STATISTICS IN TRANSITION-new series, August 2011
Vol. 12, No. 1, pp. 127—138

ESTIMATION OF DOMAIN MEANS ON THE BASIS OF STRATEGY DEPENDENT ON DEPTH FUNCTION OF AUXILIARY VARIABLES' DISTRIBUTION

Janusz L. Wywiał[1]

ABSTRACT

The paper deals with the problem of estimation of a domain means in a finite and fixed population. We assume that observations of a multidimensional auxiliary variable are known in the population. The proposed estimation strategy consists of the well known Horvitz-Thompson estimator and the non-simple sampling design dependent on a synthetic auxiliary variable whose observations are equal to the values of a depth function of the auxiliary variable distribution. The well known spherical and Mahalanobis depth functions are considered. A sampling design is proportionate to the maximal order statistic determined on the basis of the synthetic auxiliary variable observations in a simple sample drawn without replacement. A computer simulation analysis leads to the conclusion that the proposed estimation strategy is more accurate for domain means than the well known simple sample means.

Key words: sampling design, order statistic, auxiliary variable, sampling scheme, Horvitz-Thompson estimator, small area estimation, area sampling, depth function

1. Sampling strategy

Let $U=(1,2,...,N)$ be a fixed population of the size N. An observation of a variable under study (an auxiliary variable) attached to the i-th population element will be denoted by y_i (z_i), $i=1,...,N$. The sample of the size n, drawn without replacement from the population, will be denoted by s. The sampling design is denoted by $P(s)$ and inclusion probabilities of the first and second orders – by π_k, for k=1,....,N and $\pi_{k,t}$ for k≠t, k=1,...N, t=1,...N, respectively. Let $\boldsymbol{S}$ be the sample space of the samples of size n, drawn without replacement. We are

[1] Department of Statistics, Katowice University of Economics, Poland.

going to consider the following sampling designs of simple samples drawn without replacement: $P_0(s) = \binom{N}{n}^{-1}$ for all $s \in \boldsymbol{S}$.

There are several sampling designs on the value of a non-negative auxiliary variable. Some reviews of them are presented, e.g. by Tillé (2006). We are going to consider the following one analysed by Wywiał (2007, 2008, 2009).

Let $Z_{(r)}$ be the r-th order statistic from a simple sample drawn without replacement where the auxiliary variable's values are observed. Let $\alpha \in (0;1)$ and $[n\alpha]$ be the integer part of the value $n\alpha$. The sample quantile of the order α is defined as $Q_{s,\alpha} = Z_{(r)}$ where $r = [n\alpha]+1$ and $(r-1)/n \leq \alpha < r/n$. Let $G(i,r)=\{s: Z_{(r)}=x_i\}$ be the set of all samples s whose r-th order statistic of the auxiliary variable is equal to x_i where $r \leq i \leq N-n+r$. Let $U_1=(1,...,i-1)$ be a subpopulation of the population U and let s_1 be the simple sample of the size $(r-1)$, drawn without replacement from U_1. Similarly, let $U_2=(i+1,...,N)$ be a subpopulation of the population U and let s_2 be a simple sample of the size $(n-r)$, drawn without replacement from U_2. The sample space of the samples of the type s_1 will be denoted by $\mathbf{S}(U_1,s_1)$ and the sample space of the samples of the type s_2 will be denoted by $\mathbf{S}(U_2,s_2)$. Let $s = (s_1 \cup \{i\} \cup s_2)$ be such a sample that the value of the r-th order statistic of an auxiliary variable - observed in the sample - equals x_i. Hence, the sample space of such a sample s is as follows: $\mathbf{S}_{r,i} = \mathbf{S}(U_1,s_1) \times \{i\} \times \mathbf{S}(U_2,s_2)$, where $\times$ is the symbol of the Cartesian product. So, the sample space for all $i=r,...,N-n+r$ is as follows $\mathbf{S} = \bigcup_{i=r}^{N-n+r} \mathbf{S}_{r,i}$.

Wywiał (2008) proposed the following sampling design proportional to the z_i value of the $Z_{(r)}$ statistic.

$$P_1(s \mid r) = \frac{Z_{(r)}}{\sum_{j=r}^{N-n+r} \binom{j-1}{r-1}\binom{N-j}{n-r} z_j} \qquad \text{for } s \in \boldsymbol{S}, \qquad (1)$$

or

$$P_1(s \mid r) = \frac{z_i}{\sum_{j=r}^{N-n+r} \binom{j-1}{r-1}\binom{N-j}{n-r} z_j} \qquad \text{for } s \in \boldsymbol{S}_{r,i},\ i=r,...,N-n+r.$$

The conditional version of the sampling design is as follows.

$$P_1(s \mid z_u \le Z_{(r)} \le z_v) = P_1(s \mid r; u, v) = \frac{Z_{(r)}}{\sum_{j=u}^{v} \binom{j-1}{r-1}\binom{N-j}{n-r} z_j} \qquad \text{for } s \in \boldsymbol{S}_{r/u,v},$$

where $\mathbf{S}_{\mathbf{r}/u,v} = \bigcup_{i=u}^{v} \mathbf{S}_{r,i}$. Equivalently,

$$P_1(s \mid r; u, v) = \frac{z_i}{\sum_{j=u}^{v} \binom{j-1}{r-1}\binom{N-j}{n-r} z_j} \qquad \text{for } s \in \boldsymbol{S}_{r,i},\ i=r,\dots,N\text{-}n+r. \tag{2}$$

for $s \in G(i,r)$ and $r \le u \le i \le v \le N\text{-}n+r$.

Let us define such a function $\delta(t)$ that if $t \le 0$ ($t>0$) then $\delta(t)=0$ ($\delta(t)=1$). Moreover, let

$$a_r(u,v) = \sum_{j=u}^{v} z_j \binom{j-1}{r-1}\binom{N-j}{n-r} \tag{3}$$

The inclusion probabilities of the first order are as follows:

$$\pi_k = \frac{\delta(u-k)\delta(r-1)\delta(v-1)\delta(u-1)}{a_r(u,v)} \sum_{i=u}^{v} \binom{i-2}{r-2}\binom{N-i}{n-r} z_i +$$

$$+ \frac{\delta(k-u+1)\delta(v-k+1)}{a_r(u,v)} \left(\delta(n-r)\delta(k-u)\delta(k-1) \sum_{i=u}^{k-1} \binom{i-1}{r-1}\binom{N-i-1}{n-r-1} z_i + \right.$$

$$\left. + \binom{k-1}{r-1}\binom{N-k}{n-r} z_k + \delta(r-1)\delta(v-k) \sum_{i=k+1}^{v} \binom{i-2}{r-2}\binom{N-i}{n-r} z_i \right) +$$

$$\frac{\delta(k-v)\delta(n-r)\delta(N-v)}{a_r(u,v)} \sum_{i=u}^{v} \binom{i-1}{r-1}\binom{N-i-1}{n-r-1} z_i \tag{4}$$

for k=1,...,N. The probabilities of the second order have been derived by Wywiał (2008).

The sampling scheme implementing the $P_2(s|r;u,v)$ conditional sampling design is as follows. Population elements are ordered according to the increasing values of the auxiliary variable. Next, the *i*-th element of the population is drawn with this probability:

$$p_1(i\,|\,r;u,v)=\frac{\binom{i-1}{r-1}\binom{N-i}{n-r}z_i}{a_r(u,v)}, \qquad \text{i=u,...,v} \tag{5}$$

Next, the simple sample s_1 of the size *(r-1)* is drawn from the subpopulation U_1 and the simple sample s_2 of the size *(n-r)* – from the subpopulation U_2. Let us note that the expression (6) shows the truncated distribution of the order statistic $Z_{(r)}$ from the sample drawn according to the sampling design $P_1(s|r;u,v)$. Hence, $p_1(i\,|\,r;u,v)=P_1(Z_{(r)}=z_i\,|\,z_u\le Z_{(r)}\le z_v)$.

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

$$\bar{y}_{HTS}=\frac{1}{N}\sum_{k=1}^{N}\frac{I_k y_k}{\pi_k} \tag{6}$$

If $I_k = 1$, if the *k*-th population element was drawn into a sample. When $I_k= 0$, the *k*-th element was not drawn into the sample. It is well known that the strategy $(\bar{y}_{HTS}, P(s))$ is unbiased for the population mean when all inclusion probabilities are positive. The expression for the variance of the Horvitz-Thompson is well known and it can be found in all survey sampling textbooks. The strategy $(\bar{y}_{HTS}, P_0(s))=(\bar{y}_S, P_0(s))$ where $\bar{y}_S=\frac{1}{n}\sum_{i\in S}y_i$ is called a simple sample mean and its variance is: $D^2(\bar{y}_S, P_0(s))=\frac{N-n}{Nn}v$,

2. Synthetic auxiliary variable

Now, we are going to consider an auxiliary variable with possible negative values which will be denoted by *x*. Let us define the following transformation of the auxiliary variable.

$$z_i = \frac{1}{1+(x_i-\bar{x})^2} \quad , \text{i=1,…,N} \tag{7}$$

So, the obtained variable *z* can be treated as a new synthetic auxiliary variable. Let us note that the observations of a symmetric variable *x,* close to its mean value $\bar{x}$, are transformed into the large values of the variable *z*. Let the two-dimensional variable *(y,x)*, where *y* denotes the variable under study, be symmetric. So, when the variables are highly correlated, we can expect the sampling design (proportionate to values of the variable *z*) more frequently gives the sample with observations of the variable *y* close to its mean value.

Now, we are going to construct more general synthetic auxiliary variables. A multidimensional auxiliary variable will be denoted by *x* and its value attached to the i-th population element will be denoted by $\mathbf{x}_{i*}=[x_{i1} \ldots x_{iq}]$, *i=1,…,N.*

The matrix of all observations of the auxiliary variable is $\mathbf{x} = \begin{bmatrix} \mathbf{x}_{1*} \\ \ldots \\ \mathbf{x}_{N*} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_{*1} & \ldots & \mathbf{x}_{*q} \end{bmatrix}$ where $\mathbf{x}_{*j} = \begin{bmatrix} \mathbf{x}_{11} \\ \ldots \\ \mathbf{x}_{Nj} \end{bmatrix}$, *i=1,…,N.* Let $\bar{\mathbf{x}} = \begin{bmatrix} \bar{\mathbf{x}}_1 & \ldots & \bar{\mathbf{x}}_q \end{bmatrix} = \frac{1}{N}\mathbf{J}_N^T\mathbf{x}$

where $\boldsymbol{J}_N$ be a unit column-vector whose all *N*-elements are equal to one. Moreover, let $\mathbf{d} = \mathbf{x} - \mathbf{J}_N\bar{\mathbf{x}}$ and $\mathbf{c}(x) = \frac{1}{N-1}\mathbf{d}^T\mathbf{d}$ be the population variance-covariance matrix of the auxiliary variables. Similarly, $\mathbf{c}(x,y) = \frac{1}{N-1}\mathbf{d}^T(\mathbf{y} - \mathbf{J}_N\bar{y})$ where yT =[y1…yN]. The sample auxiliary variable observation matrix is denoted by $\boldsymbol{x}_s$ and it consists of such rows $\boldsymbol{x}_{i*}$,i∈s. So, the sample variance-covariance matrix is $\mathbf{c}_s(x) = \frac{1}{n-1}\mathbf{d}_s^T\mathbf{d}_s$, $\mathbf{c}_s(x,y) = \frac{1}{n-1}\mathbf{d}_s^{\,T}(\mathbf{y}_s - \mathbf{J}_N\bar{y}_s)$ and y_s is the column vector of values of the variable under study observed in the sample.

There are many definitions of depth functions, see e.g. Liu, Parelius and Singh (1999). We will consider the so called spherical depth function and the one based on the well known Mahalanobis (1936) distance. They are defined by the following equations respectively.

$$z(x_i) = \frac{1}{1+\mathbf{d}_i\mathbf{d}_i^{\,T}}, \quad z(x_i) = \frac{1}{1+\mathbf{d}_i\mathbf{c}^{-1}(x)\mathbf{d}_i^{\,T}}, \qquad \text{i=1,…,N} \tag{8}$$

So, the maximal depth of an observation of the auxiliary variable is equal to one. Let us note that the synthetic variable given by the expression (8) is a one-dimensional version of the spherical depth function.

Let us suppose that the distribution of the variables *(y,**x**)* where $\boldsymbol{x}=[x_1,x_2,\ldots,x_q]$ is symmetric. Moreover, let the multiple correlation coefficient between the variable under study **y** and the q-dimensional auxiliary one $\boldsymbol{x}$ be high. So, the observations of the auxiliary variable $\boldsymbol{x}$, close to the mean values $\bar{\mathbf{x}}$, are transformed into high values of the synthetic variable z. Hence, similarly to the one-dimensional auxiliary variable, when the variables are highly correlated, we can expect the sampling design (proportionate to values of the variable z) more frequently gives the sample with observations of the variable y close to its mean value. Let us note that the foregoing analysis shows that the symmetry of a multidimensional auxiliary variable distribution is not necessary.

3. Accuracy comparison of estimation strategies

The accuracy of the $(t_S, P(s))$ estimation strategy was measured by means of the relative efficiency (*deff*):

$$e = \frac{D^2(\bar{y}_{HTS}, P(s))}{D^2(\bar{y}_S, P_0(s))}$$

The parameter $D^2(\bar{y}_{HTS}, P(s))$ is assessed as the variance of the estimator's values calculated on the basis of the values of the variables observed in replicated sampling according to the sampling design *P(s)*.

Let *es* and *eM* denote the relative efficiency for the strategy based on the synthetic auxiliary variable evaluated on the basis of spherical depth function and the Mahanalobis one, respectively. Moreover, we assume that r=n, because for $r<n$ the relative efficiency coefficients are usually not greater than in the case when $n=r$, as it was shown by Wywiał (2007).

The simulation analysis was based on three–dimensional observations which were obtained on the basis of the generator of random values of three-dimensional normal distribution. The synthetic auxiliary variable was evaluated on the basis of the spherical or Mahanalobis depth functions defined by the expression (8). The distribution considered was the mixture of the three-dimensional normal random variables with the same weights equal to *1/3*. The normal distribution was dented by $N(m_1, m_2, m_3, V)$, where m_1 is the mean value of the variable under study, m_2, m_3 are mean values of auxiliary variables, and finally V is the variance-covariance matrix. The parameters of the distributions were $N(4,4,4,V_1)$, $N(4,8,8,V_3)$ and $N(4,12,4,V_3)$, where

$$V_1 = \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 0.6 & 0.5 \\ 0.5 & 0.5 & 0.6 \end{bmatrix}, \qquad V_2 = \begin{bmatrix} 1 & 0.95 & 0.9 \\ 0.95 & 1 & 0.8 \\ 0.9 & 0.8 & 1 \end{bmatrix}, \; V_3 = \begin{bmatrix} 1 & -0.9 & 0.97 \\ -0.9 & 1 & -0.8 \\ 0.97 & -0.8 & 1 \end{bmatrix}$$

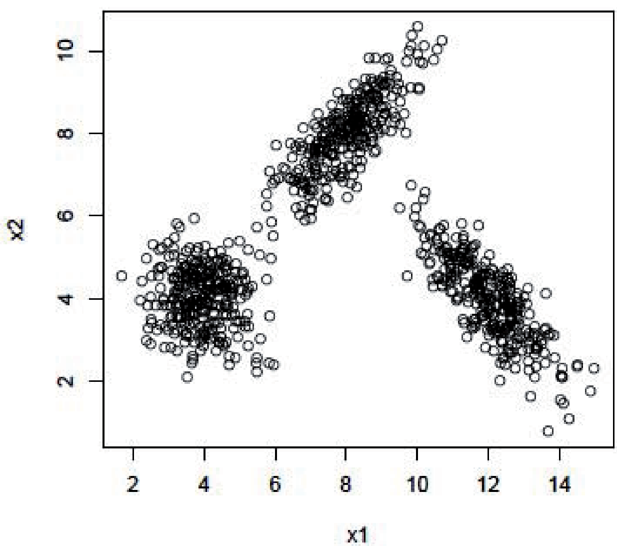

Figure 1. The spread function of the auxiliary variables.

The multiple correlation coefficients between variable under study and the auxiliary variables were calculated on the basis of the expression: $\rho_i = 1 - \det(V_i)/diag(V_i)$, i=1,2,3. Their values are ρ_1=0.913, ρ_2=0.978, ρ_3=0.992.

According to the probability distribution considered, 900 observations were generated by means of a standard program available in the package R.

The two-dimensional distribution of the auxiliary variable is the marginal distribution of the three-dimensional distribution has been just defined. It is the mixture of the following normal distributions: N(4,4,V_{x1}), N(8,8,V_3) and N(12,4,V_{x3}), where

$$V_{x1} = \begin{bmatrix} 0{,}6 & 0{,}5 \\ 0{,}5 & 0{,}6 \end{bmatrix}, \quad V_{x2} = \begin{bmatrix} 1 & 0{,}9 \\ 0.9 & 1 \end{bmatrix}, \quad V_3 = \begin{bmatrix} 1 & 0.97 \\ 0.97 & 1 \end{bmatrix}.$$

Figure 1 shows the spread of the auxiliary variable’s observations.

Table 1. The relative efficiency coefficients. The spherical depth function. The size population: N=900. The sampling scheme for r=n=u.

n	e_0	e_1	e_2	e_3
90	0,027	0,033	0,027	0,026
81	0,029	0,036	0,030	0,029
72	0,032	0,040	0,032	0,031
63	0,036	0,044	0,037	0,035
54	0,042	0,052	0,042	0,041
45	0,050	0,062	0,051	0,048
36	0,062	0,075	0,063	0,061
27	0,083	0,099	0,085	0,080
18	0,121	0,148	0,125	0,116
9	0,227	0,280	0,227	0,222

Source: Author's own calculations.

Table 2. The relative efficiency coefficients. Mahalanobis depth function. The population of the size N=900. The sampling scheme for r=n=u.

n	e_0	e_1	e_2	e_3
90	0,027	0,033	0,026	0,027
81	0,029	0,035	0,029	0,028
72	0,033	0,041	0,033	0,032
63	0,038	0,045	0,037	0,037
54	0,042	0,053	0,043	0,041
45	0,050	0,061	0,052	0,048
36	0,062	0,073	0,062	0,060
27	0,081	0,098	0,082	0,079
18	0,121	0,154	0,121	0,118
9	0,223	0,274	0,216	0,220

Source: Author's own calculations.

The analysis of the Tables 1 and 2 lead to conclusion that there is not any significant difference between the relative efficiency coefficients of the estimation strategy based on auxiliary variable dependent on spherical depth functions and the appropriate relative efficiency coefficients of the estimation strategy based on an auxiliary variable dependent on the Mahanalobis depth functions. The Tables 1 and 2, as well as Figure 2 show that those coefficients decrease when the sample size increases.

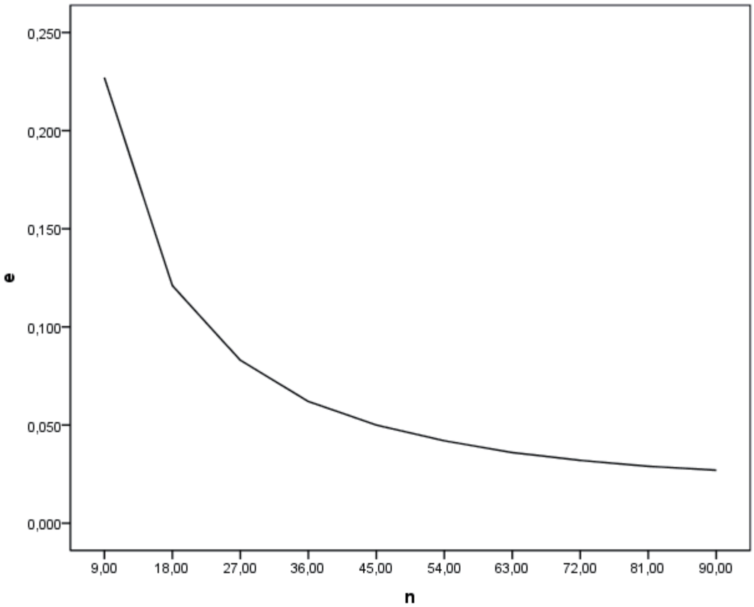

Figure 2. The relative efficiency coefficients for an increasing sample size.

Table 3. The relative efficiency coefficients. The spherical depth function. The population of the size *N=900.* The sampling scheme for *r=n=9* and *u≥9*.

u	e_0	e_1	e_2	e_3
9	0,227	0,280	0,227	0,222
50	0,225	0,280	0,226	0,219
100	0,221	0,272	0,228	0,212
150	0,224	0,277	0,226	0,217
200	0,224	0,277	0,224	0,218
300	0,224	0,279	0,225	0,218
400	0,225	0,278	0,230	0,216
500	0,225	0,287	0,231	0,215
600	0,225	0,286	0,226	0,218
700	0,225	0,288	0,230	0,213
800	0,235	0,294	0,234	0,229
850	0,255	0,298	0,247	0,253

Source: Author's own calculations.

Table 4. The relative efficiency coefficients.
The Mahanalobis depth function. The population of the size $N=900$. The sampling scheme for $r=n=9$ and $u \geq 9$.

u	e_0	e_1	e_2	e_3
9	0,223	0,274	0,216	0,220
50	0,222	0,277	0,216	0,218
100	0,220	0,283	0,214	0,215
150	0,224	0,286	0,219	0,220
200	0,222	0,276	0,216	0,218
250	0,221	0,277	0,217	0,217
300	0,224	0,282	0,216	0,221
400	0,224	0,279	0,219	0,221
500	0,225	0,281	0,219	0,221
600	0,225	0,283	0,219	0,0221
700	0,229	0,288	0,215	0,228
800	0,233	0,288	0,219	0,234
850	0,233	0,295	0,223	0,232

Source: Author's own calculations.

The Tables 1 and 2 consist of the relative efficiency coefficients of the conditional strategy $(\bar{y}_{HTS}, P(s \mid n, u, N))$. The analysis of these tables let us say that the conditional variant of the sampling design does not significantly lead to improving efficiency of the estimation.

4. Conclusions

The simulation analysis leads to the conclusion that the direct Horvitz-Thompson estimator of the domain mean under the proposed sampling designs dependent on depth functions of auxiliary variables is evidently more accurate than the simple sample mean. The sampling designs dependent on the spherical depth function and the Mahanalobis one approximately lead to the same accuracy of the estimation. The conditional versions of the sampling designs do not significantly improve the accuracy of the estimation. So, in practice, it is sufficient to use the direct Horvitz-Thompson estimator of the domain mean under the sampling designs dependent on spherical depth functions of auxiliary variables.

Acknowledgement

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

REFERENCES

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

LIU R.Y., PARELIUS J.M., SINGH K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *The Annals of Statistics*, vol. 27, No. 3, 783-858.

MAHALANOBIS P. C. (1936). On the generalized distance in statistics. *Proc. Nat. Acad. Sci. India* vol. 12, 49-55.

Tillé 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, 125-137.

WYWIAŁ J. L. (2008). Sampling design proportional to order statistic of auxiliary variable. *Statistical Papers,* vol. 49, No. 2/April, 277-289.

WYWIAŁ J. L. (2009). Sampling design proportional to positive function of order statistics of auxiliary variable. *Studia Ekonomiczne-Zeszyty Naukowe,* 2009, 53, 35-60.