right)$ be sample space of the sample s_3. Moreover, $\mathbf{S}=\mathbf{S}\left(U,s\right)$).

The sampling scheme is given by the following probabilities:

$$P_{r,u}(s|c) = P_1(s_1|i)p_{r,u}(i|c)P_2(s_2|i,j)p_{r,u}(j|i,c)P_3(s_3|j) \tag{10}$$

where

$$P_1(s_1|i) = \binom{i-1}{r-1}^{-1}, \quad P_2(s_2|i,j) = \binom{j-i-1}{u-r-1}^{-1},$$

$$P_3(s_3|j) = \binom{N-j}{n-u}^{-1},$$

$$p_{r,u}(j|i,c) = \frac{p_{r,u}(i,j|c)}{p_{r,u}(i|c)}, \tag{11}$$

$$p_{r,u}(i,j|c) = \sum_{s \in G(r,u,i,j)} P_{r,u}(s) = \frac{f(x_j, x_i, c)g(r,u,i,j)}{z(r,u,c)}, \tag{12}$$

$$p_{r,u}(i|c) = \frac{1}{z(r,u,c)} \sum_{j=i+u-r}^{N-n+u} f(x_j, x_i, c)g(r,u,i,j). \tag{13}$$

In order to select the sample s, firstly the i-th element of the population should be selected according to the probability function $p_{r,u}(i|c)$. Next, the j-the element of the population should be drawn according to the probability function $p_{r,u}(j|i,c)$. Finally, the samples s_1, s_2 and s_3 should be selected according to the sampling designs $P_1(s_1)$, $P_2(s_2)$ and $P_3(s_3)$, respectively.

4. Some sampling strategies

The well known Horvitz-Thompson estimator (1952) is as follows:

$$\bar{y}_{HT,s} = \frac{1}{N} \sum_{k \in s} \frac{y_k}{\pi_k}. \tag{14}$$

The statistic is unbiased estimator of the population mean value if $\pi_k > 0$ for $k = 1, ..., N$. The variance and its estimator are determined by the expressions (20) and (22), respectively.

The particular case of the above estimator is the well known sampling design of the simple random sample drawn without replacement whose sampling design is: $P_0(s) = \binom{N}{n}^{-1}$. The variance of the mean from the simple random sample $\bar{y}_s = \frac{1}{n}\sum_{k \in s} y_k$ drawn without replacement is $D^2(\bar{y}_s, P_0(s)) = \frac{N-n}{nN} v(y)$ where $v(y) = \frac{1}{N-1}\sum_{k=1}^{N}(y_k - \bar{y})^2$.

Let us construct the following ratio sampling strategy for the population mean $\bar{y} = \frac{1}{N}\sum_{i \in U} y_i$. We assume that $y_i = bx_i + e_i$ for all $i \in U$, $\sum_{i \in U} e_i =$

0 and the residuals of that linear regression function are not correlated with the auxiliary variable. Thus, it has been assumed that the intercept of the linear regression function is equal to zero. The linear correlation coefficient between the variables y and x will be denoted by ρ. Let $\left(X_{(r)}, Y_r\right)$ be two dimensional random variables where $X_{(r)}$ is the r-th order statistic of an auxiliary variable and Y_r is the variable under study. Let us define the following ratio type estimator:

$$\bar{y}_{r,u,s} = \bar{y}_s \frac{E\left(X_{(r)} + X_{(u)}|c\right)}{X_{(r)} + X_{(u)}} \tag{15}$$

where on the basis of the expressions (3) and (5) we have

$$\begin{aligned} &E\left(X_{(r)} + X_{(u)}|c\right) = \\ &= \sum_{i=r}^{N-n+r} \sum_{j=i+u-r}^{N-n+u} f(x_j, x_i, c) P\left(X_{(r)} = x_i, X_{(u)} = x_j | X_{(r)} + X_{(u)} \geq c\right) = \\ &\qquad \sum_{i=r}^{N-n+r} \sum_{j=i+u-r}^{N-n+u} f(x_j, x_i, c) \frac{P\left(X_{(r)} = x_i, X_{(u)} = x_j\right)}{P\left(X_{(r)} + X_{(u)} \geq c\right)} = \\ &\qquad = \frac{\sum_{i=r}^{N-n+r} \sum_{j=i+u-r}^{N-n+u} f(x_j, x_i, c) g(r,u,i,j)}{\sum_{i=r}^{N-n+r} \sum_{j=i+u-r}^{N-n+u} \gamma(x_j, x_i, c) g(r,u,i,j)} = \frac{z(r,u,c)}{\alpha(r,u,c)}, \end{aligned} \tag{16}$$

$$\alpha(r,u,c) = \sum_{i=r}^{N-n+r} \sum_{j=i+u-r}^{N-n+u} \gamma(x_j, x_i, c) g(r,u,i,j) = \binom{N}{n} P\left(X_{(r)} + X_{(u)} \geq c\right),$$

$$\gamma(x_i, x_j, c) = \begin{cases} 1 & \text{for} \quad x_i + x_j \geq c \\ 0 & \text{for} \quad x_i + x_j < c. \end{cases}$$

Let $\mathbf{S}_c = \{s : x_{(r)} + x_{(u)} \geq c\}$ and $\mathbf{S}_{\bar{c}} = \{s : x_{(r)} + x_{(u)} < c\} = \mathbf{S} - \mathbf{S}_c$ where $(x_{(r)}, x_{(u)})$ are values of the order statistics $(X_{(r)}, X_{(u)})$. Moreover, let $\bar{y}_c = \frac{1}{Card(\mathbf{S}_c)} \sum_{s \in \mathbf{S}_c} \bar{y}_s$, $\bar{y}_{\bar{c}} = \frac{1}{Card(\mathbf{S}_{\bar{c}})} \sum_{s \in \mathbf{S}_{\bar{c}}} \bar{y}_s$ where $Card(\mathbf{S}_c) = \alpha(r,u,c)$. Under the stated assumptions we have:

$$\begin{aligned} &E\left(\bar{y}_{r,u,s}, P_{r,u}(s|c)\right) = \sum_{s \in \mathbf{S}_c} \bar{y}_s \frac{E\left(X_{(r)} + X_{(u)}|c\right)}{x_{(r)} + x_{(u)}} P_{r,u}(s|c) = \\ &\quad = \sum_{s \in \mathbf{S}_c} \bar{y}_s \frac{z(r,u,c)}{\alpha(r,u,c)\left(x_{(r)} + x_{(u)}\right)} \frac{x_{(r)} + x_{(u)}}{z(r,u,c)} = \frac{1}{Card(\mathbf{S}_c)} \sum_{s \in \mathbf{S}_c} \bar{y}_s = \bar{y}_c. \end{aligned}$$

Particularly, if $c=0$ then $P_{r,u}(s|0)=P_{r,u}(s)$ and

$$E\left(\bar{y}_{r,u,s},P_{r,u}(s)\right)=\sum_{s\in\mathbf{S}}\bar{y}_s\frac{E\left(X_{(r)}+X_{(u)}\right)}{x_{(r)}+x_{(u)}}P_{r,u}(s)=\frac{1}{\binom{N}{n}}\sum_{s\in\mathbf{S}}\bar{y}_s=\bar{y}.$$

These results and the decomposition:

$$\begin{aligned}\bar{y}=\frac{1}{\binom{N}{n}}\sum_{s\in\mathbf{S}}\bar{y}_s=\frac{1}{\binom{N}{n}}\left(\sum_{s\in\mathbf{S}_c}\bar{y}_s+\sum_{s\in\mathbf{S}_{\bar{c}}}\bar{y}_s\right)=&\\ =\frac{1}{\binom{N}{n}}\left(\alpha(r,u,c)\bar{y}_c+(1-\alpha(r,u,c))\bar{y}_{\bar{c}}\right)=&\\ =\bar{y}_cP\left(X_{(r)}+X_{(u)}\geq c\right)+\bar{y}_{\bar{c}}P\left(X_{(r)}+X_{(u)}<c\right)&\end{aligned}$$

lead to the following expression:

$$E\left(\bar{y}_{r,u,s},P_{r,u}(s|c)\right)=\begin{cases}\bar{y} & \text{for}\quad c=0\\ \bar{y}+\left(\bar{y}_c-\bar{y}_{\bar{c}}\right)P\left(X_{(r)}+X_{(u)}<c\right) & \text{for}\quad c>0.\end{cases}\tag{17}$$

Hence, under the unconditional sampling design $P_{r,u}(s)$ the strategy $(\bar{y}_{r,u,s},P_{r,u}(s|c))$ is unbiased.

The next ratio type estimator, see e.g. Särndal at. all (1992), is as follows:

$$\tilde{y}_s=\bar{y}_{HT,s}\frac{\bar{x}}{\bar{x}_{HT,s}}\tag{18}$$

The parameters of the strategy $(\tilde{y}_s,P_{r,u}(s|c))$ are approximately as follows:

$$E\left(\tilde{y}_s,P_{r,u}(s|c)\right)\approx\bar{y},$$

$$\begin{aligned}D^2\left(\tilde{y}_s,P_{r,u}(s|c)\right)\approx D^2\left(\bar{y}_{HT,s},P_{r,u}(s|c)\right)-2hCov\left(\bar{y}_{HT,s},\bar{x}_{HT,s},P_{r,u}(s|c)\right)+&\\ +h^2D^2\left(\bar{x}_{HT,s},P_{r,u}(s|c)\right)&\end{aligned}\tag{19}$$

where $h=\frac{\bar{y}}{\bar{x}}$ and

$$Cov\left(\bar{y}_{HT,s},\bar{x}_{HT,s},P_{r,u}(s|c)\right)=\frac{1}{N^2}\left(\sum_{k\in U}\sum_{l\in U}\Delta_{k,l}\frac{y_k}{\pi_k}\frac{x_l}{\pi_l}\right),\tag{20}$$

$\Delta_{k,l}=\pi_{k,l}-\pi_k\pi_l,\qquad D^2\left(\bar{x}_{HT,s},P_{r,u}(s|c)\right)=Cov\left(\bar{x}_{HT,s},\bar{x}_{HT,s},P_{r,u}(s|c)\right),$
$D^2\left(\bar{y}_{HT,s},P_{r,u}(s|c)\right)=Cov\left(\bar{y}_{HT,s},\bar{y}_{HT,s},P_{r,u}(s|c)\right).$

The variance: $D^2\left(\bar{y}_{r,u,s}, P_{r,u}(s|c)\right)$ can be estimated by the following approximately unbiased estimator:

$$\hat{D}^2\left(\tilde{y}_s, P_{r,u}(s|c)\right) = \hat{D}^2\left(\bar{y}_{HT,s}, P_{r,u}(s|c)\right) + \\ - 2h_{r,u,s}\widehat{Cov}\left(\bar{y}_{HT,s}, \bar{x}_{HT,s}, P_{r,u}(s|c)\right) + h^2_{r,u,s}\hat{D}^2\left(\bar{x}_{HT,s}, P_{r,u}(s|c)\right) \quad (21)$$

where

$$h_{r,u,s} = \frac{\bar{y}_{HT,s}}{\bar{x}_{HT,s}},$$

$$\widehat{Cov}\left(\bar{y}_{HT,s}, \bar{x}_{HT,s}, P_{r,u}(s|c)\right) = \frac{1}{N^2}\left(\sum_{k\in s}\sum_{l\in s}\Delta_{*,k,l}\frac{y_k}{\pi_k}\frac{x_l}{\pi_l}\right), \quad (22)$$

$\Delta_{*k,l} = \frac{\Delta_{k,l}}{\pi_{k,l}}$, $\quad \hat{D}^2\left(\bar{x}_{HT,s}, P_{r,u}(s|c)\right) = \widehat{Cov}\left(\bar{x}_{HT,s}, \bar{x}_{HT,s}, P_{r,u}(s|c)\right)$,
$\hat{D}^2\left(\bar{y}_{HT,s}, P_{r,u}(s|c)\right) = \widehat{Cov}\left(\bar{y}_{HT,s}, \bar{y}_{HT,s}, P_{r,u}(s|c)\right)$.

Let us remind the following ordinary ratio estimator.

$$\hat{y}_s = \bar{y}_s\frac{\bar{x}}{\bar{x}_s}. \quad (23)$$

The approximate value of the variance is as follows:

$$D^2\left(\hat{y}_s, P_0(s)\right) \approx \frac{N(N-n)}{n}\left(v(y) + h^2v(x) - 2hv(x,y)\right)$$

where $v(x,y) = \frac{1}{N-1}\sum_{k=1}^{N}(x_k-\bar{x})(y_k-\bar{y})$ and particularly $v(y) = v(y,y)$, $v(x) = v(x,x)$.

The approximately unbiased estimator of the variance is as follows:

$$\hat{D}^2\left(\hat{y}_s, P_0(s)\right) = \frac{N(N-n)}{n}\left(v_s(y) + h_s^2v_s(x) - 2h_sv_s(x,y)\right),$$

where $h_s = \frac{\bar{y}_s}{\bar{x}_s}$, $v_s(x,y) = \frac{1}{n-1}\sum_{k\in s}(x_k-\bar{x}_s)(y_k-\bar{y}_s)$ and particularly $v_s(y) = v_s(y,y)$, $v_s(x) = v_s(x,x)$. The strategy $(\hat{y}_s, P_0(s))$ is approximately unbiased for the population mean $\bar{y}$.

It is well known that the strategy $(\hat{y}_s, P_1(s))$ is unbiased for $\bar{y}$ where

$$P_1(s) = \frac{\bar{x}_s}{\binom{N}{n}} \quad (24)$$

is the sampling design of Lahiri (1951), Midzumo (1952) and Sen (1953).

5. Simulation analysis of strategies' accuracy

The population taken into account consists of the municipalities in Sweden whose number is $N = 284$. The value x_k, $k = 1, ..., N$, of the auxiliary variable x is equal to the size (in thousands) of people population in the k-th municipality in 1975. The value y_k, $k = 1, .., N$, of the variable under study y is the taxation revenues (in millions of kronor) from the k-th municipality in 1985. Their observations were published by Särndal, Swenson and Wretman (1992), pp. 652-659. number is $N = 284$.

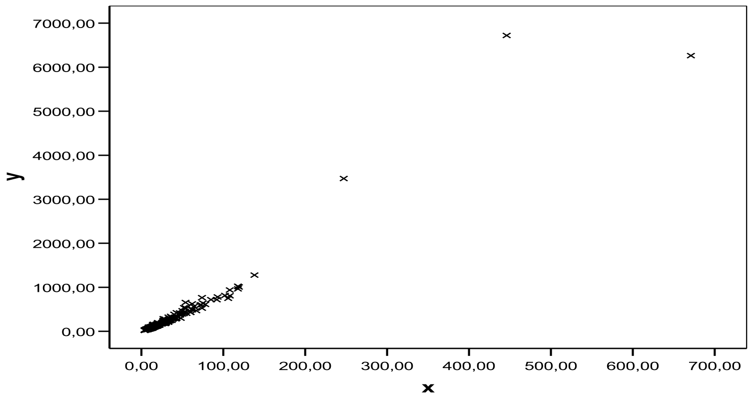

FIGURE 1. Scatterplot of y versus x.

There are three outlier observations of the variable as it is shown in Figure 1. Let σ and β_3 be the standard deviation and the skewness coefficient, respectively, of the auxiliary variable in the population. In case of data without outliers (the number of municipalities is 281) $\bar{x} = 24,263$, $\sigma = 24,153$ and $\beta_3 = 0,043$. In case of data with outliers (the number of municipalities is N=284) $\bar{x} = 28.810$, $\sigma = 52,873$ and $\beta_3 = 8,427$. Thus, the distribution of the variable without the outliers is almost symmetric. But the distribution of the variables with outliers is highly right skewed. The samples according to the preassigned sampling design were drawn from the just presented population. The samples were replicated 1000 times.

The Figure 1 shows that the dependence between the variable under study and the auxiliary variable can be approximated by means of a linear regression with its constance equal to zero. In this case, as it is well-known, the accuracies of regression type estimators of a population mean are similar to the accuracy of the ratio type estimators. Moreover, the ratio estimators are simpler than the regression ones. Thus, that is why we consider only the ratio estimator in the analysis.

Let $MSE(t, P(s))$ be the mean square error of the strategy $(t, P(s))$ used to estimate the population mean $\bar{y}$. The coefficient of the relative efficiency is defined as follows:

$$e(t, P(s)) = \frac{MSE(t, P(s))}{D^2(\bar{y}_s, P_0(s))} 100\%$$

The results of the simulation analysis are presented by Tables and Figures 2, 3. The tables show the relative efficiency coefficients. The distributions of the estimators, generated on the basis of samples of the size $n = 14$ drawn by means of appropriate sampling schemes are presented by means of the well-known box-plots on Figures 2 and 3. Table 1 shows the notation of the strategies.

TABLE 1. The symbols of the strategies.

strategy	symbol	efficiency
$\bar{y}_s, P_0(s)$	$ty0$	$-$
$\hat{y}_s, P_0(s)$	$ty1$	e_1
$\hat{y}_s, P_1(s)$	$ty2$	e_2
$\bar{y}_{HT,s}, P_1(s)$	$ty3$	e_3
$\bar{y}_{HT,s}, P_{n-1,n}(s)$	$ty4$	e_4
$\bar{y}_{HT,s}, P_{n-1,n}(s\|3\bar{x})$	$ty4d3$	e_4
$\tilde{y}_s, P_{n-1,n}(s)$	$ty5$	e_5
$\tilde{y}_s, P_{n-1,n}(s\|3\bar{x})$	$ty5d3$	e_5
$\bar{y}_{n-1,n,s}, P_{n-1,n}(s)$	$ty6$	e_6
$\bar{y}_{n-1,n,s}, P_{n-1,n}(s\|3\bar{x})$	$ty6d3$	e_6

TABLE 2. The relative efficiency coefficients (%) of the strategies.

N :	281			284		
n	$e1$	$e2$	$e3$	$e1$	$e2$	$e3$
2 (0,7%)	2.50	2.24	17.62	1.56	1.15	5.60
3 (1%)	2.88	2.58	28.10	1.88	1.35	8.64
6 (2%)	3.09	2.86	46.62	2.73	1.96	14.87
9 (3%)	3.28	3.09	57.95	3.78	2.66	20.49
14 (5%)	3.42	3.30	73.59	5.04	3.62	29.33
29 (10%)	3.22	3.14	84.51	7.22	5.92	47.44

Firstly, let us consider the strategies under the unconditional sampling designs. Table 2 shows the relative efficiency coefficients of the strategies which do not depend on the sampling design $P_{r,u}(s|c)$. The ratio estimator under the sampling design $P_1(s)$ is slightly better than the ratio estimator under the simple random sample and they are both significantly more accurate than the Horvtiz-Thompson estimator under the sampling design $P_1(s)$. In general, Tables 2, 3 and 4 let us infer that the ratio type strategies are significantly better than the Horvitz-Thompson ones. This conclusion is strongly confirmed by the box-plots shown by Figures 2 and 3. The strategy $(\hat{y}_s, P_1(s))$ is the best among the six considered strategies except for the case of the population with outliers when the strategy $(\tilde{y}_s, P_{n-1,n}(s))$ is the best for $n = 14$ $n = 29$.

Wywiał(2007) considered the accuracy of estimation on the basis of the sampling design proportional to one order statistic denoted by $P_r(s)$. For some of possible values r of the sampling design $P_r(s)$ the mean squares of the estimators were determined on the basis of the simulation analysis. The results of the analysis lead to the conclusion that the considered strategies dependent on sampling design $P_r(s)$ are the most accurate when $r = n - 1$ or $r = n$. That is why Tables 3 and 4 deal only with the case when $r = n-1$ or $r = n$. In general, all the inclusion probabilities of the first order are expected to be proportional to the appropriate values of a positively valued auxiliary variable. Thus, in the case considered we can suppose that if $r < n - 1$ and $r < u \leq n$, the inclusion probabilities $\pi_k(r, u, c)$ are not so proportional to x_k as in the case when $r = n - 1$ and $u = n$ for all $k = 1, ..., N$.

TABLE 3. The relative efficiency coefficients (%) of the conditional strategies for $P_{n-1,n}(s|k\bar{x})$, $k = 0, 1, 2, 3$. The population with outliers, $N = 284$.

$k\bar{x}$	0			$\bar{x}$			$2\bar{x}$			$3\bar{x}$		
n	$e4$	$e5$	$e6$	$e4$	$e5$	$e6$	$e4$	$e5$	$e6$	$e4$	$e5$	$e6$
2	6.5	1.3	2.0	1.9	8.6	2.2	4.3	0.7	3.2	9.8	5.4	3.2
3	9.2	1.5	1.9	3.5	1.1	2.0	3.9	0.9	3.0	6.3	0.5	3.5
6	18.2	2.3	3.0	11.8	2.0	2.9	5.6	1.5	3.0	5.7	0.9	3.8
9	21.4	2.9	5.5	17.8	2.8	5.4	9.0	2.0	4.4	6.2	1.3	4.8
14	25.6	3.4	13.9	23.2	3.2	13.5	15.3	2.8	10.0	8.4	1.9	7.7
29	34.5	4.5	56.8	34.9	4.6	58.2	29.1	4.4	48.7	19.8	3.6	35.1

The analysis of Tables 3, 4 and Figures 2, 3 leads to the following conclusion. The relative accuracies of the sampling strategies for sampling designs $P_{r,u}(s|c)$ are usually better in case of the population with outliers or extreme values.

The Figures 2 and 3 let us infer that in the case of the population without outliers values the distributions of the estimators are almost symmetric. In the case of the population with outliers values the estimators are distributed with a large number of outliers. In this situation we cannot expect an accurate estimation. But the analysis of the Figures 2 and 3 lets us say that in the case of the population with outliers the conditional sampling design (for $c > 0$) leads to reduction of the number of outliers observations in the distribution of appropriate estimators. The conditional strategies which depend on the sampling design $P_{r,u}(s|c)$ are usually more accurate than their appropriate unconditional versions (for $c = 0$). The estimators of the conditional strategies are unbiased or negligible biased estimators of the population mean. When the conditioning value c increases (the considered levels: $c = 0$, $c = \bar{x}$, $c = 2\bar{x}$, and $c = 3\bar{x}$) the accuracies of the strategies $(\tilde{y}_s, P_{r,u}(s|c))$ and $(\bar{y}_{HT,s}, P_{r,u}(s|c))$ usually increase, too. The accuracy of the strategy $(\tilde{y}_s, P_{r,u}(s|c))$ is the best among the conditional ones. The strategy $(\bar{y}_{r,u,s}, P_{r,u}(s|c))$ is not worse than $(\bar{y}_{HT,s}, P_{r,u}(s|c))$ except for the case of the population with outliers and of the sample size $n = 29$.

TABLE 4. The relative efficiency coefficients (%) of the conditional strategies for $P_{n-1,n}(s|k\bar{x})$, $k = 0, 1, 2, 3$. The population without outliers, $N = 281$.

$k\bar{x}$	0			$\bar{x}$			$2\bar{x}$			$3\bar{x}$		
n	$e4$	$e5$	$e6$	$e4$	$e5$	$e6$	$e4$	$e5$	$e6$	$e4$	$e5$	$e6$
2	17.7	2.3	2.7	3.8	1.6	2.1	17.6	2.5	1.2	32.3	3.0	0.8
3	24.5	2.3	3.1	8.0	1.7	3.0	15.9	2.0	2.2	14.3	2.4	1.9
6	41.8	2.5	12.7	30.3	2.2	11.8	17.1	2.2	2.3	14.8	2.3	1.8
9	51.8	2.8	25.3	50.6	2.8	22.8	26.3	2.3	16.8	26.8	2.3	11.3
14	63.8	3.1	44.6	59.9	3.0	39.6	39.9	2.8	28.7	36.4	2.2	22.1
29	81.7	3.1	84.3	75.0	3.1	82.9	71.8	2.9	69.1	56.9	2.9	52.4

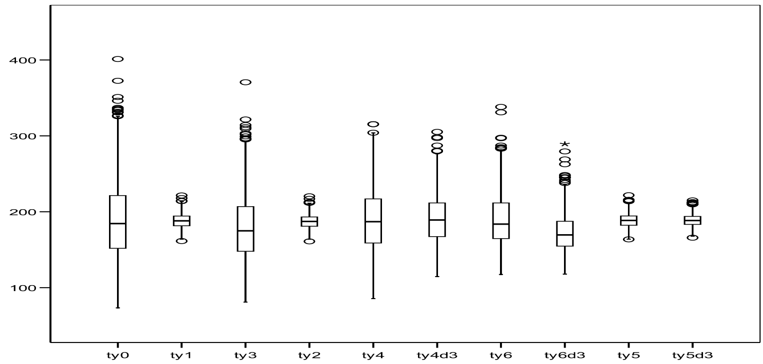

FIGURE 2. Boxplot of the estimator distributions in the population without outliers for n=14

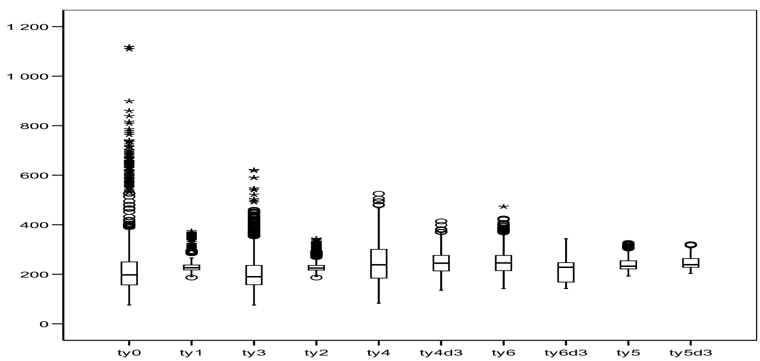

FIGURE 3. Boxplot of the estimator distributions in the population without outliers for n=14

6. Conclusions

The inclusion probabilities of the conditional sampling design proportionate to the sum of two order statistics are presented. They let us determine the variance of the Horvitz-Thompson estimator as well as its estimate.

The simulation analysis lets us expect the estimation strategies with the sampling design $P_{n-1,n}(s|c)$ not to be less accurate than the strategies with sampling design $P_{r,u}(s|c)$.

In general, the accuracies of the considered ratio type strategies: $(\tilde{y}_s, P_{n-1,n}(s|c))$, $(\hat{y}_s, P_1(s))$ or $(\hat{y}_s, P_0(s))$ are the best among the all strategies considered in the analysis. The accuracies of these three strategies are comparable. Moreover, the conditional strategies for $c = k\bar{x}$, $k > 0$ are slightly better than the appropriate unconditional ones.

Let us underline that the above conclusions cannot be treated as sufficiently general because they have been derived on the basis of a partial computer simulation analysis based on special data set taken into account. But it seems that those results can be an inspiration for larger simulation analyses or studies of the formal properties of the sampling design.

Acknowledgement

The research was supported by the grant number N N111 434137 from the Ministry of Science and Higher Education.

REFERENCES

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, 663–685.

LAHIRI G. W., (1951). A method for sample selection providing unbiased ratio estimator. *Bulletin of the International Statistical Institute,* Vol. 33, pp. 133–140.

MIDZUNO H., (1952). On the sampling system with probability proportional to the sum of sizes. *Annals of the Institute of Statistical Mathematics*, Vol. 3, pp. 99–107.

SA¨RNDAL C. E., SWENSSON B., WRETMAN J., (1992). *Model Assisted Survey Sampling.* Springer Verlag, New York-Berlin-Heidelberg-London-Paris-Tokyo-Hong Kong-Barcelona-Budapest.

SEN A. R., (1953). On the estimate of variance in sampling with varying probabilities *Journal of the Indian Society of Agicultural Statistics*, **5**, 2, pp. 119–127.

TILLE´ Y., (1999). Estimation in Surveys Using Conditional Inclusion Probabilities: Complex Design. *Survey Methodology*, Vol. 25, No 1, pp. 57–66.

TILLE´ Y., (2006). *Sampling algorithms*. Springer.

WYWIAŁ, J. L., (2007). Simulation analysis of accuracy estimation of population mean on the basis of strategy dependent on sampling design proportionate to the order statistic of an auxiliary variable. *Statistics in Transition-new series*, Vol. 8, No. 1, pp. 125–137.

WYWIAŁ, J. L., (2008). Sampling design proportional to order statistic of auxiliary variable. *Statistical Papers,* Vol. 49, No. 2, pp. 277–289.

WYWIAŁ, J. L., (2009). Performing quantiles in regression sampling strategy. *Model Assisted Statistics and Applications,* Vol. 4, No. 2, pp. 131–142.