A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 292, 2013

Janusz L. Wywiał[*]

ON SPACE SAMPLING DESIGNS[**]

1. INTRODUCTION

A population with identifiable units will be denoted by $U=\{1,...,N\}$. Fixed size samples drawn without replacement will be considered. An unordered sample of size n is defined by the subset $s=\{k_1,...,k_n\}\subset U$ of the population U. The sample space is denoted by $\mathbf{S}(U)$ or $\mathbf{S}$. Sampling design is as follows:

$$\bigwedge_{s\in\mathbf{S}} P(s)\geq 0 \text{ and } \sum_{s\in\mathbf{S}} P(s)=1. \tag{1}$$

Let us define the following sets:

$$\begin{aligned} B(k) &= \{s:\ k\in s\},\ k=1,\ ...,\ N,\\ B(k,t) &= \{s:\ k,\ t\in s\},\ k\neq t=1,\ ...,\ N. \end{aligned} \tag{2}$$

Inclusion probabilities are defined by the following expressions:

$$\pi_k=\sum_{s\in B(k)} P(s),\quad \pi_{k,t}=\sum_{s\in B(k,t)} P(s). \tag{3}$$

The well-known sampling design of the simple sample drawn without replacement is:

$$P_0(s)=\binom{N}{n}^{-1} \quad \text{for all } s\in\mathbf{S}. \tag{4}$$

Let $x=(x_{ij})$ be the matrix of dimensions $m\times N$, $i=1,\ldots,m$, $j=1,\ldots,N$. It consists of values of an m-dimensional auxiliary variable. Let $\mathbf{x}_{*j}^{\mathbf{T}}=[x_{1j}\ ...\ x_{mj}],\ \ j=1,...,N,$ be an observation of the m-dimensional variable

[*] Professor, Katowice University of Economics, Department of Statistics.
[**] The research was supported by the grant number N N111 434137 from the Ministry of Science and Higher Education.

attached to a *j*-th population element. The vector $\mathbf{x}_{i*} = [x_{i1} \dots x_{iN}]$, i=1,...,m, consists of observations of the i-th auxiliary variable. Then:

$$\mathbf{x} = (x_{*1}\ x_{*2} \dots x_{*N}) \quad \text{or} \quad \mathbf{x} = \begin{bmatrix} x_{1*} \\ \dots \\ x_{m*} \end{bmatrix}. \tag{5}$$

Values of auxiliary variables observed in a sample s of size n can be written as the matrix $\mathbf{x}_S = [x_{*j_1} \dots x_{*j_n}]$. Let $\bar{\mathbf{x}}^{\mathbf{T}} = [\bar{x}_1 \dots \bar{x}_i \dots \bar{x}_m]$ be the vector of the population means, where:

$$\bar{x}_i = \frac{1}{N}\sum_{j=1}^{N} x_{ij}, \quad i = 1, \dots, m, \quad \mathbf{z} = (x_{ij} - \bar{x}_i), \quad i = 1, \dots, m \text{ and } \mathbf{z}_s = [z_{*j_1} \dots z_{*j_n}],$$

where:

$$\mathbf{z}^{\mathbf{T}}_{*j_k} = [(x_{1j_k} - \bar{x}_1) \dots (x_{mj_k} - \bar{x}_m)], \quad k = 1, \dots, n, \quad \mathbf{u} = (x_{ij} - \bar{x}_i(s)),$$

where:

$$\bar{x}_{i(S)} = \frac{1}{N}\sum_{j \in S}^{N} x_{ij}, \quad \mathbf{u}_\mathbf{S} = [u_{*j_1} \dots u_{*j_n}],$$

where:

$$\mathbf{u}^{\mathbf{T}}_{*j_k} = [(x_{1j_k} - \bar{x}_1(s)) \dots (x_{mj_k} - \bar{x}_m(s))], \quad k = 1, \dots, n.$$

Then, $\mathbf{z}_S$ and $\mathbf{u}_S$ are sub-matrices of the matrices $\mathbf{z}$ and $\mathbf{u}$, respectively. Hence, the sub-matrices $\mathbf{z}_S$ and $\mathbf{u}_S$ can be obtained through dropping all the columns of the matrices $\mathbf{z}$ and $\mathbf{u}$ except those which correspond to the population elements drawn to the sample s.

The population variance-covariance matrix is as follows:

$$\mathbf{V}(\mathbf{x}) = \mathrm{N}^{-1}\mathbf{z}\mathbf{z}^{\mathrm{T}}, \quad \mathbf{V}_*(\mathbf{x}) = (\mathrm{N}-1)^{-1}\mathbf{z}\mathbf{z}^{\mathrm{T}}, \quad \mathbf{V}(\mathbf{x}) = (\mathrm{N}-1)\mathbf{V}_*(\mathbf{x})/\mathrm{N}. \tag{6}$$

The sample variance-covariance matrices can be defined as follows:

$$\mathbf{V}_s(\mathbf{x}) = \mathrm{n}^{-1}\mathbf{u}_\mathrm{S}(\mathbf{u}_\mathrm{S})^{\mathrm{T}}, \mathbf{V}_{*\mathrm{s}}(\mathbf{x}) = (\mathrm{n}-1)^{-1}\mathbf{u}_\mathrm{S}(\mathbf{u}_\mathrm{S})^{\mathrm{T}}, \mathbf{V}_s(\mathbf{x}) = (\mathrm{n}-1)\mathbf{V}_{*\mathrm{s}}(\mathbf{x})/\mathrm{n}. \tag{7}$$

$$\mathbf{V}_{\#\mathrm{s}}(\mathbf{x}) = (\mathrm{n}-1)^{-1}\mathbf{z}_\mathrm{S}(\mathbf{z}_\mathrm{S})^{\mathbf{T}}, \ \mathbf{V}_{\#\mathrm{S}}(\mathbf{x}) = \mathbf{V}_{*\mathrm{s}}(\mathbf{x}) + \frac{n}{n-1}(\bar{\mathbf{x}}_\mathbf{s} - \bar{\mathbf{x}})(\bar{\mathbf{x}}_\mathbf{s} - \bar{\mathbf{x}})^{\mathbf{T}}. \tag{8}$$

The trace of the variance-covariance matrix is defined as follows:

$$q^2(x) = tr(V(x)), \quad q_*^2(x) = tr(V_*(x)), \quad q^2(x) = \frac{N-1}{N} q_*^2(x). \tag{9}$$

The parameter $q^2(x)$ can be rewritten as follows.

$$q^2(x) = \frac{1}{N}\sum_{j=1}^{N} q_j^2, \quad q_j^2 = \sum_{i=1}^{m}\left(x_{ij} - \overline{x}_i\right)^2, \quad q_j = \sqrt{\sum_{i=1}^{m}\left(x_{ij} - \overline{x}_i\right)^2}. \tag{10}$$

The elements of the column vector x_{*j} can be treated as coordinates of a *j*-th point in m-dimensional space. Let the point with coordinates $\overline{\mathbf{x}}$ be the centre of population. Hence, qj is the distance between a *j*-th space point and the centre. Hence $q^2(x)$ is the mean squared distance between a space points and the centre. The parameter $q(x)$ is called the mean radius of the m-dimensional variable *x*. So, $q^2(x)$ is the squared mean radius of the m-dimensional variable *x*.
The sample squared mean radius is as follows:

$$q_S^2(x) = tr(V_S(x)), \quad q_{*S}^2(x) = tr(V_{*S}(x)), \quad q_{\#S}^2(x) = tr(V_{\#S}(x)). \tag{11}$$

Let us note that $q_{*s}^2(x) = \frac{n}{n-1} q_{s*}^2(x)$.
The parameter $q_s(x)$ is called the sample mean radius of the *m*-dimensional variable *x*. The parameter $q^2(x)$ can be rewritten as follows:

$$q_s^2(x) = \frac{1}{n}\sum_{j\in s} q_j^2(s), \quad q_j(s) = \sqrt{\sum_{i=1}^{m}\left(x_{ij} - \overline{x}_i(s)\right)^2}. \tag{12}$$

Moreover, we can easily show that:

$$q_s^2(x) = \frac{1}{2n(n+1)}\sum_{j\in s}\sum_{t\in s}\sum_{i=1}^{m}\left(x_{ij} - x_{it}\right)^2. \tag{13}$$

The population generalised variance is defined by the following expression:

$$\mathbf{g} = N^{-m}\left|\mathbf{z}\ \mathbf{z}^{\mathrm{T}}\right| \tag{14}$$

The sample generalised variances are defined by the formulas:

$$\mathbf{g}_{\mathbf{S}} = n^{-m}\,|\,\mathbf{u}_{\mathrm{S}}\,^{\mathbf{u}_{\mathrm{S}}^{\mathrm{T}}}\,|, \tag{15}$$

$$\mathbf{g}_{\#\mathbf{S}} = n^{-m} \mid \mathbf{z}_{\mathrm{S}}^{\mathbf{z}_{\mathrm{S}}^{\mathrm{T}}} \mid . \tag{16}$$

Particularly, if m=1:

$$g = V(x) = v = q^2(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

and

$$g_S = V_S(x) = v_S = q_S^2(x) = \frac{1}{n}\sum_{i\in S}(x_i - \bar{x}_S)^2.$$

2. SAMPLING DESIGN PROPORTIONAL TO THE DISTANCE FROM THE CENTRE OF SPACE POPULATION

The sampling design proportional to the sample mean of the auxiliary variable proposed by Lahiri (1951, pp. 133–140) and Midzuno (1952, pp. 99–107) and Sen (1953, pp. 119–127) adapted to our problem is as follows:

$$P_1(s) = \frac{\bar{q}_s}{\bar{q}}\binom{N}{n}^{-1} \quad \text{for all } s \in \mathbf{S}. \tag{17}$$

where: $q_s = \frac{1}{n}\sum_{i\in s} q_i$, $\bar{q} = \frac{1}{N}\sum_{k\in U} q_k$, see the expression (10). Hence, the sampling design $P_1(s)$ is proportional to the mean sample distance between space points (selected to the sample s) and the centre with coordinates denoted by $\bar{\mathbf{x}}^{\mathbf{T}} = [\bar{x}_1 ... \bar{x}_i ... \bar{x}_m]$. Lahiri (1951, pp. 133-140) and Midzuno (1952, pp. 99–107) proposed the following sampling scheme implementing the $P_1(s)$. The first element of the sample is selected with the probability $p(k) = \frac{q_k}{N\bar{q}}$, k=1,...,N and the next (n–1) elements are selected in the same way as the simple sample of size (n–1), drawn without replacement.

The inclusion probabilities of the first and second order are as follows, see Brewer and Hanif (1983), Wywiał (1991, pp. 21–23), Wywiał (2003):

$$\pi_k = \frac{N-n}{(N-1)N}\frac{q_k - \bar{q}}{\bar{q}} + \frac{n}{N}, \tag{18}$$

$$\pi_{kt} = \frac{n(n-1)}{N(N-1)} + \frac{(n-1)(N-n)}{(N-2)(N-1)N} \frac{q_k + q_t - 2\overline{q}}{\overline{q}}. \tag{19}$$

where $k \neq t = 1,...,N$.

Hence, the probability that the k-th population element will be selected to the sample is proportional to the distance q_k. This and the expression (17) implies that the sampling design $P_1(s)$ prefers drawing the elements which are far from the centre of the space population.

3. SAMPLING DESIGN PROPORTIONATE TO THE SQUARED SAMPLE MEAN RADIUS OF AUXILIARY VARIABLE

The considered sampling design is the following straightforward generalization of the well-known sampling design proposed by Singh and Srivastava (1980, pp. 205–209):

$$P_2(s) = \frac{1}{\binom{N}{n}} \frac{q_{*s}^2(\mathrm{x})}{q_*^2(\mathrm{x})}, \tag{20}$$

where the parameters $q_{*S}(x)$ and $q_{*S}(x)$ are given by the expressions (9)–(13). Hence, the defined sampling design is proportional to the mean of squared distances between space points (selected to the sample s) and the centre with coordinates denoted by $\overline{\mathbf{x}}^{\mathbf{T}} = \left[\overline{x}_1(s)...\overline{x}_i(s)...\overline{x}_m(s)\right]$, where: $\overline{x}_{i(S)} = \frac{1}{N}\sum_{j\in S}^{N} x_{ij}$. On the basis of the expression (13) we can say that the sampling design $P_2(s)$ is proportional to the mean of squared distances between all space points selected to the sample s. Hence, the sampling design prefers drawing samples with population objects (elements) which are far each from other.

If m=1 the above sampling design reduces to the Singh and Srivastava's (1980, pp. 205–209) sampling design. The sampling scheme implementing this sampling design is as follows. The first two elements are selected to the sample with probabilities:

$$\alpha(j,t) = \frac{\sum_{i=1}^{m}\left(x_{ij} - x_{it}\right)^2}{2\sum_{j=1}^{N-1}\sum_{t=j+1}^{N}\sum_{i=1}^{m}\left(x_{ij} - x_{it}\right)^2}, \tag{21}$$

where $j<t,\ j=1,\ldots,N-1,\ t=j+1,\ldots,N$. The next n-2 elements are selected using simple random sampling design without replacement from the set U-$\{j,t\}$. The probabilities of inclusion are as follows (see: Wywiał (1995)):

$$\pi_k=\frac{n}{N}+\frac{N-n}{N(N-2)}a_k, \tag{22}$$

where: $q_j^2=\sum_{i=1}^{m}\left(x_{ij}-\overline{x}_i\right)^2$, $a_k=\dfrac{q_i^2-q^2(x)}{q^2(x)}$, for $k=1,...,N$.

$$\begin{aligned}\pi_{kt}=&\frac{n(n-1)}{N(N-1)}+\frac{(N-n)}{N(N-2)}\left(1-\frac{(N-1)(N-n-1)}{N(N-3)}\right)\left(a_k+a_t\right)+\\&-\frac{2(N-n)(N-n-1)}{N^2(N-2)(N-3)}b_kb_t-\frac{2(N-n)(N-n-1)}{N^2(N-1)(N-2)(N-3)},\end{aligned} \tag{23}$$

where:

$$b_k=\frac{\sum_{i=1}^{m}\left(x_{ik}-\overline{x}_i\right)}{\sqrt{q^2(x)}},\quad \text{for } k\neq t=1,....,N.$$

The expression (22) implies that the k-th population element is drawn from a population with probability proportional to the squared distance between k-th population element and the central element with coordinates $\overline{\mathbf{x}}$.

Let us note that the defined by the equations (10) and (11) parameter $q^2(x)$ can be rewritten as follows:

$$q_s^2(x)=\frac{1}{2n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}q^2(\mathbf{x}_{*j},\mathbf{x}_{*k}),$$

where:

$$q^2(\mathbf{x}_{*j},\mathbf{x}_{*k})=\left(\mathbf{x}_{*j}-\mathbf{x}_{*k}\right)^T\left(\mathbf{x}_{*j}-\mathbf{x}_{*k}\right)=\sum_{i=1}^{m}\left(x_{ij}-x_{ik}\right)^2,$$

is the squared distance between the population the j-th and k-th population elements which are identified by the coordinates x_{*j} and x_{*j}, respectively. This and the expression (20) implies that the sampling design $P_2(s)$ is proportional to the sum of the squared distances between all population elements selected to the sample. Hence, the sampling design $P_2(s)$ prefers drawing the samples with elements which are largely spread. It means that each element of the sample is far from the other.

4. SAMPLING DESIGN PROPORTIONATE TO THE GENERALIZED VARIANCE OF AUXILIARY VARIABLE

Wywiał (1997, pp. 129–143) Wywiał (1999, pp. 73–87) Wywiał (1999a, pp. 259–281) generalized the Singh-Srivastava's sampling design to the following sampling designs proportional to the sample generalized variances of multidimensional auxiliary variable.

$$P_3(s) = \frac{|\mathbf{z}_s \mathbf{z}_s^{\mathbf{T}}|}{\binom{N-m}{n-m} N^m g}, \quad s \in \mathbf{S}, \tag{24}$$

where the matrix $\mathbf{z}_s \mathbf{z}_s^{\mathbf{T}}$ is defined in the first paragraph 1.

Let $\mathbf{z}(k_1,....k_r)$ be a sub-matrix obtained through eliminating the columns of numbers $k_1,...,k_r$ from the matrix $\mathbf{z}$. Wywiał (1997, pp. 129–143), Wywiał (1999, pp. 73–87) derived the sampling scheme and the probabilities of inclusion. Particularly, the inclusion probabilities of the first and second order are as follows:

$$\pi_k^{(3)} = 1 - \frac{N-n}{N-m} \quad \frac{|\mathbf{z}(k)\,\mathbf{z}^{\mathbf{T}}(k)|}{N^{(m)}\, g}, \quad k = 1, ..., N, \tag{25}$$

$$\pi_{kh}^{(3)} = 1 - \frac{N-n}{(N-m)\,N^m g}\left[\left|\mathbf{z}(h)\,\mathbf{z}^{\mathbf{T}}(h)\right| + \left|\mathbf{z}(k)\,\mathbf{z}^{\mathbf{T}}(k)\right| - \frac{N-n-1}{N-m-1}\left|\mathbf{z}(k,h)\,\mathbf{z}^{\mathbf{T}}(k,h)\right|\right]. \tag{26}$$

The next sampling design is as follows:

$$P_4(s) = \frac{n\,|\mathbf{u}_s \mathbf{u}_s^{\mathbf{T}}|}{\binom{N-m-1}{n-m-1} N^{m+1} g}, \quad s \in \mathbf{S}. \tag{27}$$

When m=1, the sampling design is reduced to the sampling design proportional to sample variance considered by Singh and Srivastava (1980, pp. 205–209). Let $\mathbf{x}(k_1,..., k_r)$ be a sub-matrix obtained through eliminating the columns of numbers $k_1, ..., k_r$ from the matrix $\mathbf{x}$. Moreover, let:

$$\mathbf{v}(k_1,...,k_w) = \mathbf{x}(k_1,...,k_w) - \overline{\mathbf{x}}(k_1,...,k_w)\,\mathbf{J}_{N-w}^{\mathbf{T}}.$$

$\mathbf{J}_{N\text{-}w}$ is the column vector with all its (N-w) elements equal to one and:

$$\overline{\mathbf{x}}(k_1,...,k_w) = \frac{1}{N-w}\,\mathbf{x}\,(k_1,...,k_w)\,\mathbf{J}_{N-w}.$$

Wywiał (1997, pp. 129–143, 1999, pp. 73–87) derived the probability of drawing without replacement elements $k_1,...,k_r$ to a sample s during the r fixed selections from a population. The inclusion probabilities of order r=1 and r=2 are as follows:

$$\pi_k^{(4)} = 1 - \frac{\binom{N-m-2}{n-m-1}(N-1)}{\binom{N-m-1}{n-m-1}N^{m+1}\,g}\left|\mathbf{v}(k)\,\mathbf{v}^{\mathrm{T}}(k)\right|, \tag{28}$$

$$\pi_{kh}^{(4)} = 1 - \frac{1}{\binom{N-m-1}{n-m-1}N^{m+1}\,g}\left\{\binom{N-m-2}{n-m-1}(N-1)\left[\left|\mathbf{v}(k)\,\mathbf{v}^{\mathrm{T}}(k)\right| + \left|\mathbf{v}(h)\,\mathbf{v}^{\mathrm{T}}(h)\right|\right] - \binom{N-m-3}{n-m-1}(N-2)\left|\mathbf{v}(k,h)\,\mathbf{v}^{\mathrm{T}}(k,h)\right|\right\}. \tag{29}$$

Let $\underline{\mathbf{x}}(k_1,...,k_{m+1})$ be a sub-matrix obtained through eliminating all the columns from the matrix $\mathbf{x}$ except the columns indexed by numbers k_1, ..., k_{m+1}. It is well known based on the multidimensional geometry that the squared volume of the parallelotop spanned on the points which coordinates are columns of the matrix $\underline{\mathbf{x}}(k_1,...,k_{m+1})$ is as follows:

$$\Delta^2(k_1,...,k_{m+1}) = \det^2\left[\underline{\mathbf{x}}(k_1,...,k_{m+1})\mathbf{J}_h\right].$$

It is well known, see e.g. Wywiał (2003) , that the generalized variance of an h-dimensional variable is proportional to the sum of the squared volumes of parallelotops:

$$g_s \propto \sum_{(k_1,...,k_{m+1})\in K_s}\Delta^2(k_1,...,k_{m+1}),$$

where: K_s is the set of all h-elements combinations without repetitions of elements of the sample s. Thus, when the observed in the sample population elements are well spread in the population then the generalized variance take rather large value. Hence, the sampling design $P_4(s)$ prefers selecting the sample which are well spread in the population. Similar conclusion can be made on the sampling design $P_3(s)$.

5. SAMPLING DESIGN DEPENDENT ON THE ORDER STATISTIC OF AUXILIARY VARIABLE FUNCTION

Let $Q(r)$ be the r-th order statistic from a simple sample drawn without replacement. Let $\alpha \in (0;1)$ and $(n\alpha)$ is the integer part of the value $n\alpha$. The sample quantile of the order α is defined as $Qs,\alpha=Q(r)$ where $r = (n\alpha)+1$ and $(r-1)/n \le \alpha < r/n$. Let $U_1=(1,...,i-1)$ be a subpopulation of the population U and let s_1 be the simple sample of 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 size $(n-r)$, drawn without replacement from U_2. Hence, $s = (s_1 \cup \{i\} \cup s_2)$ is such a sample that the value of the r-th order statistic of the distance variable - observed in the sample - equals q_i and the z-th order statistic of an auxiliary variable - observed in the sample - equals x_j. Let $S_{r,j} = \{s: Q(r)=q_i\}$. Hence, the set $S_{r,j}$ is the sample space of all such samples that $Q(r)=q_i$. Wywiał (2008, pp. 277–289) proposed the following sampling design proportional to the value of an order statistic of an auxiliary variable. In our case, it is proportional to a value qi of the $Q(r)$ order statistic:

$$P_5(s \mid r) = \frac{q_i}{\sum_{j=r}^{N-n+r} \binom{j-1}{r-1}\binom{N-j}{n-r} q_j}. \tag{30}$$

for $s \in \mathbf{S}(r;i)$, $i=r,\ldots,N-n+r$. Particularly, if $x_i = c>0$ for all $i=1,...,N$, the above sampling design reduces to the $P_0(s)$ simple sampling design. The conditional version of the sampling design is as follows:

$$P_5(s \mid r;u,v) = \frac{q_i}{\sum_{j=u}^{v} \binom{j-1}{r-1}\binom{N-j}{n-r} q_j}, \quad \text{for } s \in \mathbf{S}(r,i), \tag{31}$$

where $r \le u \le i \le v \le N-n+r$.
Let us define such a function $\delta(t)$ that if $t<0$, $\delta(t)=0$ else $\delta(t)=1$. Moreover, let:

$$z_r(u,v) = \sum_{j=u}^{v} q_j \binom{j-1}{r-1}\binom{N-j}{n-r}.$$

The inclusion probabilities of the first order are as follows:

$$\pi_k = \frac{\delta(u-k)\delta(r-1)\delta(v-1)\delta(u-1)}{z_r(u,v)} \sum_{i=u}^{v} \binom{i-2}{r-2}\binom{N-i}{n-r} q_i +$$
$$+ \frac{\delta(k-u+1)\delta(v-k+1)}{z_r(u,v)} \left(\delta(n-r)\delta(k-u)\delta(k-1) \sum_{i=u}^{k-1} \binom{i-1}{r-1}\binom{N-i-1}{n-r-1} q_i + \right.$$
$$+ \binom{k-1}{r-1}\binom{N-k}{n-r} q_k + \delta(r-1)\delta(v-k) \left. \sum_{i=k+1}^{v} \binom{i-2}{r-2}\binom{N-i}{n-r} q_i \right) + \tag{32}$$
$$+ \frac{\delta(k-v)\delta(n-r)\delta(N-v)}{z_r(u,v)} \sum_{i=u}^{v} \binom{i-1}{r-1}\binom{N-i-1}{n-r-1} q_i, \quad \text{for } k = 1,...,N.$$

The probabilities of the second order are derived by Wywiał (2008, pp. 277–289). The sampling scheme implementing the $P_5(s|u,v)$ conditional sampling design is as follows. Firstly, population elements are ordered according to the increasing values of the auxiliary variable. Secondly, the *i*-th element of the population is drawn with this probability:

$$p_2(i \mid r;u,v) = \frac{\binom{i-1}{r-1}\binom{N-i}{n-r} q_i}{\sum_{j=u}^{v} \binom{j-1}{r-1}\binom{N-i}{n-r} q_j}, \quad i = u,...,v. \tag{33}$$

Finally, the simple sample s_1 of size $(r-1)$ is drawn from the subpopulation U_1 and the simple sample s_2 of size $(n-r)$ from the subpopulation U_2.

Particularly, the sampling design $P_5(s|u,N-n+r)$ prefers such samples that *r*-th order statistic of the distance variable is greater than *u*.

The construction of the sampling design leads to the conclusion that probability of selecting the population element is proportional to its distance qi from the central element of the population with coordinates $\bar{\mathbf{x}}$. Hence, the elements placed far from the centre of the population are preferred to be drawn to the sample.

6. SAMPLING DESIGN DEPENDENT ON THE NEIGHBOUR MATRIX

It is assumed that the neighborhood of the population elements is fixed and identified by so called neighborhood matrix. Wywiał (1996, pp. 1185–1191) and Wywiał (2003) constructed sampling designs on the basis of that matrix. First sampling designs prefers drawing population elements, which are neighbors.

Next one prefers sampling elements which are not adjacent to each other. The position of population elements can be identified by neighborhood matrix

A=(a_{ij}). If the elements (i,j) are neighbors (are not neighbors) then a_{ij}=1 (a_{ij}=0). The sampling design, which prefers the neighbor elements to be drawn without replacement, is as follows:

$$P_6(s) = \frac{\sum_{i,j\in s} a_{ij}}{\sum_{s\in \mathbf{S}} \sum_{i,j\in s} a_{ij}}. \tag{34}$$

where the sampling space is denoted by **S**.

The design that prefers drawing without replacement the elements which are not neighbors is as follows:

$$P_7(s) \propto \frac{1}{2}(n^2 - n) + \alpha - \sum_{i,j\in s} a_{ij},$$

$P_7(s)>0$ provided that $\alpha>0$.

$$P_7(s) = \frac{\frac{1}{2}n(n-1) + \alpha - \sum_{i,j\in s} a_{ij}}{\binom{N}{n}\left[\frac{1}{2}n(n-1) + \alpha\right] - \beta}, \tag{35}$$

where:

$$\beta = \sum_{s\in S} \sum_{i,j\in s} a_{ij}.$$

It is obvious that:

$$\lim_{\alpha\to\infty} P_7 = \frac{1}{\binom{N}{n}}.$$

Hence, if $\alpha\to\infty$, the sampling design $P_7(s)$ tends to be a simple sampling design. So, in practice the parameter α should be assumed to be small.

Figure 1. The population.

1	2	3	4	5

Source: author's own.

According the above definition the matrix **A** is as follows:

$$\mathbf{A}=\begin{array}{|c|c|c|c|c|}\hline 1&1&0&0&0\\\hline 1&1&1&0&0\\\hline 0&1&1&1&0\\\hline 0&0&1&1&1\\\hline 0&0&0&1&1\\\hline\end{array}$$

When we assume that α=1 on the basis of the expression (35) we calculate that β=13 and the probabilities of selecting the e.g. the samples s_1=(1,2,3), s_1=(2,3,5) are $P_8(s_1)$=2/27, $P_8(s_2)$=3/27, respectively.

Finally, let us propose the following neighborhood matrix $\mathbf{B}=(b_{ij})$. If the elements (i,j) are neighbors (are not neighbors) then b_{ij}=0 (b_{ij}=1), i=1,…,N. Moreover, we assume that b_{ii}=0, i=1,…,N. The sampling design preferring the elements which are not neighbor to be drawn without replacement, is as follows:

$$P_8(s)=\frac{\sum\limits_{i,j\in s} b_{ij}}{\sum\limits_{s\in\mathbf{S}}\sum\limits_{i,j\in s} b_{ij}}. \tag{36}$$

For instance, according the above definition, the matrix **B** for the population shown by Figure 1 is as follows:

$$\mathbf{B}=\begin{array}{|c|c|c|c|c|}\hline 0&0&1&1&1\\\hline 0&0&0&1&1\\\hline 1&0&0&0&1\\\hline 1&1&0&0&0\\\hline 1&1&1&0&0\\\hline\end{array}$$

Hence, the probabilities of selecting the e.g. the samples s_1=(1,2,3) and s_1=(2,3,5) are $P_8(s_1)$=1/18, $P_8(s_2)$=2/9, respectively.

Finally, let us define the elements of the matrix $\mathbf{B}=(b_{ij})$ as follows. Let k be the minimal number of population elements which separate the i-th and j-th population elements. We assume that $b_{ij}=k$, i, j=1,…,N, k=0,.1,2,…,S. If k=0, the elements (i,j) are neighbors. We assume that b_{ii}=0 for i=1,…,N. When k=1, there is one population element between an i-th and a j-th elements and so on. For instance, in the case of the spatial population the neighborhood shown by Figure 1 the matrix $\mathbf{B}=(b_{ij})$ is as follows.

$$\mathbf{B}=\begin{array}{|c|c|c|c|c|}\hline 0&0&1&2&3\\\hline 0&0&0&1&2\\\hline 1&0&0&0&1\\\hline 2&1&0&0&0\\\hline 3&2&1&0&0\\\hline\end{array}$$

So, the probabilities of selecting the e.g. the samples s_1=(1,2,3) and s_1=(2,3,5) are $P_8(s_1)$=1/30, $P_8(s_2)$=0.1, respectively.

The considered in this section sampling designs are defined on the basis of the neighborhood matrixes which can be treated as a kind nonparametric distance matrix between elements of a spatial population. The sampling designs $P_7(s)$ and $P_8(s)$ prefer to draw the samples consisting of population elements which are not neighbors.

7. HORVITZ-THOMPSON ESTIMATOR

An observation of a variable under study (an auxiliary variable) attached to the i-th population element will be denoted by y_i $(x_i>0)$, i=1,...,N. The well-known Horvitz-Thompson (1952, pp. 663–685) estimator is as follows:

$$t_{HTS} = \frac{1}{N}\sum_{k\in s}\frac{y_k}{\pi_k} \tag{37}$$

It is well known that the strategy $(t_{HTS}, P(s))$ is unbiased for the population mean when all inclusion probabilities are positive. Moreover, let us note that $(t_{HTS}, P_0(s)) = (\overline{y}_s, P_0(s))$ is the mean from the simple sample drawn without replacement. The variance of the strategy is as follows:

$$D^2\left(t_{HTS}\right) = \frac{1}{N^2}\sum_{k=1}^{N}\left(\frac{y_k}{\pi_k}\right)^2\pi_k\left(1-\pi_k\right) + \frac{1}{N^2}\sum_{k\neq i}^{N}\sum_{=1}^{N}\frac{y_k y_i}{\pi_k\pi_i}\left(\pi_{ki}-\pi_i\pi_k\right) \tag{38}$$

The unbiased estimator of the Horvitz-Thompson statistic's variance is as follows:

$$D_S^2\left(t_{HTS}\right) = \frac{1}{N^2}\sum_{k\in S}\left(\frac{y_k}{\pi_k}\right)^2\left(1-\pi_k\right) + \frac{1}{N^2}\sum_{k\neq i}^{N}\sum_{=1}^{N}\frac{y_k y_i}{\pi_k\pi_i}\frac{\pi_{ki}-\pi_i\pi_k}{\pi_{ki}} \tag{39}$$

Hence, the presented estimator can be used to estimate a mean value of a variable under study in a space population.

8. CONCLUSIONS

Presented sampling designs can be useful in the case when the observations of a multidimensional auxiliary variable are known in the entire population. For instance, the auxiliary variable can be defined as coordinates of population elements in space. Such auxiliary information let assess distance between space

population elements in several aspects. The proposed sampling designs are functions of distance measures. Usually, the presented sampling designs prefer drawing samples including population elements which are far from the space population centre. Another sampling design prefers to select the population elements in such a way that the distances between them are large. Moreover, there were considered sampling designs preferring to draw the samples consisting of population elements which are not neighbors.

REFERENCES

Brewer K. R.W., Hanif M. (1983), *Sampling with unequal probabilities*. Springer Verlag, New York-Heidelberg-Berlin 1983.

Horvitz D. G., Thompson D. J. (1952), *A generalization of sampling without replacement from finite universe*. Journal of the American Statistical Association, vol. 47.

Lahiri G. W. (1951), *A method for sample selection providing unbiased ratio estimator*. Bulletin of the International Statistical Institute, vol. 33.

Midzuno H. (1952), *On the sampling system with probability proportional to the sum of sizes*, Annals of the Institute of Statistical Mathematics .

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

Singh P., Srivastava A.K. (1980), *Sampling schemes providing unbiased regression estimators*. Biometrika, vol. 67.

Wywiał J. L. (1991), *On sampling design proportional to mean value of an auxiliary variable (in Polish).* Wiadomości Statystyczne, nr 6, 1991.

Wywiał J. L. (1996), *On space sampling.* Statistics in Transition. vol. 2, nr 7.

Wywiał J. L. (1997), *Sampling Design proportional to the Sample Generalized Variance of Auxiliary Variables.* Proceedings of 16th International Conference on Multivariate Statistical Analysis- MSA'97. Edited by Cz. Domanski and D. Parys. November 27-29 1997. Department of Statistical Methods, Institute of Econometrics and Statistics, University of Łódź, Polish Statistical Association. November 27–29. 1997r., pp. 129–143

Wywiał J. L. (1999), *Sampling designs dependent on the sample generalized variance of auxiliary variables*. Journal of the Indian Statistical Association. Vol. 37.

Wywiał J. L. (1999a), *Generalization of Singh and Srivastava's schemes providing unbiased regression estimatiors*, Statistics in Transition vol. 2, No. 2.

Wywiał J. L. (2003), *Some Contributions to Multivariate Methods in Survey Sampling*, Katowice University of Economics, Katowice.

Wywiał J. L. (2008), *Sampling design proportional to order statistic of auxiliary variable*, Statistical Papers, vol. 49, Nr. 2/April.

Janusz L. Wywiał

ON SPACE SAMPLING DESIGNS

Statistical research dealing the regional economic problems are based on the spatial data. When spatial populations are large then data about the population elements have to be observed in random samples. In the paper a review of the sampling designs used to draw samples from spatial populations is presented. Especially, the complex sampling designs dependent on auxiliary varia-

bles are considered. It is well known that a spatial population should be well covered by the sample. We show that this property is fulfilled by the sampling designs considered in the paper. Moreover, it is mentioned that the sampling designs can be applied to the estimation of the population average in a finite spatial population by means of the well-known Horvitz-Thompson statistic. In general, sampling design proportional to the value of positive function of multidimensional auxiliary variable is considered. It is assumed that all observations of the auxiliary variables are known. Observations of the auxiliary variable can be treated as coordinates of appropriate points in multidimensional space. The sampling designs proportional to the mean of distances between population points and a population centre, to the trace of variance-covariance matrix, to the generalized variance of the auxiliary variable are considered. Some sampling designs proportional to functions of order statistics of the auxiliary variable are presented, too. Finally, the sampling designs dependent on a neighborhood matrix are considered. Sampling schemes implementing the sampling designs are shown, too.

O LOSOWANIU PRZESTRZENNYM

W pracy przedstawiono plany i schematy losowania prób nieprostych z populacji skończonej i ustalonej zależne od obserwacji wielowymiarowej zmiennej dodatkowej. Zakłada się, że obserwacje tej zmiennej są ustalone (nielosowe) i znane w całej populacji. W szczególności geometrycznym obrazem obserwacji zmiennych dodatkowych mogą być współrzędne punktów na płaszczyźnie (w przestrzeni) euklidesowej. Zaprezentowano następujące plany losowania: Plan proporcjonalny do średniej odległości obserwowanych w próbie punktów przestrzeni od jej punktu traktowanego jako centralnym. Plany proporcjonalne do: śladu macierzy wariancji i kowariancji z próby wektorowej zmiennej dodatkowej albo jej uogólnionej wariancji. Następny plan jest proporcjonalny do wartości statystyki pozycyjnej zmiennej dystansowej. W końcu przedstawiono plany zależne od pewnej macierzy sąsiedztwa elementów obserwowanych w próbie. W pracy również zasygnalizowano, że prezentowane plany losowania są użyteczne przy estymacji wartości średniej zmiennej badanej w populacji za pomocą znanego estymatora Horvitza-Thompsona.