ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 156, 2002

*Janusz Wywial**

ON STRATIFICATION OF POPULATION ON THE BASIS OF AUXILIARY VARIABLE AND THE SELECTED SAMPLE

Abstract. Survey sampling conditional methods are usually connected with post-stratification estimators for domains and with inference on the basis of regression models or contingency tables. These problems were considered e.g. by Rao (1985), Tillé (1998), Williams (1962). The problem of stratification of a population on the basis of observations on the variable under study in a sample was considered by e.g. Dalenius (1957).

We deal with the problem of appropriate division of a simple sample into sub-samples of equal sizes. This partition of the sample leads to clustering a population into sub-populations. Each of these sub-populations includes one and only one previously created sub-sample. The linear combinations of statistics from the sub-samples are used for estimation of a population mean. The coefficients of this linear combination are proportionate to the sizes of the sub-populations. This statistic is the unbiased estimator of the population mean. The variance of the estimator has been derived. The example of determining of the estimator parameters is presented. Moreover, some generalisations of proposed estimators are suggested.

Key words: stratification after sample selection, conditional estimation, conditional mean, conditional variance.

I. ESTIMATOR

Let us assume that the values of an auxiliary variable are known in a population of size N. Its i-th value is denoted by x_i, $i = 1, ..., N$. An i-th value of a variable under study is denoted by y_i, $i = 1, ..., N$. Let us assume that the elements of the population $U = [1, ..., N]$ are ordered in such a way that $x_i < x_j$ for each $i < j = 1, ..., N$. The simple sample s of the size n is drawn without replacement from a fixed and finite population U.

* Prof., Department of Statistics, University of Economics in Katowice.

Let us divide each sample $s = \{i_1, ..., i_k, i_{k+1}, ..., i_n\}$, where $i_j < i_h$ if $i < h$, into H following sub-samples of size m: $s_h(x_k) = \{i_{m[h-1)}+1, ..., i_{mh}\}$, $h = 1, 2, ..., H < N$. Hence, $s_h(x_k) \cap s_t(x_k) = \emptyset$ for each $h \neq t = 1, ..., H$ and $\bigcup_{h=1}^{H} s_h = s$. Let $U_{s_1} = \{i: x_i \leqslant x_h\}$, $U_{s_h} = \{i: x_{h-1} < x_i \leqslant x_h\}$, $h = 2, ..., H-1$ and $U_{s_H} = \{i: x_i > x_{H-1}\}$. Hence $U_{s_H} \cap U_{s_t} = \emptyset$ for each $h \neq t = 1, ..., H$, and $\bigcup_{h=1}^{H} U_{s_h} = U$, $k = 1, ..., H$.

Let $\Omega = \{s\}$ be a space sample. In our case the set Ω consists of $\binom{N}{n}$ samples s. Let $\Omega(\mathbf{x}_k)$ be the set of such samples $s \in \Omega$ that $\mathbf{x}_k = \lfloor x_{k_1}, x_{k_2} ... x_{k_{H-1}} \rfloor$ is fixed. Hence, $\Omega = \bigcup_k \Omega(\mathbf{x}_k)$ and $\Omega(\mathbf{x}_k) \cap \Omega(\mathbf{x}_h) = \emptyset$ for $k \neq h$ (see general considerations e.g. in Flachsmeyer J. (1977)). The value $\mathbf{x}_k$ can be treated as the outcome of the random vector $\mathbf{X} = [X_1 ... X_{H-1}]$. Its probability distribution function is determined by the expression:

$$P(\mathbf{X} = \mathbf{x}_k) = \frac{\text{size}(\Omega(\mathbf{x}_k))}{\binom{N}{n}} \tag{1}$$

Let us assume that the simple sample s is drawn without replacement and its size is $n = Hm$, where $m \geqslant 1$ and $n < N$. Moreover, let $s = \{i_1, ..., i_n\}$ and $x_{i_j} < x_{i_e}$ and $i_j < i_e$ if and only if $j < e$. The sample s is divided into H sub-samples $s_h = \{i_{m(h-1)+1}, ..., i_{mh}\}$, $h = 1, ..., H$. Let us assume that $i_{mh} = k_h$, $h = 1, ..., H-1$. Hence, x_{k_h} is the sample quantil of order mh/n of the auxiliary variable. The number k_h identifies the position of the sample quantil in the population.

Wilks (1962), p. 252, considered the distribution of the order statistics in the simple sample drawn without replacement from a finite population. The particular case of this distribution is the probability distribution of the random vector $\mathbf{K} = [K_1, ..., K_{H-1}]$. If $m \geqslant 1$ and $n = Hm < N$:

$$P(K_1 = k_1, ..., K_{H-1} = k_{H-1}) = \frac{\binom{k_1 - 1}{m-1}\binom{k_1 - k_2 - 1}{m-1} ... \binom{k_{H-2} - k_{H-1} - 1}{m-1}\binom{N - k_{H-1}}{m}}{\binom{N}{Hm}} \tag{2}$$

where: $m \leqslant k_1 < k_2 ... k_{H-1} < N - m$ or

$$P(K_1 = k_1, \ldots, K_{H-1} = k_{H-1}) = \frac{\binom{N - k_{H-1}}{m} \prod_{h=1}^{H-1} \binom{k_h - k_{h-1} - 1}{m-1}}{\binom{N}{Hm}} \tag{3}$$

where $k_0 = 0$.

Particularly, if $H = 2$ the $k_1 = k$ is the sample median and:

$$P(K = k) = \frac{\binom{k-1}{m-1}\binom{N-k}{m}}{\binom{N}{2m}}, \quad k = m, \ldots, N - m \tag{4}$$

If $m = 1$, $n = H$:

$$P(K_1 = k_1, \ldots, K_{H-1} = k_{H-1}) = \frac{N - k_{H-1}}{\binom{N}{H}}, \quad k = 1, \ldots, N - 1$$

In the case when $H = n = 3$, $m = 1$ and $N = 5$ the distribution of the variable $[K_1\ K_2]$ is determined by the Table 1.

Table 1

(k_1, k_2)	$k_1 = 1$	$k_1 = 2$	$k_1 = 3$
$k_2 = 2$	0.3	0	0
$k_2 = 3$	0.2	0.2	0
$k_2 = 4$	0.1	0.1	0.1

If $H = 2$, $m = 1$ and $n = 2$, the distribution is reduced to one determined by the equation:

$$P(K = k) = \frac{2(N-k)}{(N-1)N}, \quad k = 2, \ldots, N - 1$$

If $H = 2$, $m = 1$ and $n = 3$ then

$$P_2(K = k) = \frac{6(N-k)}{(N-2)(N-1)N}, \quad k = 2, \ldots, N - 1$$

For instance, if $N = 5$, $m = 1$ and $n = H = 2$ then $P_2(K = 1) = 0.4$, $P_2(K = 2) = 0.3$, $P_2(K = 3) = 0.2$, $P_2(K = 4) = 0.1$. If $N = 5$, $H = 2$, $m = 1$

and $n = 3$ then $P_2(K = 2) = 0.3$, $P_2(K = 3) = 0.4$, $P_2(K = 4) = 0.3$. If $N = 5$, $m = 2$, $H = 2$ and $n = 4$: $P_2(K = 2) = 0.6$, $P_2(K = 3) = 0.4$. For $N = 6$, $m = 2$, $H = 2$ and $n = 4$: $P_2(K = 2) = 0.4$, $P_2(K = 3) = 0.4$, $P_2(K = 4) = 0.2$.

Let us consider the following conditional estimator of the population average $\bar{y}$:

$$\tilde{y}_{S/K} = \sum_{h=1}^{H-1}\left(\frac{K_h - K_{h-1} - 1}{N}\right)\bar{y}_{S_h'} + \frac{1}{N}\sum_{h=1}^{H-1} y_{K_h} + \left(1 - \frac{K_{H-1}}{N}\right)\bar{y}_{S_H} \tag{5}$$

where: $S_h' = S_h - \{K_h\}$ and

$$\begin{cases} \bar{y}_{S_h'} = \dfrac{1}{m-1}\sum_{i \in S_h'} h_i, \quad h = 1, \ldots, H \\ \bar{y}_{S_H} = \dfrac{1}{m}\sum_{i \in S_H} y_i \end{cases} \tag{6}$$

The expected value of this statistic is derived in the following way:

$$E(\tilde{y}_{S/K}) = E_K(E_{S/K}(\bar{y}_{S/K}/K) =$$

$$= E_K\left(\sum_{h=1}^{H-1}\left(\frac{K_h - K_{h-1} - 1}{N}\right)E_{S/K}(\bar{y}_{S_h'}) + \frac{1}{N}\sum_{h=1}^{H-1} y_{K_h} + \left(1 - \frac{K_{H-1}}{N}\right)E_{S/K}(\bar{y}_{S_H})\right) =$$

$$= E_K\left(\sum_{h=1}^{H-1}\left(\frac{K_h - K_{h-1} - 1}{N}\right)\bar{y}_{U_h'} + \frac{1}{N}\sum_{h=1}^{H-1} y_{K_h} + \left(1 - \frac{K_{H-1}}{N}\right)\bar{y}_{U_H}\right) =$$

$$= E_K(\bar{y}) = \bar{y}$$

where: $U_h' = U_h - \{K_h\}$ and

$$\begin{cases} y_{U_h'} = \dfrac{1}{K_h - K_{h-1} - 1}\sum_{i \in U_h'} y_i, \quad h = 1, \ldots, H-1 \\ y_{U_H} = \dfrac{1}{N - K_{H-1}}\sum_{i \in U_H} y_i \end{cases} \tag{7}$$

Hence:

$$\begin{cases} E_{S/k}(\tilde{y}_{S/k}) = \bar{y} \\ E(\tilde{y}_{S/k}) = \bar{y} \end{cases} \tag{8}$$

In conclusion, the statistic $\tilde{y}_{S/K}$ is a conditionally and unconditionally unbiased estimator of the population mean.

The derivation of the variance is as follows:

$$D^2(\bar{y}_{S/K}) = E_K(D^2_{S/K}(\bar{y}_{S/K}|K)) + D^2_K(E_{S/K}(\bar{y}_{S/K}|K)) = E_K(D^2_{S/K}(\bar{y}_{S/K}|K)) + 0 =$$

$$= E_K\left(\sum_{h-1}^{H-1}\left(\frac{K_h - K_{h-1} - 1}{N}\right)^2 D^2_{S/K}(y_{S_h'}) + 0 + \left(1 - \frac{K_{H-1}}{N}\right)^2 D^2_{S/K}(\bar{y}_{S_H})\right) =$$

$$= E_K\left(\sum_{h=1}^{H-1}\left(\frac{K_h - K_{h-1} - 1}{N}\right)^2 \frac{K_h - K_{h-1} - m}{(K_h - K_{h-1})(m-1)} v_{U_h'} + \left(1 - \frac{K_{H-1}}{N}\right)^2 \frac{N - K_{H-1} - m}{(N - K_{H-1})m} v_{U_H}\right) \tag{9}$$

The unbiased estimator of the variance $D^2(\bar{y}_{S/K})$ is shown by the equation:

$$d^2_S(\bar{y}_{S/K}) = \sum_{h=1}^{H-1}\left(\frac{K_h - K_{h-1} - 1}{N}\right)^2 \frac{K_h - K_{h-1} - m}{(K_h - K_{h-1})(m-1)} v_{S_h'} +$$

$$+ \left(1 - \frac{K_{H-1}}{N}\right)^2 \frac{N - K_{H-1} - m}{(N - K_{H-1})m} v_{S_H} \tag{10}$$

where:

$$\begin{cases} v_{S_h'} = \dfrac{1}{m-1} \sum\limits_{i \in S_h'} (y_i - y_{S_h'})^2, \quad h = 1, \ldots, H-1 \\ v_{S_H} = \dfrac{1}{m} \sum\limits_{i \in S_H} (y_i - y_{S_H})^2 \end{cases} \tag{11}$$

II. EXAMPLE OF SIMULATION STUDY OF THE ESTIMATION EFFICIENCY

Let us consider the particular case when $H = 2$. The distribution of 30 observations $(x; y)$ of a two-dimensional variable is shown by the Fig. 1. The basic parameters of this variable in the population consisting of 30 elements are as follows: the average of auxiliary variable $\bar{x} = 68.6824$, the mean of the variable under study $\bar{y} = 93.6536$, the variances of auxiliary variable and the variable under study $v_x = (89.1094)^2$, $v_y = (17.6015)^2$, respectively and finally the correlation coefficient between these variables $r = 0.9940$.

Let the population average be estimated by means of the estimators $\bar{y}_{S/K}$. The simple sample drawn without replacement has 5 elements. The sample space consists of $\binom{30}{5}$ samples. On the basis of all these possible samples, the conditional (and the unconditional) expected values and variances of both estimators have been calculated. The variance of the simple sample mean is $D^2(\bar{y}_S) = 51.6356$ and $D^2(\bar{y}_{S/K}) = 42.9823$.

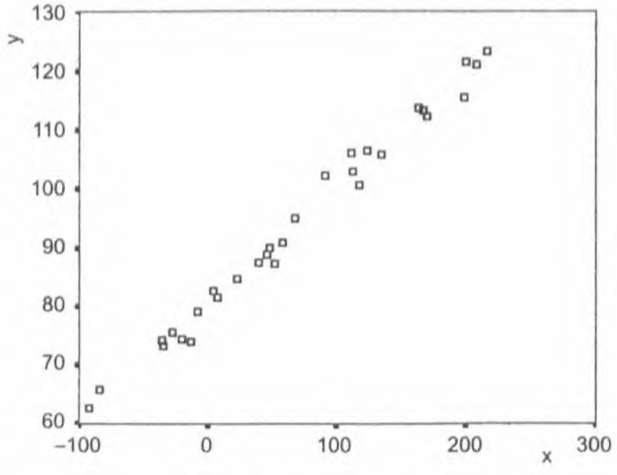

Fig. 1. The scatter plot for variables x and y in the population

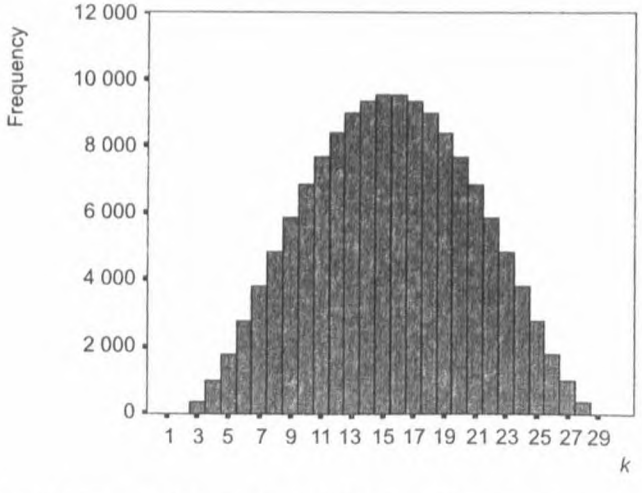

Fig. 2. The distribution of the random variable K in the case of the estimator $\tilde{y}_{S/K}$

The relative efficiency is defined by the expression: $e = (100\%)D^2(\tilde{y}_{S/K})/D^2(\bar{y}_S)$. In our case $e = 83.24\%$. Hence, the precision of the conditional estimators $\tilde{y}_{S/K}$ is better than the precision of the simple sample mean.

As it was defined by the expression (4), the outcome k of the random variable K is the number of the population element dividing the sample into two sub-samples. The probability distribution of the random variable K is presented by the Fig. 2.

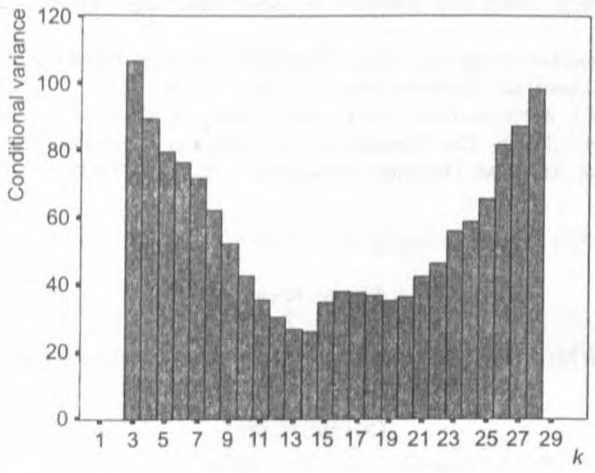

Fig. 3. The conditional variances of the estimator $\bar{y}_{s/K}$

The conditional variances of the estimator $\bar{y}_{s/K}$ are represented by the Fig. 3.

The above considered conditional method of estimation can be generalised in several directions. Firstly, in the case of two auxiliary variables the sample quantils let us divide the population into m^2 non-empty and disjoint sub-populations. Secondly, instead of a one-dimensional auxiliary variable and a variable under study, the multidimensional ones can be considered because, usually, the vector of population means is estimated and the vector of auxiliary variables can be available. For instance in this case the precision of the estimation of mean vector can be determined by trace of variance-covariance matrix or by generalised variance.

ACKNOWLEDGEMENT

The research was supported by the grant number 1 H02B 008 16 from the State Committee for Scientific Research (KBN).

REFERENCES

Dalenius T. (1957), *Sampling in Sweden. Contribution to Methods and Theories of Sample Survey Practice*, Almqwist & Wiksells, Stockholm.

Flachsmeyer J. (1977), *Combinatorics* (in Polish), PWN, Warszawa.

Rao J. N. K. (1985), *Conditional Inference in Survey Sampling*, „Survey Methodology", **11**, 1, p. 15–31.
Tillé (1998), *Estimation in Surveys Using Conditional Inclusion Probabilities: Simple Random Sampling*, „International Statistical Review", **66**, 3, p. 303–322.
Wilks S. S. (1962), *Mathematical Statistics*, John Wiley & Sons, Inc. New York, London.
Williams W. H. (1962), *The Variance of an Estimator with Part-Stratified Weighting*, „Journal of the American Statistical Association", **57**, p. 622–627.

Janusz Wywiał

O WARSTWOWANIU POPULACJI NA PODSTAWIE ZMIENNEJ POMOCNICZEJ I PRÓBY PO JEJ WYLOSOWANIU

(Streszczenie)

Problem estymacji wartości przeciętnej w populacji na podstawie próby prostej losowanej bezzwrotnie z populacji ustalonej i skończonej jest rozważany. Zakładamy, że wartości zmiennej pomocniczej są obserwowane w całej populacji. Próba prosta, po jej wylosowaniu, jest porządkowana zgodnie z rosnącymi wartościami zmiennej pomocniczej. Następnie próba ta jest dzielona na $H > 1$ równolicznych podprób. Potem zlicza się, ile jest elementów populacji pomiędzy elementami rozdzielającymi podpróby. Udziały tych liczebności stanowią współczynniki kombinacji liniowej, m.in. średnich z podprób. Taki warunkowy estymator daje nieobciążone (warunkowo i bezwarunkowo) oceny wartości średniej w populacji. Pokazano przykład oceny wartości średniej w populacji z wyznaczeniem wartości wariancji warunkowych i bezwarunkowych estymatora.