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 269, 2012

*Dorota Bartosińska**

STATISTICAL INFERENCE FROM COMPLEX SAMPLE WITH SAS ON THE EXAMPLE OF HOUSEHOLD BUDGET SURVEYS

Abstract. Many sample surveys are not based on simple or unrestricted random samples, but usually on complex samples with stratification, clustering, unequal inclusion probabilities and multistage sampling. To estimate a parameter, all individual data from complex sample must be weighted by weights connected with sample selection scheme and statification and adjusted for nonresponse and noncoverage errors. Standard statistical computer software is correct for statistical inference from unrestricted random sample, but not from complex sample. The aim of this paper is to present utility of SAS software to statistical inference from complex sample. Data from Household Budget Survey 2008 were used in examples.

Key words: complex sample, household budget survey, statistical inference

I. INTRODUCTION

Results of sample surveys are used in different studies, but not always statistical inference is correct. First, parameter estimates are sometimes used as they would be parameters from the full-scale surveys (censuses) and no sampling errors are not taken into account. Second, usually used standard statistical computer software is good for statistical inference from unrestricted random sample. If you use standard statistical computer software for statistical inference from complex samples, parameter and their variance estimates can be biased, so both parameter estimation and hypothesis testing are incorrect. Third, we must remember that non-sampling errors usually occur both in surveys and censuses and they affect the accuracy and quality of statistical data.

The aim of this paper is to present utility of SAS software to statistical inference from complex sample surveys. SAS software (version 9.2) has six procedures to complex samples, (SAS, 2011). SURVEYSELECT procedure is used to select complex sample. SURVEYMEANS and SURVEYFREQ procedures are used to statistical inference from complex sample for quantitative and qualitative variable, respectively. SURVEYREG and SURVEYLOGISTIC

* Ph.D., Institute of Statistics and Demography, Warsaw School of Economics

procedures are used to regression analysis and logistic regression analysis, respectively, from complex sample. SURVEYPHREG procedure is used to event history analysis for from complex sample. These five above mentioned SAS survey procedures make statistical inference for the whole population parameters, taking weights, stata, clusters, and also for domain parameters. In order to estimate the variance of parameter estimates obtained from complex sample surveys involving complex sampling and complex estimation procedures, it is necessary to use one of indirect methods of variance estimation. In order to estimate variance of parameter estimators, SAS software (version 9.1) used only the Taylor linearization, but SAS software (version 9.2) uses the Taylor linearization on default, and also optionally: balanced repeated replication and jackknife methods. In this paper SURVEYSELECT, SURVEYMEANS and SURVEYFREQ procedures were described.

Examples of application SAS software to statistical inference from complex samples were done on the data from Household Budget Survey, conducted by the Central Statistical Office in Poland in 2008. The population was about 13 million households in Poland, sample about 37 thousand households (0.3%). Sample selection scheme was two-stage, stratified, with different inclusion probabilities for the first stage. Some results were presented and interpreted: parameter estimates, estimated coefficient of variation of parameter estimates and estimated design effects. Coefficient of variation of parameter estimator is ratio of standard error of parameter estimator and parameter estimator. Standard error of parameter estimator is root of estimator variance. Design effect is ratio of estimator variance under applied sample selection design and estimator variance under simple random sampling with the same sample size.

II. SAMPLING WITH SAS

SAS software has a special procedure to select complex sample: SURVEYSELECT procedure. It provides a variety of methods for selecting probability-based random samples. The procedures can select a simple random sample or can sample according to a complex multistage sample design that includes stratification, clustering and unequal inclusion probabilities. The syntax of this procedure is:

PROC SURVEYSELECT options;

STRATA stratification variables;

CONTROL ordering variable in systematic and sequential sampling;

SIZE variable containing the size measure in sampling with probability proportional to size;

ID variable to copy from the input data set to output data set;

The PROC SURVEYSELECT options are: sample selection method, sample size or sample rate and other sample designs parameters. Sample selection methods in SAS software are: **SRS** (Simple Random Sampling), **URS** (Unrestricted Random Sampling), **SYS** (Systematic Random Sampling), **SEQ** (Sequential Random Sampling), **PPS** (Probability Proportional to Size without Replacement), **PPS_WR** (Probability Proportional to Size with Replacement), **PPS_SYS** (Probability Proportional to Size Systematic Sampling), **PPS_SEQ** or **CHROMY** (Probability Proportional to Size Sequential Sampling), **PPS_BREWER** or **BREWER** (Probability Proportional to Size Brewer's Method), **PPS_MURTHY** or **MURTHY** (Probability Proportional to Size Murthy's Method Sampling), **PPS_SAMPFORD** or **SAMPFORD** (Probability Proportional to Size Sampford's Method Sampling).

The statements and options of SURVEYSELECT procedures were described in (Gołata, 2009 and SAS, 2011).

The output of SURVEYSELECT procedure has: output data file including selection probabilities and sampling weights. According to sample selection scheme and stratification, primary sampling weights are constructed before sampling and they are the inverse selection probability:

$$w_i = \frac{1}{p_i}. \tag{1}$$

For multistage sampling, the probability of selection is computed as the product of the selection probabilities for each stage of sampling. For example, for two stage sampling, the probability of selecting secondary unit *j* in primary unit *i* is:

$$p_{ij} = p_i \cdot p_{j/i}. \tag{2}$$

The primary weights are usually adjusted for nonresponse and noncoverage error (Levy, Lemeshow, 2008).

Weights can be interpreted as the number of individuals in the target population represented by the *i*th sample unit. The sum of weights for all units in the sample is an estimator of number of units in population:

$$\hat{N} = \sum_{i=1}^{n} w_i. \tag{3}$$

III. STATISTICAL INFERENCE FOR QUANTITATIVE VARIABLE FROM COMPLEX SAMPLE WITH SAS

SURVEYMEANS procedure is used to statistical inference from complex sample for quantitative variable for parameters, such as population means, totals, ratio of two means or totals. Estimator of the population total X can be put into the following form:

$$\hat{x} = \sum_{h=1}^{H} \sum_{j=1}^{m_{hi}} \sum_{i=1}^{n_h} x_{hij} w_{hij} \; ; \tag{3}$$

where:

$h = 1, 2, \ldots, H$ is the stratum index,

$i = 1, 2, ..., n_h$ is the cluster index within stratum h,

$j = 1, 2, ..., m_{hi}$ is the unit index within cluster i of stratum h,

x_{hij} – the observed values of the analysis variables X for unit j in cluster i of stratum h,

w_{hij} – the sampling weight for unit j in cluster i of stratum h.

Estimator of the population mean is the ratio estimator of estimator of population total and estimator of number of population units), given by:

$$\hat{\bar{x}} = \frac{\hat{x}}{\hat{N}} \; ; \tag{4}$$

(Bracha, 1996; Levy, Lemeshow, 2008).

The syntax of this procedure is:

PROC SURVEYMEANS options statistics-keywords;
BY grouping variable;
CLASS qualitative variables;
CLUSTER variable of the first stage of sampling;
DOMAIN domain variable;
RATIO numerator/denominator;
STRATA stratification variable;
VAR analysis variable;
WEIGHT weighting variable.

The statements and options of SURVEYMEANS procedures were described in (Gołata, 2009 and SAS, 2011).

In this paper, variances of parameter estimates were estimated in SAS using the Taylor linearization method. The SAS SURVEYMEANS procedure estimates the variance of mean as:

$$\hat{V}(\hat{\bar{x}}) = \sum_{h=1}^{H} \hat{V}_h(\hat{\bar{x}}) = \frac{n_h(1-f_h)}{n_h - 1} \sum_{i=1}^{n_h} (e_{hi.} - \bar{e}_{h..})^2; \quad (5)$$

where:

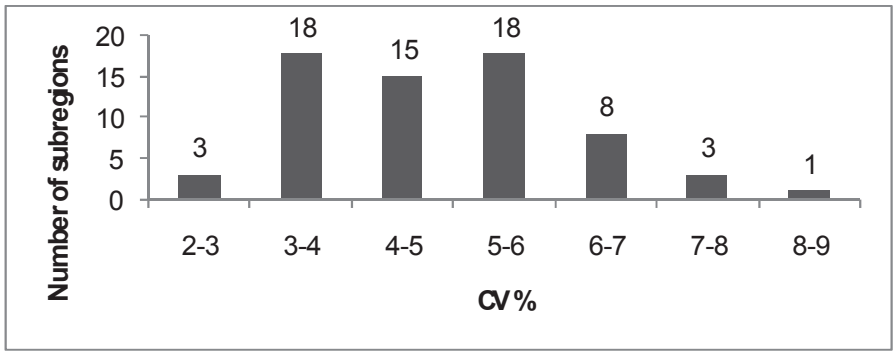
$$e_{hi.} = \frac{\sum_{j=1}^{m_{hi}} w_{hij}(x_{hij} - \hat{\bar{x}})}{w_{\_}},$$

$$\bar{e}_{h..} = \frac{\sum_{i=1}^{n_h} e_{hi.}}{n_h},$$

f_h – the sampling rate for stratum h, which is used in Taylor series variance estimation, is the fraction of first-stage units selected for the sample.

Mean expenditures in household in Poland, by 16 regions and by 66 subregions were estimated using SURVEYMEANS procedure. The results of mean estimation for Poland and by regions are presented in table 1. Estimated CVs by regions were between 1.8% and 4.8%. Estimated deff for regions ranged from 1.7 to 4.9. Estimated CVs of mean expenditures by subregions are presented on figure 1. Estimated CVs of means by subregions were between 2.1% and 9.0%, average 4.9%. Expenditures per capita in household were also estimated as ratio of expenditure and number of persons in household. Estimated expenditures per capita in household in Poland was 904,27 PLN with coefficient of variation 0,6%.

Figure 1. Estimated coefficients of variation of mean expenditures by subregions in Poland



Source: own calculations

Table 1. The results of mean expenditures estimation in Poland and by regions

Region	Mean	CV%	deff
Dolnośląskie	2 519.18	2.4	1.937
Kujawsko-pomorskie	2 362.41	2.7	2.693
Lubelskie	2 403.27	3.0	2.328
Lubuskie	2 540.51	2.9	1.882
Łódzkie	2 434.15	2.4	2.430
Małopolskie	2 668.83	2.3	2.673
Mazowieckie	3 044.13	1.9	2.633
Opolskie	2 750.70	3.0	1.743
Podkarpackie	2 436.48	2.2	2.135
Podlaskie	2 357.65	4.8	4.904
Pomorskie	2 725.01	2.6	2.171
Śląskie	2 413.99	1.8	2.506
Świętokrzyskie	2 283.64	3.1	2.287
Warmińsko-mazurskie	2 116.49	3.5	3.080
Wielkopolskie	2 576.81	2.1	2.271
Zachodniopomorskie	2 394.98	2.8	2.592
Poland	2 558.46	0.7	2.459

Source: own calculations.

IV. STATISTICAL INFERENCE FOR QUALITATIVE VARIABLE FROM COMPLEX SAMPLE WITH SAS

SURVEYFREQ procedure is used to statistical inference from complex sample for qualitative variable for parameters, such as population totals and proportions. Estimator of the population number and proportion of interest category is given by formulas (3) and (4) respectively, if analysed variable is given by:

$$y_i = \begin{cases} 1 \\ 0 \end{cases}; \tag{6}$$

where 1 – interest category, 0 – other categories.

The syntax of the procedure is:

PROC SURVEYFREQ options;
BY grouping variable;
CLUSTER variable of the first stage of sampling;
STRATA stratification variable;
TABLES qualitative variables;
WEIGHT weighting variable;

This procedure also is used to contingency analysis in the case of complex sample, if we write in TABLES instruction two variables with "*". The statements and options of SURVEYFREQ procedures were described in (SAS, 2011).

Household structure in Poland, by 16 regions and by 66 subregions were estimated using SURVEYFREQ procedure. The results of household structure for Poland are presented in table 2. About 25% household in Poland is one-person-household. Estimated CVs for Poland were between 1.0% and 2.4%, and deff between 1.011 to 2.156. Estimated CVs of household structure by regions were between 2.3% and 18.8%, average 6.6%. Estimated deff for regions ranged from 0.843 to 3.196, average 1.363. Estimated CVs of household structure by subregions were between 3.2% and 54.7%, average 12.8%.

Table 2. The results of household structure estimation in Poland

Number of person in household	Proportion	CV%	deff
1	24.8	1.3	2.156
2	23.2	1.0	1.011
3	19.9	1.1	1.121
4	18.0	1.2	1.252
5	8.1	1.9	1.246
6+	5.9	2.4	1.351

Source: own calculations.

V. CONCLUSIONS

1) There were obtained precise estimates of the parameters for Poland and by regions (CVs up to 5% for mean expenditures, but up to 19% for the household structure).

2) Parameters estimates for subregions are less precise (CVs for mean expenditures up to 9%, for the household structure up to 55%). There would be useful methods of small area estimation.

3) The design effect was up to 5 for estimates for Poland and by region. It means that, the estimated variances in applied complex sampling are up to 5 times greater than in simple random sampling. So computer statistical software using simple or unrestricted random samples would affect erroneous statistical inference from complex samples.

4) In statistical inference from complex sample we should take into account: sampling selection scheme, stratification and weights, and use proper statistical software and procedures.

5) In order to infer from complex sample surveys we have to consider both sampling and non-sampling errors that allows to be close to the true value of parameters.

REFERENCES

Bracha C. (1996), *Teoretyczne podstawy metody reprezentacyjnej*, Wydawnictwo Naukowe PWN, Warszawa.

Gołata E. (2009), *Wybrane zagadnienia metody reprezentacyjnej*, W: Frątczak E. (Editor), *Wielowymiarowa analiza statystyczna. Teoria – przykłady zastosowań z systemem SAS*, Warsaw School of Economics, Warsaw, chapter 3, pp. 83-116.

Levy P.S., Lemeshow S. (2008), *Sampling of population, methods and applications*, Wiley.

SAS. (2011), http://support.sas.com/documentation/onlinedoc/91pdf/index.html.

Dorota Bartosińska

WNIOSKOWANIE STATYSTYCZNE Z PRÓB ZŁOŻONYCH Z SAS NA PRZYKŁADZIE BADANIA BUDŻETÓW GOSPODARSTW DOMOWYCH

Wiele badań reprezentacyjnych nie opiera się na prostych próbach losowych, a zwykle na próbach złożonych: z nierównymi prawdopodobieństwami wyboru jednostek, z warstwowaniem, z zespołami i wieloma stopniami losowania. Aby uzyskać oszacowania parametrów, wszystkie dane indywidualne są ważone nie tylko wagami wynikającymi z zastosowanego schematu losowania próby i warstwowania przed wylosowaniem próby, lecz także ze względu na braki odpowiedzi i błędy pokrycia. Standardowe statystyczne pakiety komputerowe są odpowiednie dla prób prostych niezależnych. W przypadku prób złożonych dają one obciążone wyniki i zniekształcają wariancje estymatorów parametrów, czyli wnioskowanie statystyczne o populacji jest niepoprawne. Celem referatu jest prezentacja możliwości wykorzystania oprogramowania SAS w metodzie reprezentacyjnej. W przykładach wykorzystano dane z Badania Budżetów Gospodarstw Domowych 2008.