Janusz L. Wywiał
Uniwersytet Ekonomiczny w Katowicach

SAMPLING DESIGNS PROPORTIONATE TO NON-NEGATIVE FUNCTIONS OF TWO QUANTILES OF AUXILIARY VARIABLE

1. Sampling design

Let U be a fixed population of size N. The observation of a variable under study and an auxiliary variable are denoted by y_i and x_i, $i = 1, \ldots, N$, respectively. Moreover, let $x_i \leq x_{i+1}$, $i = 1, \ldots, N-1$. Our problem is estimation of the population average $\bar{y} = \frac{1}{N}\sum_{k\in U} y_k$.

Let us consider the sample space $\mathbf{S}$ of the samples s of the fixed effective size $1 < n < N$. The sampling design is denoted by $P(s)$. We assume that $P(s) \geq 0$ 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 order a $\alpha \in (0;1)$ 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 $[n\alpha]$, $r = 1, 2, \ldots, n$. 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 conveniently to consider the order statistic than the quantile.

Let $G(r,u,i,j) = \{s : X_{(r)} = x_i, X_{(u)} = x_j\}$, $r = 1, \ldots, n-1$; $u = 2, \ldots, n$, $r < u$ 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 $r \leq i < j \leq N-n+u$. Moreover,

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

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

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

Let $f(x_j, x_j, c)$ be non-negative function of values of the order statistics $X_{(r)}$ and $X_{(r)}$ and

$$g(r,u,i,j) = \binom{i-1}{r-1}\binom{j-i-1}{u-r-1}\binom{N-j}{n-u} \tag{3}$$

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:

$$P_{r,u}(s \mid c) = \frac{f(X_{(u)}, X_{(r)}, x)}{z(r,u,c)} \tag{4}$$

where $i < j$ and $r \le i \le N-n+r$ and $r < u \le j \le N-n+u$.

Particularly, let $f(x_j, x_i, c) = x_j - x_i$ and

$$f(x_j, x_i, c) = \begin{cases} x_j - x_i \ \textit{for}\ x_j - x_i \ge c, \\ 0 \ \textit{for}\ x_j - x_i < c. \end{cases} \tag{5}$$

We say that the above sampling design is (conditional) unconditional when $(c > 0)$ $c = 0$. In general this concept is agree with definition of the conditional sampling design considered by Tillé (1999; 2006).

As it is well known the inclusion probability of the first order is determined by the equations: $\pi_k(r,u,c) = P_{r,u}(s : k \in s) = \sum_{\{s:k\in s\}} P_{r,u}(s \mid c)$, $k = 1, \ldots, N$. We assume that if $x \le 0$ then $\delta(x) = 0$ otherwise $\delta(x) = 1$. Let us note that $\delta(x)\delta(x-1) = \delta(x-1)$.

Theorem 1.1. Under the sampling design $P_{r,u}(s)$ the inclusion probabilities of the first degree are as follows. If $k < r$,

$$\pi_k(r,u,c) = \frac{\delta(r-1)\delta(r-k)}{z(r,u,c)} \sum_{i=r}^{N-n+r} \sum_{j=i+u-r}^{N-n+u} \binom{i-2}{r-2}\binom{j-i-1}{u-r-1}\binom{N-j}{n-u} f(x_j, x_i, c) \tag{6}$$

If $r \le k \le N-n+u$,

$$\pi_k(r,u,c) = \frac{\delta(r-1)\delta(r-k)}{z(r,u,c)} \cdot$$
$$\cdot \Bigg(\delta(k-u)\delta(n-u) \sum_{i=r}^{k-u+r-1} \sum_{j=i+u-r}^{k-1} \binom{i-1}{r-1}\binom{j-i-1}{u-r-1}\binom{N-j-1}{n-u-1} f(x_j, x_i, c) +$$
$$+ \delta(k-u+1)\binom{N-k}{n-u} \sum_{i=r}^{k-u+r} \binom{i-1}{r-1}\binom{k-i-1}{u-r-1} f(x_j, x_i, c) +$$
$$+ \delta(k-r)\delta(N-n+u-k)\delta(u-r-1) \sum_{i=r}^{k-1} \sum_{j=k+1}^{N-n+u} \binom{i-1}{r-1}\binom{j-i-1}{u-r-1}\binom{N-j}{n-u} f(x_j, x_i, c) + \quad (7)$$
$$+ \delta(N-n+r-k+1)\binom{k-1}{r-1} \sum_{j=k+u-r}^{N-n+u} \binom{j-k-1}{u-r-1}\binom{N-j}{n-u} f(x_j, x_k, c) +$$
$$+ \delta(N-n+r-k)\delta(r-1) \sum_{i=k+1}^{N-n+r} \sum_{j=i+u-r}^{N-n+u} \binom{i-2}{r-2}\binom{j-i-1}{u-r-1}\binom{N-j}{n-u} f(x_j, x_i, c) \Bigg)$$

If $k > N-n+u$,

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

The inclusion probabilities of the second order are defined by

$$\pi_{k,t}(r,u,c) = \pi_{t,k}(r,u,) = P_{r,u}(s : k \in s, t \in s) = \sum_{\{s:k \in s, t \in s\}} P_{r,u}(s \mid c)$$

where $k < t = 1, ..., N$.

Theorem 1.2. The inclusion probabilities of the second degree of the sampling design $P_{r,u}(s \mid c)$ are as follows.

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

where

$$P(k,t\in s_3)=\frac{\delta(n-u-1)}{z(r,u,c)}\big(\delta(k-N+n-u)\cdot$$
$$\cdot\sum_{i=r}^{N-n+r}\sum_{j=i+u-r}^{N-n+u}\binom{i-1}{r-1}\binom{j-i-1}{u-r-1}\binom{N-j-2}{n-u-2}f(x_j,x_i,c)+\delta(N-n+u-k+1\_\delta(k-u)\cdot$$
$$\cdot\sum_{i=r}^{N-u+r}\sum_{j=k-1}^{N-n+u}\binom{i-1}{r-1}\binom{j-i-1}{u-r-1}\binom{N-j-2}{n-u-2}f(x_j,x_i,c)\Big), \tag{10}$$

$$P(X_{(u)}=x_k,t\in s_3)=\frac{\delta(N-k)\delta(n-u)\delta(N-n+u-k+1)}{z(r,u,c)}\cdot$$

$$\cdot\binom{N-k-1}{n-u-1}\sum_{i=r}^{k-n+r}\binom{i-1}{r-1}\binom{k-i-1}{u-r-1}f(x_j,x_i,c), \tag{11}$$

$$P(k\in s_2,t\in s_3)=\frac{\delta(u-r-1)\delta(n-u)\delta(t-k-u+r+1)}{z(r,u,c)}\cdot$$
$$\cdot\Big(\delta(N-n+u-t+1\sum_{i=r}^{k-1}\sum_{j=i+u-r}^{t-1}\binom{i-1}{r-1}\binom{j-i-2}{u-r-2}\binom{N-j-1}{n-u-1}f(x_j,x_i,c)+ \tag{12}$$
$$+\delta(t-N+n-u)\sum_{i=r}^{k-1}\sum_{j=i+u-r}^{N-n+u}\binom{i-1}{r-1}\binom{j-i-2}{u-r-2}\binom{N-j-1}{n-u-1}f(x_j,x_i,c)\Big),$$

$$P(X_{(r)}\in x_k,t\in s_3)=\frac{\delta(n-u)\delta(N-k)\delta(k-r+1)}{z(r,u,c)}\cdot$$
$$\cdot\Big(\delta(t-k-u+r)\delta(N-n+u-t+1)\binom{k-1}{r-1}\sum_{j=k+u-r}^{t-1}\binom{j-k-1}{u-r-1}\binom{N-j-1}{n-u-1}f(x_j,x_i,c)+ \tag{13}$$
$$+\delta(N-n-k+r+1)\delta(t-N+n-u)\binom{k-1}{r-1}\sum_{j=k+u-r}^{N-n+u}\binom{j-k-1}{u-r-1}\binom{N-j-1}{n-u-1}f(x_j,x_k,c)\Big),$$

$$P(X\in s_1,t\in s_3)=\frac{\delta(r-1)\delta(n-u)}{z(r,u,c)}\big(\delta(r-k)\delta(t-N+n-u)\cdot$$
$$\cdot\sum_{i=r}^{N-n+r}\sum_{j=i+u-r}^{N-n+u}\binom{i-2}{r-2}\binom{j-i-1}{u-r-1}\binom{N-j-1}{n-u-1}f(x_j,x_i,c)+\delta(r-k)\delta(N-n+u-t+1)\delta(t-u)\cdot$$
$$\cdot\sum_{i=r}^{t-u+r-1}\sum_{j=i+u-r}^{t-1}\binom{i-2}{r-2}\binom{j-i-1}{u-r-1}\binom{N-j-1}{n-u-1}f(x_j,x_i,c)+ \tag{14}$$
$$+\delta(k-r+1)\delta(t-N+n-u)\delta(N-n+r-k)\delta(t-u)\cdot$$
$$\cdot\sum_{i=k+1}^{t-u+r-1}\sum_{j=i+u-r}^{t-1}\binom{i-2}{r-2}\binom{j-i-1}{u-r-1}\binom{N-j-1}{n-u-1}f(x_j,x_i,c)\Big),$$

$$P(X \in s_2, X_{(u)} = x_t) = \frac{\delta(t-u)\delta(u-r)}{z(r,u,c)} \binom{N-t}{n-u} \sum_{i=r}^{t-u+r-1} \binom{i-1}{r-1} \binom{t-i-1}{u-r-1} f(x_j, x_i, c) \cdot$$
$$\cdot \sum_{i=r}^{t-u+r-1} \sum_{j=i+u-r}^{t-1} \binom{i-2}{r-2} \binom{j-i-1}{u-r-1} \binom{N-j-1}{n-u-1} f(x_j, x_i, c) + \tag{15}$$
$$+ \delta(k-r+1)\delta(t-N+n-u)\delta(N-n+r-k)\delta(t-u) \cdot$$
$$\cdot \sum_{i=k+1}^{t-u+r-1} \sum_{j=i+u-r}^{t-1} \binom{i-2}{r-2} \binom{j-i-1}{u-r-1} \binom{N-j-1}{n-u-1} f(x_j, x_i, c) \Bigg),$$

$$P(X_{(r)} = x_k, X_{(u)} \in x_t) =$$
$$= \frac{\delta(k-r+1)\delta(N-n+u-t+1)}{z(r,u,c)} \binom{k-1}{r-1} \binom{t-k-1}{u-r-1} \binom{N-t}{n-u} f(x_j, x_i, c), \tag{16}$$

$$P(k \in s_1, X_{(u)} = x_t) = \frac{\delta(r-1)\delta(u-r)}{z(r,u,c)} \cdot$$
$$\cdot \Bigg(\delta(r-k)\delta(t-u+1) \binom{N-t}{n-u} \sum_{i=r}^{t-u+r} \binom{i-2}{r-2} \binom{t-i-1}{u-r-1} f(x_j, x_i) + \tag{17}$$
$$\cdot \delta(k-r+1)\delta(t-u+r-k) \binom{N-t}{n-u} \sum_{i=k+1}^{t-u+r} \binom{i-2}{r-2} \binom{t-i-1}{u-r-1} f(x_j, x_i) \Bigg),$$

$$P(k, t \in s_2) = \frac{\delta(u-r-t+k-1)\delta(u-r-2)}{z(r,u,c)} \cdot$$
$$\cdot \sum_{i=r}^{k-1} \sum_{j=t+1}^{N-n+u} \binom{i-1}{r-1} \binom{j-i-3}{u-r-3} \binom{N-j}{n-u} f(x_j, x_i, c), \tag{18}$$

$$P(X_{(r)} \in x_k, t \in s_2) = \frac{\delta(N-n-k+r+1)\delta(u-r-1)}{z(r,u,c)} \cdot$$
$$\cdot \binom{k-1}{r-1} \sum_{j=k+u-r}^{N-n+u} \binom{j-k-2}{u-r-2} \binom{N-j}{n-u} f(x_j, x_i, c), \tag{19}$$

$$P(k \in s_1, t \in s_2) = \frac{\delta(r-1)\delta(u-r-1)\delta N-n+u-t)}{z(r,u,c)} \cdot$$

$$\cdot \left(\delta(r-k) \sum_{i=r}^{t-1} \sum_{j=t+1}^{N-n+u} \binom{i-2}{r-2} \binom{j-i-2}{u-r-2} \binom{N-j}{n-u} f(x_j, x_i, c) + \right. \tag{20}$$

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

$$P(k \in s_1, X_{(r)} = x_t) = \frac{\delta(r-1)\delta(k-1)\delta(N-n+u-t)}{z(r,u,c)} \cdot$$

$$\cdot \binom{k-2}{r-2} \sum_{j=t+1}^{N-n+u} \binom{j-k-1}{u-r-1} \binom{N-j}{n-u} f(x_j, x_k, c), \tag{21}$$

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

$$\left. + \delta(t-r-1) \sum_{i=t+1}^{N-n+r} \sum_{j=i+u-r}^{N-n+u} \binom{i-3}{r-3} \binom{j-i-1}{u-r-1} \binom{N-j}{n-u} f(x_j x_i, c). \right. \tag{22}$$

2. Sampling scheme

The sampling scheme implementing the sampling design $P_{r,u}(s \mid c)$ P is as follows. Firstly, population elements are ordered according to increasing values of the auxiliary variable. Let $s = s_1 \cup \{i\} \cup s_3 \cup \{j\} \cup s_3$, $s_1 = \{k : k \in U, x_k < x_i\}$, $s_2 = \{k : k \in U, x_j > x_k > x_i\}$ and $s_3 = \{k : k \in U, x_k > x_j\}$. Moreover, let $U = U(1, i-1) \cup \{i\} \cup U(i+1, j-1) \cup \{j\} \cup U(j+1, N)$ where $U(1, i-1) =$ $= (1, \ldots, i-1)$, $U(i+1, j-1) = (i+1, \ldots, j-1)$, $U(j+1, N) = (j+1, \ldots, N)$. Let $S(U(1, i-1); s)$ be sample space of the sample s_1 of size $r-1, S(U(i+1, j-1); s)$ be sample space of the sample s_2 of size $u-r-1, S(U(j+1, N); s)$ be sample space of the sample s_2 of size $n-u$. Similarly, $S = S(U, s)$.

The sampling scheme is given by the following of probabilities:

$$P_{r,u}(s \mid c) = P_1(s_1) p_{r,u}(i \mid c) P_2(s_2) p'_{r,u}(j \mid c) P_3(s_3) \tag{23}$$

where

$$P_1(s_1)=\binom{i-1}{r-1}^{-1},P_2(s_2)=\binom{j-i-1}{u-r-1}^{-1},P_3(s_3)=\binom{N-j}{n-u}^{-1} \tag{24}$$

$$P_{r,u}(i\mid c)=P(X_{(r)}=x_i\mid X_{(u)}=x_j,c)=\frac{P_{r,u}(X_{(r)}=x_i,X_{(u)}=x_j,c)}{P_{r,u}(X_{(u)}=x_j,c)} \tag{25}$$

$$p'_{r,u}(j\mid c)=P_{r,u}(X_{(u)}=x_j,c)=\frac{1}{z(r,u,)}\sum_{i=r}^{N-n+r}f(x_jx_i,c)g(r,u,i,j) \tag{26}$$

$$P_{r,u}(X_{(r)}=x_i\mid X_{(u)}=x_j,c)=\sum_{s\in G(r,u,i,j)}P_{r,u}(s\mid c)\frac{f(x_jx_i,c)g(r,u,i,j)}{z(r,u,c)} \tag{27}$$

In order to select the sample s, firstly the j-th element of the population should be selected, according to the probability function $p'_{r,u}=(j\mid c)$. Next, the i-the element of the population should be drawn according to the probability function $p'_{r,u}=(i\mid c)$. 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.

3. Some sampling strategies

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

$$\bar{y}_{HT,s}=\frac{1}{N}\sum_{k\in s}\frac{y_k}{\pi_k} \tag{28}$$

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 (32) and (34), respectively.

The particular case of the above estimator is the well known sampling design of the simple sample drawn without replacement is as follows: $P_0(s)=\binom{N}{n}^{-1}$. The variance of the mean from the simple 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 $v_y = \frac{1}{N-1}\sum_{k=1}^{N}(y_k - \bar{y})^2$.

The results of the previous chapter lead to construction of the regression sampling strategy for the population mean $\bar{y}_s = \frac{1}{n}\sum_{ks} y_i$. We assume that $y_i = a + bx_i + e_i$ 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. The linear correlation coefficient between the variables y and x will be denoted by ρ . Let $(X_{(r)}, Y_r)$ be two dimensional random variable where $X_{(r)}$ is the *r*-th order statistic of an auxiliary variable and Y_r is the variable under study. Wywiał (2009) considered the following estimator:

$$\bar{y}_{r,u,s} = \bar{y}_{HT,s} + b_{r,u,s}(\bar{x} - \bar{x}_{HT,s}) \tag{29}$$

where

$$b_{r,u,s} = \frac{Y_u - Y_r}{X_{(u)} - X_{(r)}} \tag{30}$$

Wywiał (2009) showed that under the sampling design stated in the definition 1.1 the parameters of the following strategies $\bar{y}_{r,u,s}, P_{r,u}(s \mid c))$, are approximately as follows.

$$E\left(\bar{y}_{r,u,s}, P_{r,u}(s \mid c)\right) \approx \bar{y}$$

$$D^2(\bar{y}_{r,u,s}, P_{r,u}(s \mid c)) \approx D^2(\bar{y}_{HT,s}, P_{r,u}(s \mid c)) - 2bCov(\bar{y}_{HT,s}, \bar{x}_{HT,s}, P_{r,u}(s \mid c)) + \\ + b^2 D^2(\bar{x}_{HT,s}, P_{r,u}(s \mid c)) \tag{31}$$

where

$$Cov(\bar{y}_{HT,s}, \bar{x}_{HT,s}, P_{r,u}(s \mid c)) = \frac{1}{N^2}\left(\sum_{k\in U}\sum_{l\in U} \Delta_{k,l} \frac{y_k}{\pi_k}\frac{x_l}{\pi_l}\right) \tag{32}$$

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

The approximately unbiased estimator of the variance: $D^2 = \bar{y}_{r,u,s}, P_{r,u}(s \mid c))$ is as follows

$$\hat{D}^2(\bar{y}_{r,u,s}, P_{r,u}(s \mid c)) = \hat{D}^2(\bar{y}_{HT,s}, P_{r,u}(s \mid c)) - 2b_{r,u,s}\hat{Cov}(\bar{y}_{HT,s}, \bar{x}_{HT,s}P_{r,u}(s \mid c)) + \\ + b_{r,u,s}^2 \hat{D}^2(\bar{x}_{HT,s}, P_{r,u}(s \mid c)) \tag{33}$$

where

$$\hat{Cov}(\bar{y}_{HT,s}, \bar{x}_{HT,s}, P_{r,u}(s \mid c)) = \frac{1}{N^2}\left(\sum_{k\in s}\sum_{l\in s}\Delta_{*,k,l}\frac{y_k}{\pi_k}\frac{x_l}{\pi_l}\right) \tag{34}$$

$$\Delta_{*k,l} = \frac{\Delta_{k,l}}{\pi_{k,l}},\ \hat{D}^2(\bar{x}_{HT,s}, P_{r,u}(s \mid c)) = \hat{Cov}(\bar{x}_{HT,s}, \bar{x}_{HT,s}, P_{r,u}(s \mid c)),$$

$$\hat{D}^2(\bar{y}_{r,u,s}, P_{r,u}(s \mid c)) = \hat{Cov}(\bar{y}_{HT,s}, \bar{x}_{HT,s}P_{r,u}(s \mid c))$$

The next regression type estimator is given by:

$$\tilde{y} = \bar{y}_{HT,s} + \bar{b}_s(\bar{x} - \bar{x}_{HT,s}) \tag{35}$$

$$\tilde{b}_s = \frac{\tilde{v}_{x,y,s}}{\tilde{v}_{x,s}} \tag{36}$$

where

$$\tilde{v}_{x,y,s} = \frac{1}{\tilde{N}-1}\sum\frac{(x_k - \tilde{x}_{HT,s})(y_k - \tilde{y}_{HT,s})}{\pi_k},\ \tilde{v}_{x,s} = \tilde{v}_{x,x,s},$$

$\tilde{N} = \sum_{k\in s}\frac{1}{\pi_k}$, $\tilde{x}_{HT,s} = N\bar{x}_{HT,s}/\tilde{N}$, $\tilde{y}_{HT,s} = N\bar{y}_{HT,s}/\tilde{N}$. The statistics $\tilde{v}_{x,y,s}$ and $\tilde{v}_{x,s}$ are the consistent estimators of the population covariance and variance, $v_{xy} = \frac{1}{N-1}\sum_{k\in U}(x_k - \bar{x})(y_k - \bar{y})$, respectively (see e.g. Särndal et al., 1997, p. 187).

The strategy $(\tilde{y}_s, P_{r,u}(s \mid c))$ has approximately the same parameters as the above considered strategy $(\tilde{y}_{r,u,s}, P_{r,u}(s \mid c))$.

Let us remember that the ordinary regression estimator is given by:

$$\hat{y}_s = \bar{y}_s + b_s(\bar{x} - \bar{x}_s) \tag{37}$$

where

$$b_s = \frac{\sum_{k\in s}(x_k - \bar{x}_s)(y_k - \bar{y}_s)}{\sum_{k\in s}(x_k - \bar{x}_s)^2} \tag{38}$$

The approximate value of the variance is as follows.

$$D^2(\hat{y}_s, P_0(s)) = \frac{N-n}{Nn} v_y (1-\rho^2) \tag{39}$$

where $\rho = \frac{v_{x,y}}{\sqrt{v_x v_y}}$.

The approximately unbiased estimator of the variance is as follows.

$$D^2(\hat{y}_s, P_0(s)) = \frac{N-n}{Nn} v_{y,s} (1-r_s^2) \tag{40}$$

where $r_s = \frac{v_{x,y,s}}{\sqrt{v_{x,s} v_{y,s}}}$, $v_{xy,s} = \frac{1}{n-1}\sum_{k\in s}(x_k - \bar{x}_s)(y_k - \bar{y}_s)$, $v_{x,s} = v_{xx,s}$, $v_{y,s=v_{yy,s}}$

The strategy $(\hat{y}_s, P_0(s))$ is asymptotically unbiased for the population mean $\bar{y}$. It is well known that the strategy $(\hat{y}_s, P_{S,S}(s))$ is unbiased for $\bar{y}$ where

$$P_{S,S}(s) = \frac{v_{x,s}}{\binom{N}{n} v_x} \tag{41}$$

is the sampling design of Singh and Srivastava (1980). The strategies $\hat{y}_s, P_{SS}(s)$ and $(t_{HT,s}, P_{SS}(s))$ are unbiased for the population mean $\bar{y}$. The variances of the strategy $(\hat{y}_s, P_{S,S}(s))$ is approximately equal to the right side of the equation (39).

4. Simulation analysis of strategies accuracy

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 we define as follows:

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

where $\left(\bar{y}_s, P_0(s)\right)$ is ordinary simple sample mean. Let $e1, e2, e3, e4, e5$ and $e6$ be the relative efficiency coefficients of the strategies $\left(\hat{y}_s, P_0(s)\right)$, $\left(\hat{y}_s, P_{SS}(s)\right)$, $\left(\bar{y}_{HT,s}, P_{SS}(s)\right)$, $\left(\bar{y}_{r,u,s}, P_{r,u}(s \mid c)\right)$, $\left(\bar{y}_s, P_{r,u}(s \mid c)\right)$ and $\left(\bar{y}_{HT,s}, P_{r,u}(s \mid c)\right)$, respectively.

The population consists of the municipalities in Sweden. The auxiliary variable x is: 1975 municipal population (in thousands) and the variable under study y is: 1985 municipal taxation revenues (in millions of kronor). Their observations have been published by Srndal, Swenson and Wretman (1992). The size of this population is 284 municipalities. There are three outlier observations of the variables, see Figure 1. Let d and β_3 be the standard deviation and the skewnees coefficient, respectively, of the auxiliary variable in the population. In the case of data without outliers (size of the population N = 281) $\bar{x}$ = 24, 263, d = 24,153 and β_3 = 0, 043. In the case of data with outliers (size of the population N = 284) $\bar{x}$ = 28,810, d = 52, 873 and β_3 = 8, 427. The samples were replicated 1000 times.

In general, Tables 1, 2, 3 let us infer that the regression type strategies are the best among the considered ones.

Table 1

The relative efficiency coefficients (%) of the strategies

N :	281			284		
n	e1	e2	e3	e1	e2	e3
2 (0,7%)	80,5	6,9	30,6	5,0	5,8	10,0
3 (1%)	6,6	2,8	22,7	1,7	5,3	9,3
4 (1,5%)	4,7	2,6	24,8	1,8	4,0	8,6
6 (2%)	4,2	2,7	28,9	1,7	3,7	9,6
9 (3%)	4,1	2,7	35,0	1,9	3,3	12,2
11 (4%)	3,4	2,6	42,6	2,4	3,3	13,1
14 (5%)	3,8	2,6	45,0	2,6	3,3	14,6
29 (10%)	3,4	2,6	61,0	4,1	4,4	21,3

On the basis of the Table 1 we conclude that the regression strategy $(\bar{y}_{HT,s}, P_{SS}(s))$ is worse than the strategies $(\hat{y}_{HT,s}, P_0(s))$, and $(\hat{y}_S, P_{SS}(s))$. In the case of the population without outliers the Singh-Srivastava regression strategy is better than ordinary regression strategy. But in the case of the population with outliers there is an opposite situation because $(\hat{y}_S, P_0(s))$, is better than $(\hat{y}_s, P_{SS}(s))$.

The Horvitz-Thompson strategy for Singh and Srivastava's strategy $(\bar{y}_{HT,S}, P_{SS}(s))$ is better than the unconditional strategy $(\bar{y}_{HT,S}, P_{1,n}(s))$ when the outliers do not exists. But in the case when the population is with the outliers the strategy $(\bar{y}_{HT,S}, P_{1,n}(s))$ is better than $(\bar{y}_{HT,S}, P_{SS}(s))$ only for small sample sizes $n \leq 4$.

The simulation analysis of the accuracy of the conditional strategies $(\bar{y}_{r,u,s}, P_{r,u}(s \mid c))$, $(\tilde{y}_s, P_{r,u}(s \mid c))$ and $(\bar{y}_{HT,s}, P_{r,u}(s \mid c))$ leads to the conclusion that they are the best for the sampling design proportional to the difference of the last and the first order statistics. That is why the Tables 2 and 3 deals only with strategies dependent on the conditional sampling design $P_{1,n}(s \mid c)$.

Table 2

The relative efficiency coefficients (%) of the conditional strategies for $P_{1,n}(s \mid c)$, $c = kd$, $k = 0, 1, 2, 3$ (The population with outliers, N = 284)

c = kd	0			d			2d			3d		
n	e4	e5	e6	e4	e5	e6	e4	e5	e6	e4	e5	e6
2	3,3	1,1	9,7	3,5	1,7	11,4	8,2	2,2	23,2	11,6	2,1	30,7
3	1,4	1,4	5,1	1,6	1,4	6,1	4,9	1,8	13,5	5,5	1,9	16,6
4	1,6	1,5	7,2	1,8	1,6	6,1	3,1	1,8	10,4	5,2	2,0	15,7
6	1,9	1,9	9,8	1,7	1,6	6,7	3,1	1,8	9,7	4,5	2,1	14,6
9	2,3	2,1	14,0	2,1	1,9	10,0	3,1	2,1	10,2	4,2	2,2	13,0
11	2,6	2,4	16,1	2,6	2,4	11,9	3,1	2,1	10,1	4,1	2,0	12,7
14	3,4	2,9	20,1	2,9	2,7	14,8	3,6	2,6	10,5	4,2	2,3	12,4
29	4,5	3,9	27,1	4,7	3,8	26,8	5,0	3,8	17,0	4,8	3,0	15,7

In the both case of the population with outliers and without them the Horvitz-Thompson type strategy $(\bar{y}_{HT,s}, P_{1,n}(s \mid c))$ is less accurate than $(\bar{y}_{1,n,s}, P_{1,n}(s \mid c))$ and $(\bar{y}_s, P_{1,n}(s \mid c))$.

The regression strategy $(\tilde{y}_s, P_{1,n}(s \mid c))$ is better than $(\bar{y}_{1,n,s}, P_{1,n}(s))$ for all considered sample sizes and values of the parameter c (see Tables 2 and 3 or Figures 2 and 3).

It is possible to observe that for $n \geq 6$ and some conditional strategies are more accurate then the unconditional appropriate ones. For instance, the strategy $(\bar{y}_{HT,s}, P_{1,14}(s \mid 3d))$ is better than the unconditional strategy $(\bar{y}_{HT,s}, P_{1,14}(s)) = (\bar{y}_{HT,s}, P_{1,14}(s \mid 0))$. This situation explain Figures 4 and Figure 5, where the distribution spread of the unconditional strategy is gray and the conditional one is black. The distribution of the unconditional strategy $(\bar{y}_{HT,s}, P_{1,14}(s))$ has

some large outliers and that is why the variance of this strategy is larger than the variance of the strategy $(\bar{y}_{HT,s}, P_{1,14}(s \mid 3d))$. Moreover, the conditional strategy neglect some small values of the unconditional one.

Table 3

The relative efficiency coefficients (%) of the conditional strategies for $P_{1,n}(s \mid kd)$, $k = 0, 1, 2, 3$ (The population without outliers, N = 281)

c = kd	0			d			2d			3d		
n	e4	e5	e6	e4	e5	e6	e4	e5	e6	e4	e5	e6
2	3,3	1,1	9,7	3,5	1,7	11,4	8,2	2,2	23,2	11,6	2,1	30,7
3	1,4	1,4	5,1	1,6	1,4	6,1	4,9	1,8	13,5	5,5	1,9	16,6
4	1,6	1,5	7,2	1,8	1,6	6,1	3,1	1,8	10,4	5,2	2,0	15,7
6	1,9	1,9	9,8	1,7	1,6	6,7	3,1	1,8	9,7	4,5	2,1	14,6
9	2,3	2,1	14,0	2,1	1,9	10,0	3,1	2,1	10,2	4,2	2,2	13,0
11	2,6	2,4	16,1	2,6	2,4	11,9	3,1	2,1	10,1	4,1	2,0	12,7
14	3,4	2,9	20,6	2,9	2,7	14,8	3,6	2,6	10,5	4,2	2,3	12,4
29	4,5	3,9	27,1	4,7	3,8	26,8	5,0	3,8	17,0	4,8	3,0	15,7

Conclusions

The inclusion probabilities of the conditional sampling design proportionate to the difference of two order statistics are presented. They let determine the variance of the Horvitz-Thompson statistic as well as its estimator. The construction of the sampling design can be easy modified in such a way that definition of the sampling design proportionate to e.g. the sum of two order statistics is straightforward.

The simulation analysis let us infer that in general the considered regression type strategies are the best among the considered ones. Moreover, the relative efficiencies of the regression ones are similar and the best among them is the strategy $(\bar{y}_s, P_{1,n}(s \mid 3d))$. In some cases the conditional strategies could be slightly better than the appropriate unconditional ones but this conclusion will be developed in separate and more deep studies.

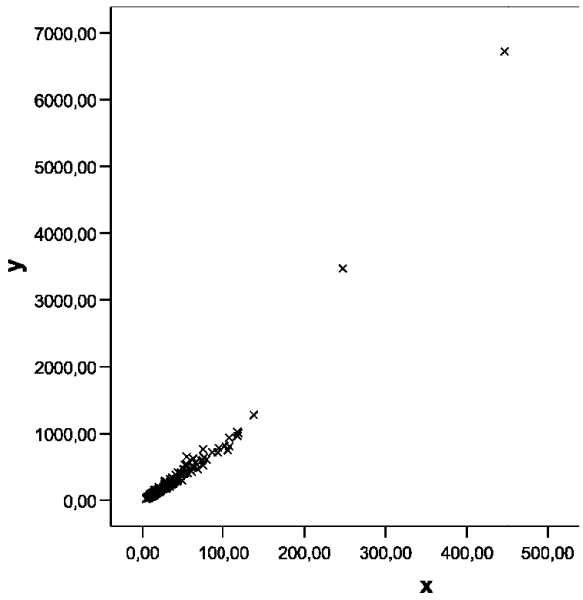

Figure 1. Spread of variables x and y

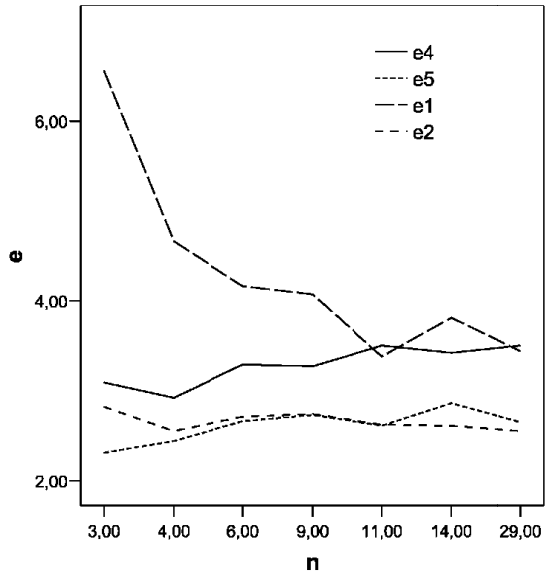

Figure 2. Efficiencies of the strategies in the case of the population without outliers

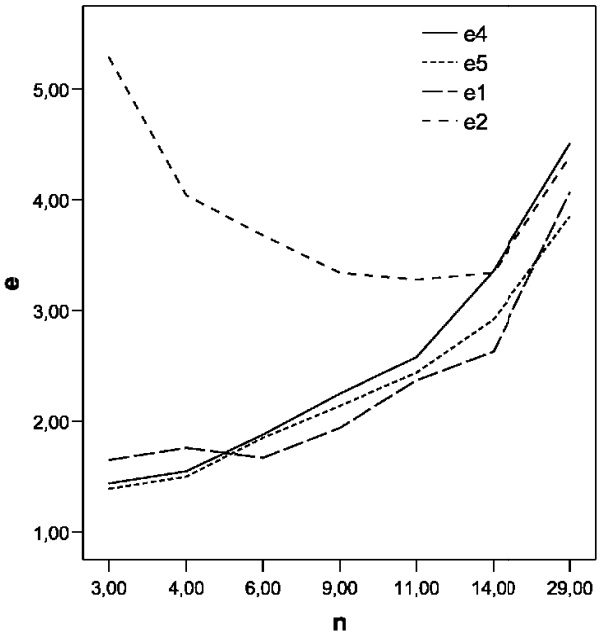

Figure 3. Efficiencies of the strategies in the case of the population with outliers

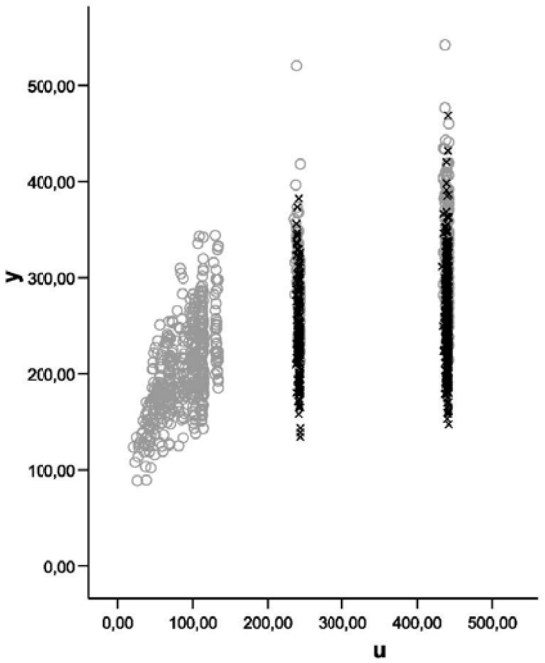

Figure 4. Distribution spreads of the strategy $u = (\bar{y}_{HT,s}, P_{r,u}(s \mid d))$ on $x = x_{(n)} - x_{(n)}$ for $d = 0$ or $d = 3$ and $N = 284$, $n = 14$

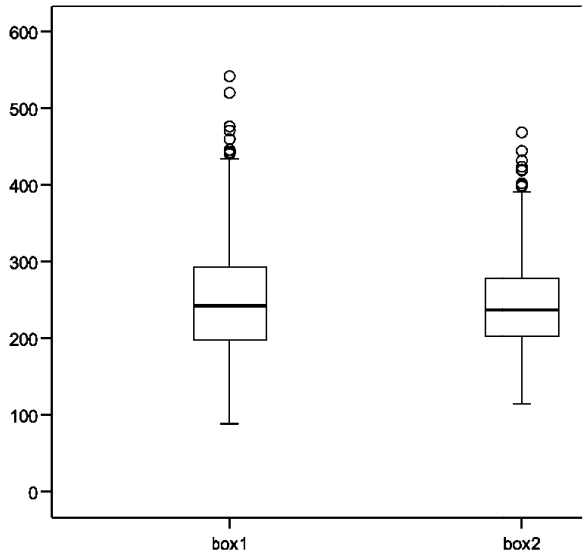

Figure 5. Distribution boxplots of the strategy $u = (\bar{y}_{HT,s}, P_{r,u}(s \mid d))$ for $d = 0$ or $d = 3$ and $N = 284$, $n = 14$

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", 47, s. 663-685.

Särndal C.E., Swensson B., Wretman J. (1992): *Model Assisted Survey Sampling*. Springer Verlag, New York – Berlin – Heidelberg.

Singh P., Srivastava A.K. (1980): *Sampling Schemes Providing Unbiased Regression Estimators*. „Biometrika" 67 (1), s. 205-9.

Tillé Y. (1999): E*stimation in Surveys Using Conditional Inclusion Probabilities: Complex Design*. „Survey Methodology", Vol. 25, No, 1, s. 57-66.

Tillé Y. (2006): *Sampling algorithms*. Springer, New York.

Wywiał J.L. (2003): *On Conditional Sampling Strategies*. „Statistical Papers", Vol. 44, 3, s. 397-419.

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, s. 125-137.

Wywiał J.L. (2008): *Sampling Design Proportional to Order Statistic of Auxiliary Variable*. „Statistical Papers", Vol. 49, No. 2, s. 277-289.

Wywiał J.L. (2009): *Performing Quantiles in Regression Sampling Strategy*. „Model Assisted Statistics and Applications", Vol. 4, No. 2, s. 131-142.

SAMPLING DESIGNS PROPORTIONATE TO NON-NEGATIVE FUNCTIONS OF TWO QUANTILES OF AUXILIARY VARIABLE

Summary

Estimation of the population average in a finite population by means of sampling strategies dependent on sample quantiles of an auxiliary variable are considered. The sampling design proportional to an order statistic of an auxiliary variable was defined by Wywiał (2007, 2008). It was generalized into case of the sampling design proportional to the difference of two order statistics by Wywiał (2009), too. In this paper those results are generalized on the case of a conditional sampling design. Several strategies including the Horvitz-Thompson statistic and regression estimators are considered. Their accuracy is analyzed on the basis of computer simulation experiments.