PLISKA
STUDIA MATHEMATICA BULGARICA
ПЛИСКА
БЪЛГАРСКИ МАТЕМАТИЧЕСКИ СТУДИИ

For further information on
Pliska Studia Mathematica Bulgarica
visit the website of the journal http://www.math.bas.bg/~pliska/
or contact: Editorial Office
Pliska Studia Mathematica Bulgarica
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Telephone: (+359-2)9792818, FAX:(+359-2)971-36-49
e-mail: pliska@math.bas.bg

Pliska Stud. Math. Bulgar. **18** (2007), 91–102

PLISKA
STUDIA MATHEMATICA
BULGARICA

ESTIMATION OF FRACTION OF DISTINGUISHED ELEMENTS IN POPULATION BASED ON PARTIALLY REALIZED RANDOM SAMPLE

Wiesława Dabała Bronisław Lednicki

This article presents an application of representative method in public opinion polls.

1. Introduction

The finite population comprised persons aged 18 or above, residing in the territory of Poland. For organizational reasons, individual sampling could not be used. Therefore, the sample was selected using a three-stage sampling scheme, where:

– first-stage sampling units were census units (little areas with no more than 300 housing units);

– second-stage sampling units were housing units (*dwellings, flats or houses in rural areas*);

– third-stage sampling units were adults from the housing units.

2000 *Mathematics Subject Classification:* 62D05

Key words: applications of mathematical statistics, applications of representative method in public opinion polls

2. Stratification of the population

First-stage sampling units, i.e. census units, were divided into strata h ($h = 1, 2, 3, \ldots, 64$). One stratum comprised the census units belonging to the same class with regard to the population of a province (voivodship). In each of the 16 administrative provinces of Poland, the localities have been divided into the following categories:

- provincial capitals,
- other cities/towns,
- villages.

The number of strata was bigger in the provinces with a bigger adult population. The predetermined number of census units to be sampled was stratified pro rata to the number of second-stage sampling units in a stratum, i.e. proportional to the number of housing units.

Let:

n^* – total number of census units to be sampled,

n_h^* – number of census units to be sampled in the h^{th} stratum,

$n^* = \sum_{h=1}^{64} n_h^*$,

h – number of a stratum, $h = 1, 2, 3, \ldots, 64$

then:

$$n_h^* = \frac{M_h}{M} n^* \tag{1}$$

where:

M_h – number of secondary sampling units (dwellings) in the h^{th} stratum,

$M_h = \sum_{j=1}^{N_h} M_{hj}$

M_{hj} – number of dwellings in the j^{th} census unit belonging to the h^{th} stratum;

M – total number of dwellings in the population.

$M = \sum_{h=1}^{64} \sum_{j=1}^{N_h} M_{hj} = \sum_{h=1}^{64} M_h$

N_h – number of census units in the h^{th} stratum;

N – total number of census units, $N = \sum_{h=1}^{64} N_h$

3. Sampling of first-stage units, i.e. sampling of census units

First-stage units were selected using a procedure proposed by W. G. Madow in 1949 and subsequently modified by H. O. Hartley and J. N. K. Rao in 1962. This procedure consists of systematic sampling with selection probabilities proportional to the value of the attribute adopted as the criterion of the sampling unit size, after a prior randomly ordering of these units. In this case, the numbers of housing units in census units are the size criterion. In each stratum, the sampling procedure was as follows:

a) Census units were randomly ordered;

b) First order probabilities; Π_{hj} - were determined:

$$\Pi_{hj} = n_h^* \frac{M_{hj}}{M_h} \tag{2}$$

c) A cummulative series of probabilities was created:

$$\Pi_{hj}^* = \sum_{j=1}^{k_h} \Pi_{hj} \tag{3}$$

where:
$\Pi_{h,0}^* = 0$ and $\Pi_{h,N_h}^* = n_h^*$

$j = 1, 2 \ldots, k_h \qquad k_h = 1, 2, \ldots, N_h$

d) The sampling interval $t = 1$ was determined and a random sampling starting point a was determined from the $\langle 0; 1\rangle$ range.

e) The j^{th} census unit was included in the sample, if the following inequality were satisfied:

$$\Pi_{h,j-1}^* < a + (b-1) \leq \Pi_{h,j}^* \tag{4}$$

where:
$j = 1, 2 \ldots, N_h \qquad b = 1, 2, \ldots, n_h^*.$
Census units from each stratum h were selected using the same procedure.

4. Sampling of second-stage units, i.e. sampling of housing units

Housing units (second-stage units) were selected from the previously sampled census units using simple sampling scheme without replacement. Six housing units were sampled from each census unit.

5. Sampling of third-stage units, i.e. selecting adults to be interviewed from a housing units

Adults were selected from previously sampled housing units (secondary sampling units). One adult was sampled from each previously sampled housing unit. The individuals were selected using the Kisch's method. It is a procedure of simple sampling of one person from each dwelling. Under this method, one random number is generated for the expected number of persons in a dwelling within the 1 to 9 range. These numbers are assigned to the number assigned to the housing unit during the sampling and cannot be changed. An interviewer visiting a dwelling receives these numbers together with the address, generates the sampling frame in the dwelling and selects an individual using the random numbers. In this way, the selected individual can be uniquely identified.
The table with random numbers for one housing unit is as follows:

Expected number of adults in dwelling: c =	1	2	3	4	5	6	7	8	9
Random number i.e sampled person's number in the sampling frame	RN_1	RN_2	RN_3	RN_4	RN_5	RN_6	RN_7	RN_8	RN_9

where:

RN_c – random number for the expected number of adults in the housing unit from the 1–c range.
$c = 1, 2, \ldots, 9$

Let:
L_{hjl} – real number of adults in the l-th dwelling (sampled), the j-th census unit in the h-th stratum ($L_{hjl} = 1, 2, 3, 4, \ldots, 9$).

6. Estimation and evaluation of estimator variance

Let:
U – Finite population (adult population of the country),
$U = \{u_1, u_2, \ldots, u_{30000000}\}$,
$i = 1, 2, \ldots, 30,000,000$ – the number of an individual in the population,
U_d – non-empty subset of U separated for analysis, or, in special case, the whole population. In practice, an analysis of population subsets selected for analysis purposes after the survey execution is frequently required. Due to the broad scope of the matter, in this paper we shall only consider the situation where $U_d = U$,
$d = 1, 2, \ldots, D$,
D – number of subsets in the analysis
X', Y' – functions defined on U with values 1 or 0.
The estimated parameter is a fraction of units with distinguished property. It is the ratio of global values of functions X' and Y'.

$$R = \frac{X}{Y} \tag{5}$$

where:

$X = \sum X_i$

$$X_i = \left\{ \begin{array}{ll} 1 & \text{for } u_i \in U \text{ with distinguished property} \\ 0 & \text{for other units} \end{array} \right\}$$

$Y = \sum Y_i$

$$Y_i = \left\{ \begin{array}{ll} 1 & \text{for } u_i \in U \\ 0 & \text{for other units} \end{array} \right\}$$

X_i, Y_i – values of functions X', Y' for units $u_i \in U$.

Let:
s^* – the total sample sampled from the population U,
s – the total sample after realization from the population U ($s \subset s^* \subset U$)

If $s^* = s$ then estimator R is the following statistic:

$$r = \frac{x}{y} \tag{6}$$

where:

$$x = \sum_h \sum_j \sum_l \sum_i \frac{M_{hj}}{\Pi_{hj}} \frac{L_{hjl}}{6} x_i \tag{7}$$

$$y = \sum_h \sum_j \sum_l \sum_i \frac{M_{hj}}{\Pi_{hj}} \frac{L_{hjl}}{6} y_i \tag{8}$$

x_i – values of X' for $u_i \in s$

$$x_i = \left\{ \begin{array}{ll} 1 & \text{for } u_i \in s \text{ with distinguished property} \\ 0 & \text{for other units} \end{array} \right\}$$

y_i – values of Y' for $u_i \in s$

$$y_i = \left\{ \begin{array}{ll} 1 & \text{for } u_i \in s \\ 0 & \text{for other units} \end{array} \right\}$$

Since the realized sample s is smaller than the selected sample s^*, the formulas 7 and 8 are transformed as follows:

$$x = \sum_h \sum_j \sum_l \sum_i W_{hjli} x_i = \sum_h \sum_j \sum_l \sum_i \frac{M_{hj}}{\Pi_{hj}} \frac{L_{hjl}}{m_{hj}} x_i \tag{9}$$

$$y = \sum_h \sum_j \sum_l \sum_i W_{hjli} y_i = \sum_h \sum_j \sum_l \sum_i \frac{M_{hj}}{\Pi_{hj}} \frac{L_{hjl}}{m_{hj}} y_i \tag{10}$$

where:
W_{hjli} – weight of the i^{th} person in the realized sample s, sampled from the l^{th} housing unit in the j^{th} census unit, the h^{th} stratum.

$$W_{hjli} = \frac{M_{hj}}{\Pi_{hj}} \frac{L_{hjl}}{m_{hj}} \tag{11}$$

m_{hj} – number of sampled housing units in which interviews were conducted
$m_{hj} \leq 6$ (6 housing units were sampled from each census unit).
Other symbols as previously.

Estimator r variance is as follows:

$$D^2(r) \approx \frac{D^2(x) + r^2 D^2(y) - 2COV(x,y)}{y^2} \tag{12}$$

In the case where the Hartley-Rao scheme was used for first-stage sampling, and simple sampling without replacement at the second-stage, the x_{HR} and the y_{HR} estimators variances: $D^2(x_{HR})$, $D^2(y_{HR})$ and $COV(x_{HR}, y_{HR})$, are given in literature (Bracha,1996). The approximate estimator for variance $D^2(x_{HR})$ was published by Sarndall (1992).

7. Adjusted estimator

The estimation method described in literature (section 6) has the following shortcomings:

- It does not account for non-response sufficiently;
- In the case of basic demographic distributions, it allows relatively large discrepancies between the survey results and the ongoing estimates of the Central Statistical Office. The main reasons for discrepancies include random errors, non-responses and incomplete sample execution.

Therefore, an adjusted estimator r_A was used:

$$r_A = \frac{x_A}{y_A} = \frac{\sum\limits_i w_i x_i}{\sum\limits_i w_i y_i} \tag{13}$$

where weights w_i, in addition to the varying probabilities of respondent selection (as in formula 11), also take into account the differences in response rates between the six locality categories and the population structure according to the Central Statistical Office by gender, age group (five age groups are distinguished) and place of residence (urban / rural).

Step 1

Weights for second- stage units, i.e. housing units, were determined:

$$W1_{hj} = \frac{M_{hj}}{6\Pi_{hj}} \tag{14}$$

Let us also remind that:
M_{hj} – number of housing units in the jth census unit in the hth stratum; Π_{hj} – first-order probabilities calculated using formula 2.
6 is the number of housing units sampled from the j^{th} census unit.

Step 2

Determining response rates for the six locality categories. The following categorization of localities has been adopted:
$k = 1$ – Warsaw (the capital)
$k = 2$ – other cities with more than 500,000 inhabitants
$k = 3$ – towns with 100,000 to 500,000 inhabitants
$k = 4$ – towns with 20,000 to 100,000 inhabitants
$k = 5$ – towns with up to 20,000 inhabitants
$k = 6$ – villages
The response rate for survey E_k for the k^{th} locality category was determined using the following formula:

$$E_k = \frac{B_k}{B_k + D_k} \tag{15}$$

where:

B_k – number of housing units surveyed, i.e. dwellings in which a respondent was successfully sampled and interviewed.

D_k – number of housing units in the k^{th} locality category, in which interviews were not conducted for different reasons.
D_k does not include sampling frame errors, i.e. non-existent housing units, unoccupied dwellings and housing units used for non-residential purposes, e.g. offices, warehouses etc. B_k and D_k values were determined using the following formulas:

$$B_k = \sum_h \sum_j mb_{khj} W1_{hj} \tag{16}$$

$$D_k = \sum_h \sum_j md_{khj} W1_{hj} \tag{17}$$

where:

mb_{khj} – number of housing units surveyed in the j^{th} census unit of the h^{th}stratum and the k^{th} locality category;

md_{khj} – number of housing units not surveyed in the j^{th} census unit of the h^{th} stratum and the k^{th} locality category, excluding housing units included in the sampling frame due to errors.

The calculated E_k values allowed to determine the W2 weight taking into account the response rate in the k^{th} locality category.

$$W2_{khj} = \frac{W1_{hj}}{E_k} \tag{18}$$

Step 3

During the third step, differences in the demographic distributions between the sample and the official Central Statistical Office data were eliminated. For this purpose, the sample was divided *ex post* into 20 groups distinguished by the place of residence (rural/urban), gender, and subsequently into 5 age categories. Based on data from the sample, g_p values were determined, i.e. the number of persons by gender, place of residence and age category, using the following formula:

$$g_p = \sum_k \sum_h \sum_j \sum_l \sum_i W2_{khj} L_{hjl} z_{khjli} \tag{19}$$

where:
p – group number after the sample division – by place of residence, gender and age category
$p = 1, 2, 3, \ldots, 20$

$$z_{khli} = \left\{ \begin{array}{ll} 1 & \text{for for persons included to the group p} \\ 0 & \text{for other persons} \end{array} \right\} \tag{20}$$

Subsequently, the following values were determined for each of p demographic groups:
G_p – the number of persons in the population belonging to the p group – based on the Central Statistical Office data.

Step 4

Ultimately, weight W_{pkhjli} of the i^{th} respondent, sampled from the I^{th} housing

unit in the j^{th} census unit,the k^{th} locality category and the p^{th} demographic group was as follows:

$$W_{pkhjli} = W2_{khj}\frac{G_p}{g_p}L_{hjl} \tag{21}$$

For the sake of simplification, let us adopt the following formula:

$$W_i = W_{pkhjli} \tag{22}$$

$$w_i = na\frac{W_i}{\sum\limits_i W_i} \tag{23}$$

where:
W_i – "big weight" of the i^{th} respondent in realised sample
w_i – "litttle weight" of the i^{th} respondent in realised sample
na – size of realized sample of persons, i.e. the number of interviews conducted.

The final adjusted ratio estimator, after simplifying the symbols, shall be expressed with the following formula:

$$r_A = \frac{\sum\limits_i W_i x_i}{\sum\limits_i W_i y_i} = \frac{\sum\limits_i w_i x_i}{\sum\limits_i w_i y_i} \tag{24}$$

8. Evaluation of r_A estimator variance

The complex nature of the r_A estimator resulting from its adjustment and the complex sampling scheme caused problems with finding a formula for its variance. Therefore, in order to evaluate the variance, the Jacknife method (Wolter 1985) was used, in the version appropriate for a compound estimator of random variables ratio, along with a stratified sampling scheme.
Under this method, "pseudo-values" of the parameter being estimated are calculated based on the sample less one unit. If the first-stage units were stratified before sampling, the estimator "pseudo-values" are estimated by eliminating the units from the first stratum, the second stratum etc. until all possibilities are exhausted in the last stratum. Each time, after reducing the sample by one first-stage unit (census unit), estimator "pseudo-values" are determined.
Variance estimator:

$$d^2(r_A) = \sum_h \left[n_h - \frac{1}{n_h} \right] \sum_j [r_{hj} - r_h]^2 \tag{25}$$

where:

r_{hj} – evaluation of parameter R based on the sample reduced by the j^{th} census unit of the h^{th} stratum;
r_h – average evaluation of parameter R based on the r_{hj} evaluations from the h^{th} stratum;
$r_h = \frac{1}{n_h} \sum_j r_{hj} \qquad j = 1, 2, \ldots, n_h;$
n_h – number of first-stage units in the h^{th} stratum, in which the survey was executed.
$n_h \leq n_h^*$

9. Conclusion

The presented three-stage sampling scheme and estimation methods were used in practice in the years 1990-2003. The response rate was falling gradually from over 83% to just over 70%.
It became possible to use the sampling frame which allows two-level sampling. The change of the sampling scheme did not, however, cause an increase of the response rate and simplified methods of estimator variance evaluations still had to be used.
As the response rates are falling, a question arises about the "threshold" response rate below which the parameters are no longer correctly estimated and not close enough to the actual ones in the population. What other methods could be used instead of the ones based on adjusted estimators and estimations of their variance using simplified methods based on resampling (e.g. bootstrap, jacknife etc.)?

REFERENCES

[1] Bracha Cz. Wykorzystanie informacji o cechach dodatkowych w badaniach reprezentacyjnych. ZBSE GUS i PAN (1987) Warsaw, Poland.

[2] Bracha Cz. Teoretyczne podstawy metody reprezentacyjnej. Wydawnictwo Naukowe PWN (1996) Warsaw, Poland.

[3] Bracha Cz. Metoda reprezentacyjna w badaniu opinii publicznej i marketingu. Efekt (1998) Warsaw, Poland.

[4] Bracha Cz., Lednicki B., Wieczorkowski R. Wykorzystanie złożonych metod estymacji do dezagregacji danych z badan aktywności ekonomicznej ludności w roku 2003. *ZBSE. GUS (Central Statistical Office)* (2004) Warsaw, Poland.

[5] Dabała W. How to estimate population parameters using partially realized random sample? XXV International Seminar on Stability Problems for Stochastic Models, Maiori/Salerno, Italy. *Series of the Journal of Mathematical Sciences* (2005) (Kluwer-Plenum, New York-London).

[6] Efron B., Tibshirani R.J. An Introduction to the Bootstrap Monographs on Statistics and Applied Probability, 57. Chapman and Hall, New York (USA) - London (Great Britain).

[7] Hartley H.O., Rao J.N.K. Sampling with unequal probabilities and without replacement. *AMS* **33** (1962).

[8] Hunsen M.H., Hurwitz W.N., Madow W.G. Sample Survey Methods and Theory. John Wiley & Sons, Inc. (1993) New York-Chichester-Brisbane-Toronto-Singapore.

[9] Madow W.G. On the theory of systematic sampling II. *AMS* **20** (1949).

[10] Sarndall C.E., Swensson B, Wretman J. Model Assisted Survey Sampling. (1992) Springer Verlag.

[11] Tryfos P. Sampling Methods for Applied Research. Text and Cases. (1996) John Wiley & Sons, Inc. New York, USA.

[12] Volter K.M. Introduction to Variance Estimation. (1985) Springer Verlag.

Wiesława Dabała
Public Opinion Research Centre, Warsaw, Poland
115, Universitetska Str.
e-mail: w.dabala@cbos.pl

Bronisław Lednicki
Central Statistical Office, Warsaw, Poland