STATISTICS IN TRANSITION-new series, October 2010
Vol. 11, No. 2, pp. 231—252

DEVELOPMENT OF STANDARD ERROR ESTIMATION METHODS IN COMPLEX HOUSEHOLD SAMPLE SURVEYS IN POLAND

Jan Kordos[1], Agnieszka Zięba[2]

ABSTRACT

The authors begin with a general description of estimation methods of standard error and confidence interval from complex household sample surveys in Poland. The following estimation methods have been applied in resent years: (i) the interpenetrating sub-samples, (ii) the Taylor series linearization, (iii) the jackknife, (iv) the balanced repeated replication, and (v) the bootstrap methods. A short development of each method is presented and its application in the Polish household sample surveys described. At the end some concluding remarks are given.

Key words: interpenetrating samples, Taylor series linearization, random groups, balanced repeated replication, jackknife, bootstrap, complex sample surveys.

1. Introduction

The Polish household sample surveys, such as the Household Budget Survey (HBS), the Labour Force Survey (LFS) and the EU Statistics of Income and Living Conditions Survey (EU-SILC) are complex sample surveys. These surveys typically use some of the following sampling techniques: sampling without replacement from a finite population, systematic sampling, stratification, clustering, unequal probabilities of selection, multi-stage or multi-phase sampling. As a consequence, the values of the variables of interest in a complex survey sample are neither independent nor identically distributed. In addition, survey processing, aimed at improving the quality and usability of survey data, reducing the bias of the estimates, etc. further increase the complexity of the survey data. Weighted analyses are necessary for unbiased (or nearly unbiased) estimates of population parameters. Standard error or confidence interval estimation depends upon the sampling plan specifics and requires approximate methods.

[1] Warsaw School of Economics.
[2] Warsaw School of Economics.

It is well known that sampling variance is an important measure of the quality of estimates of finite population parameters (totals, means, quantiles, ratios, etc.). It measures the amount of *sampling error* in the estimate due to observing a sample instead of the whole population. Estimates of sampling variance are needed to produce the *coefficients of variation* (cv) that are disseminated along with the survey estimates and to construct confidence intervals for finite population parameters of interest. Sampling variance is also used as the variability measure for inferences about super-population models when the *design-based approach* is recommended, that is when the sample design is *informative* (Binder and Roberts 2001, 2003, Rao, 2003).

Estimation of the sampling variance can become very complicated due to the complex sample design, use of non-linear estimators, impact of survey processing etc. Standard error estimation for estimators based on complex sample survey data must recognize the following factors:

a. most estimators are non-linear (a ratio of linear estimators is common);
b. estimators are weighted;
c. the sampling plan will generally have used stratification prior to first-stage sampling (and perhaps also at subsequent sampling stages); and
d. elements in the sample will generally not be statistically independent owing to multistage cluster sampling.

In almost all cases, it is not possible to obtain a closed-form algebraic expression for the estimated standard error or confidence intervals. Thus, the research literature on standard error estimation for complex sample survey data contains several approximate methods from which sample survey data analysts can choose [Brick et al, 2000; Eurostat, 2009; Kish and Frankel, 1974; Kovar et al, 1988; McCarthy and Snowden, 1985; Osier, 2006; Shao, 2003; Shao and Tu, 1995; Sitter, 1992; Tibshirani, 1985; Verma, 2005, 2010; Verma and Betti, 2005, Wolter, 1985].

In Poland the following estimation methods for standard error estimators from complex sample surveys have been used:

(i) the interpenetrating sub-samples (random groups),
(ii) the Taylor series linearization,
(iii) the jackknife replication techniques,
(iv) the balanced repeated replication (BRR),
(v) the bootstrap methods.

As stressed above, most household surveys are based on complex sample designs applied to very large populations. The primary sampling units (PSUs) are generally selected using probability proportional to size (PPS) without replacement sampling, making the concept of "sampling fraction" more complex.

However, the number of PSUs is often large and the PSU sampling fraction in each stratum is fairly small, giving a value close to 1.0 for all first-stage *fpc* terms. Thus, a common approximation in the analysis of complex sample survey data is one where the PSUs have been sampled with replacement in each stratum. If this approximation is made in the presence of some strata with large first-stage

sampling fractions, the variance will be overestimated to some extent. Such overestimation is often accepted in view of the complexity of standard error estimation without the approximation.

There are two basic approaches to variance estimation: i) an analytical approach using the *linearization* method, and ii) *resampling* and *replication* methods (jackknifing, balanced repeated replication, bootstrapping). The description of these methods can be found in many books on survey sampling, see for example (Osier, 2006; Shao, 2003; Shao and Tu, 1995; Tibshirani, 1985; Wolter, 1985; Lohr ,1999). The increase in computing power has made the use of the resampling techniques feasible for large survey samples. These methods are also relatively easy to implement because, regardless of the point estimator, a resampling method always uses the same procedure, replicated many times, while the linearization approach requires a development of a new formula for every estimator and weight adjustment and usually still requires additional simplifying assumptions (Binder, Kovacevic, and Roberts, 2004).

In Poland, we started applying different methods of standard error estimation in complex household sample surveys nearly forty years ago. Some of these methods were generally described in different papers mainly in Polish (GUS (1986 ,2009, 2010ab; Kordos, 1985, 1996, Kordos and Szarkowski, 1974; Kordos et al., 2002; Lednicki, 1982, 2011).

Short descriptions of standard error estimation methods for complex household sample surveys in Poland are given below. For each estimation method of standard error for complex sample surveys in Poland, a general development of the method is presented and application of the method for particular survey in Poland described.

2. Standard error estimation methods for complex sample surveys in Poland

Household sample surveys in Poland have a long tradition; however we focus our attention first on random household budget surveys which were started over fifty years ago (GUS, 1986; Kordos, 1996, 2001; Kordos et al., 2002; Lednicki 1982). A lot of attention was devoted to these surveys due to their special role in the analysis of the living conditions of the population. Various methods of surveys were experimented with and attempts were made to improve their methodology and organization. Here attention is paid to different methods of standard error estimation in complex household sample surveys.

At the beginning of 1970s we applied the interpenetrating sub-samples method for standard error estimation in the household budget surveys (Kordos and Szarkowski, 1974). Later our statisticians undertook some research in studying other methods for standard error estimation in complex household sample surveys, such as the Taylor linearisation, jackknife, balanced repeated replication, and bootstrapping.

2.1. The interpenetrating sub-samples

When two or more sub-samples are taken from the same population by the same sampling plan so that each sub-sample covers the population and provides estimators of the population parameters on application of the same sampling procedures, the sub-samples are known as *interpenetrating sub-samples* (Mahalanobis, 1946; Sukhatme & Sukhatme, 1970; Som, 1996). The sub-samples may or may not be drawn independently and there may be different levels of interpenetration corresponding to different stages in a multi-stage sample scheme.

This technique may be used to (Murthy, 1967; Som, 1996; Sukhatme, Sukhatme, 1970; Zasepa, 1972):

1) compute the sampling error from the first-stage units if these comprise one level of interpenetration (both by the standard method and also by a nonparametric test);
2) provide control in data collection and processing;
3) examine the factors causing variation, e.g. enumerators, field schedules, different methods of data collection and processing;
4) supply advanced estimates on the basis of one or more sub-samples when the total sample cannot be covered due to some emergency;
5) provide a basis of analytical studies by the method of fractile graphical analysis.

This technique has been used in a number of sample surveys in India, including the Indian National Sample Survey, in Peru, Zimbabwe, the Philippines and the USA [Som, 1996].

In Poland, the interpenetrating sub-samples were used in the HBS and the LFS for :

a) standard error estimation,
b) providing control in data processing,
c) supplying advanced estimates on the basis of one or more sub-samples when the total sample cannot be covered for different reasons, and
d) modular topics.

We used this method for standard error estimation in the HBS from 1970 till 1999, and the LFS from 1992 till 1999 (Kordos & Szarkowski, 1974; Szarkowski and Witkowski, 1994; Zasępa, 1972, 1993). . However, we realized some shortcomings of this method and some critical remarks were presented in the paper by Kordos and Szarkowski (1974).

The general idea of interpenetrating sub-samples is given below. For estimating parameter θ of population, we select from the population s sub-samples by the same sampling plan so that each sub-sample covers the population and provides estimators of the population parameters on application of the same sampling procedures. For each sub-sample the parameter θ is estimated, then we

get $\hat{\theta}_{1_1},\hat{\theta}_{2,},...,\hat{\theta}_s$, where $\hat{\theta}_h$ stands for an estimate of parameter θ from the sub-sample h (h = 1, 2, …, s). Then, the parameter θ is estimated from the formula:

$$\hat{\theta} = \frac{1}{s}\sum_{h=1}^{s}\hat{\theta}_i \tag{1}$$

It is possible to prove that if estimators $\hat{\theta}_{1_1},\hat{\theta}_{2,},...,\hat{\theta}_s$ are independent with the same expected values, then

$$s^2(\hat{\theta}) = \frac{1}{s(s-1)}\sum_{h=1}^{s}(\hat{\theta}_h - \hat{\theta})^2 \tag{2}$$

is unbiased estimator of variance of statistics given by (1).

As mentioned in the paper (Kordos, Szarkowski, 1974) there is also another advantage of using interpenetrating sub-samples. Let $\hat{\theta}_{\min,s_i}$ stands for the smallest, and $\hat{\theta}_{\max,s_i}$ for the largest of observed values of $\hat{\theta}_{1_1},\hat{\theta}_{2,},...,\hat{\theta}_s$ i.e.

$$\hat{\theta}_{\min,s,...,} = \min\{\hat{\theta}_1,...,\hat{\theta}_s\}$$
$$\hat{\theta}_{\max,s} = \max\{\hat{\theta}_1,...,\hat{\theta}_s\}.$$

Then, it is possible to show, that if estimator (1) is unbiased and symmetrical (what is fulfilled when sizes of samples are sufficiently large), than probability that the interval $\{\hat{\theta}_{\min,s_i}, \hat{\theta}_{\max,s_i}\}$ covers estimated parameter θ is equal:

$$P = 1-(\frac{1}{2})^{s-1} \tag{3}$$

It means that the interval $\{\hat{\theta}_{\min,s_i}, \hat{\theta}_{\max,s_i}\}$ may be interpreted as some kind of confidence interval for parameter θ, and level of confidence is given by (3). It is worth to add that for s = 4, P = 0.875, for s= 5, P = 0.94, for s = 8, P = 0.992, and for s = 10, P= 0.998.

The sub-samples may also be distinguished by differences in the survey procedures or processing features. These are sometimes known as *replicated samples***.** We also called them later *random groups*.

As stressed above, some of our statisticians from the Central Statistical Office of Poland (GUS) tried to apply other methods of standard error estimation, such as the Taylor linearization technique, the jackknife replication technique, the balanced repeated replication or some bootstrap methods.

2.2. The linearization method for variance estimation

The Taylor linearization technique for standard error estimation was introduced, after some experiments, in the 4th quarter of 1999 in the Polish Labour Force Surveys. At that time we had no access to ready-computer programme, and our sampling statistician, Andrzej Szarkowski, prepared such a computer-programme himself which was used on a larger scale in GUS for the Labour Force Survey from 1999 until to 2002. Since 2003 the bootstrap method has been used till present time (Popiński, 2006).

Basic idea of linearization method consists of deriving from a complex non-linear statistic a linear statistic which has the same asymptotic variance (Osier, 2006).

$$V(\hat{\theta}) \approx V(\hat{T}) = V(\sum_{i=1}^{n} w_i T_i) \tag{4}$$

where:

$\hat{\theta}$ – complex non-linear statistic

n – the sample size

w_i – sample weight of item i

Ti – linearized variable (variable whose expression depends on $\hat{\theta}$).

We would like to mention here some of our Polish experiments in case of the complex structure of the *Laeken* indicators (Zięba, Kordos, 2010). The asymptotic variance of the estimator is the variance of its linearization (Niemiro and Wieczorkowski, , 2005):

$$V(\hat{T}) = \sum_{h=1}^{L} \frac{a_h}{a_h - 1} \sum_{c=1}^{a_h} \left(\sum_{i=1}^{n_{hc}} w_{hci} T_{hci} - \frac{\sum_{c=1}^{a_h} \sum_{i=1}^{n_{hc}} w_{hci} T_{hci}}{a_h} \right)^2 \tag{5}$$

where:

w_{hci} – the survey weight attached to i^{th} sample unit (individual) in the c^{th} cluster (PSU) of the h^{th} stratum

T_{hci} – corresponding linearized variable

$\hat{T} = \sum w_{hci} T_{hci}$ – the estimator of target parameter θ

h – number of strata *(h=1, 2,...,L)*

a_h – number of clusters (PSUs) in stratum h

n_{hc} – the sample size (number of individuals) in the c^{th} cluster (PSU) of the h^{th} stratum.

In our experiments different linearization framework was applied. Estimation of the *At-risk-of-poverty rate after social transfers* and *Relative median poverty risk gap* makes use of nonparametric techniques of quantile estimation. (Zięba, Kordos, 2010). The asymptotic variances of these estimators involve the density (*f*) of the underlying probability distribution and standard kernel estimators were applied. Estimation of *S80/S20 income quintile share ratio* is closely related to the theory of M-estimators and GINI can be estimated by a U-statistic (Niemiro, Wieczorkowki, 2005).

Summing up our experience with the Taylor linearization estimation, we would like to stress that the method can be applied in general sampling designs, theory is well developed, and now software is available. However, there are also some cons: (i) finding partial derivatives may be difficult; (ii) different method is needed for each statistic; (iii) the function of interest may not be expressed a smooth function of population totals or means; (iv) accuracy of the linearization approximation.

2.3. The Jackknife Replication Technique

The jackknife replication technique was used in Poland experimentally with other estimation methods of standard errors for complex sample surveys, such as the Taylor linearisation, balanced repeated replication, and some bootstrap methods. From our experiments and from international literature, we have drawn conclusions, that the jackknife method is not as efficient as the bootstrap method used by us (Zięba, Kordos, 2010).

The jackknife, originally introduced as a method of bias estimation (Quenouille, 1949) and subsequently proposed for variance estimation (Tukey, 958), involves the systematic deletion of groups of units at a time, the re-computation of the statistic with each group deleted in turn, and then the combination of all these recalculated statistics. The simplest jackknife entails the deletion of single observations, but this delete-one jackknife is inconsistent for non-smooth estimators, such as the median and other estimators based on quantiles (Efron, 1979). Shao and Wu (1989) and Shao and Tu (1995) have shown that the inconsistency can be repaired by deleting groups of observations. Rao and Shao (1992) describe a consistent version of the delete-one jackknife variance estimator using a particular hot deck imputation mechanism to account for non-response.

Jackknife, and bootstrap are very similar, but bootstrap overshadows the others for it is a more thorough procedure in the sense that it draws many more sub-samples than the others. Through simulations, it is found that the bootstrap technique provides less biased and more consistent results than the Jackknife method does.

The jackknife is a less general technique than the bootstrap, and explores the sample variation differently. However the jackknife is easier to apply to complex

sampling schemes, such as multi-stage sampling with varying sampling weights, than the bootstrap.

The jackknife and bootstrap may in many situations yield similar results. But when used to estimate the standard error of a statistic, bootstrap gives slightly different results when repeated on the same data, whereas the jackknife gives exactly the same result each time (assuming the subsets to be removed are the same).

The jackknife works well only for linear statistics (e.g., mean). It fails to give accurate estimation for non-smooth (e.g., median) and nonlinear (e.g., correlation coefficient) cases.

Thus improvements to this technique were developed.

More information on comparisons of jackknife, linearization and bootstrap is in our last publication (Zięba, Kordos, 2010).

2.4. The balanced repeated replication (BRR)

Balanced half-sampling (McCarthy, 1969) is the simplest form of balanced repeated replication. It was originally developed for stratified multistage designs with two primary sampling units drawn with replacement in the first stage. Two main generalizations to surveys with more than $n_h = 2$ observations per stratum have been proposed. The first, investigated by Gurney and Jewett (1975), Gupta and Nigam (1987), Wu (1991) and Sitter (1993), uses orthogonal arrays, but requires a large number of replicates, making it impractical for many applications. The second generalization, a simpler more pragmatic approach, is to group the primary sampling units in each stratum into two groups, and to apply balanced repeated replication using the groups rather than individual units (Rao and Shao, 1996; Wolter, 1985). The balanced repeated replication variance estimator can be highly variable, and a solution to this suggested by Robert Fay of the US Bureau of the Census (Fay, 1989) is to use a milder reweighting scheme. Another solution (Rao and Shao, 1996, 1999) is to repeat the method over differently randomly selected groups to provide several estimates of variance, averaging of which will provide a more stable overall variance estimate.

With BRR, a half-sample replicate is formed by selecting one unit from each pair of PSUs and weighting the selected unit by 2 (so that it represents both units). Consequently, estimates from every PSU are in each replicate although half are weighted by zero. Though the number of half-samples can be quite large (2L, where L = the number of strata), the BRR method requires that only some of the possible half-sample replicates be created to obtain a variance estimator that reduces to the textbook (approximate sampling formula) variance estimator, which is an unbiased estimator of the true population variance. It is used a Hadamard matrix (a $k \times k$ orthogonal matrix consisting of 1's and -1's, where k is a multiple of 4) to specify the replicates, choosing a value of k that is greater than L. To minimize the number of replicates, usually the smallest value of k possible is chosen. With a fully balanced design, each pair of sample units is assigned to a

unique row in the Hadamard matrix. Within pairs, each PSU is assigned to one panel, and this panel assignment is used in conjunction with each column value to assign replicate factors (Wolter, 1985, pp. 111-115). To reduce the number of replicates, surveys may assign more than one strata to the same row in the Hadamard matrix. This is known as *partial* balancing (Wolter, 1985, pp. 125-131) or as the *grouped* balanced half-sample method (Rao and Shao (1996) . With grouped half-sample replication, units in the L strata are divided into G groups with approximately L/G strata in each group. Units within groups are split into two panels, and half-samples are formed for each group. With grouped balanced half-samples, the replicate variance estimator is not equivalent to the textbook estimator. Although it is still an unbiased estimator for the true variance, its variance is larger than the corresponding fully balanced method because cross-product terms between replicates no longer cancel (Wolter, 1985, pp. 127). Rao and Shao (1996) prove that the grouped half-sample replicate estimator is inconsistent.. Consequently, replicate weighting cells are often collapsed (more than in the full survey data procedure), which can induce a positive bias in the variance estimates (Rao and Shao,1999)). The Modified Half Sample (MHS) replication method developed by Robert Fay (1989) addresses this problem of overly perturbed replicate weights and drastically reduced replicate sample sizes.

The method of balanced half-samples was originally developed for the case of a large number of strata and a sample composed of only two elements (or two PSUs) per stratum. The idea of using half-samples for variance estimation was introduced at the US Bureau of the Census around 1960 (Särndal et al, 1992; Walter, 1985).

The BRR procedure consists of three steps:

1. Forming balanced half-samples from the full sample;
2. Constructing the replicate weights to be used in calculating the estimate of the parameter of interest for the subsamples; and
3. Computing the estimates of variance for the parameter of interest.

Forming Random Groups

The random groups comprise all possible balanced half-samples of PSUs. The sample design must include two PSUs per stratum (although many actual samples using this method form psuedo strata including psuedo-PSUs that meet this requirement). Each half sample is formed by deleting one PSU from each stratum. All possible combinations of such half-samples comprise the set of replicates.

Constructing Replicate Weights

Typically, BRR is implemented using replicate weights. Weights are set to zero for elements in excluded PSUs, and corresponding adjustments made to the weights of elements in the remaining PSUs. One set of such weights is constructed for each replicate.

Fay's method uses a milder adjustment to each weight. In some instances this approach improves the estimates of statistics such as medians and percentiles.

Computing Estimates of Variance

Assume that A such groups have been constructed. Then, for each group (a = 1,... ,A), the target statistics $\hat{o}_{(a)}$ is calculated based on data from half-sample a. The point estimate for the statistic may be estimated as the average of these half-sample estimates, or based on the single overall sample. The variance estimate is given by

$$V(\hat{\theta}) = \frac{1}{A}\sum_{a=1}^{A}(\hat{\theta}_a - \bar{\hat{\theta}})^2 \tag{6}$$

When using Fay's method the formula becomes

$$V(\hat{\theta}) = \frac{1}{A(1-k)^2}\sum_{a=1}^{A}(\hat{\theta}_a - \bar{\hat{\theta}})^2 \tag{7}$$

The BRR method for the Polish Household Budget Survey

The balanced repeated replication technique has been used in Poland for the Household Budget Survey since 2000 till now (Kordos et al., 2002; Lednicki, 2010). The general idea is presented below.

The basic estimated parameters for HBS are monthly expenditure (income or consumption) per head or equivalent unit (Lednicki, 2010), i.e.

$$R = \frac{X}{Y} \tag{8}$$

Estimate of parameter R is

$$\hat{r} = \frac{\sum_h \hat{x}_h}{\sum \hat{y}_h} \tag{9}$$

In turn, the values in nominator and denominator of (9) are estimated as follows:

$$\begin{aligned} \hat{x}_h &= \sum_i \sum_j w_{hij} \cdot x_{hij} \\ \hat{y}_h &= \sum_i \sum_j w_{hij} \cdot y_{hij} \end{aligned} \tag{10}$$

where: x_{hij} is value of variable X in the j-th household and the i-th PSU in stratum h, y_{hij} is value of variable Y in the j-th household and the i-th PSU in stratum h, w_{hij} – weight of the j-th household and the i-th PSU in stratum h.

Weights w_{hij} are calculated taking into account selection probability of the household and next it is adjusted to ex post stratification. For that purpose data from population census are used on size of household, separately for urban and rural areas. For each of 12 categories of households the following coefficients are calculated:

$$M_k = \frac{G_K}{\hat{g}_k}, k = 1, 2, ..., 12) \tag{11}$$

where: G_k – the number of households of category k, $\hat{g}_k$ – the number of households of category k estimated from the sample.

Multiplier M_k is used for correction of primarily selection weights.

The method of balanced half-samples is used to estimate standard error of parameter ***r*** (Särndal et al., 1992) . Using this method, first calculated are values $x_{1h}, x_{2h}; y_{1h}, y_{2h}$, which are estimates of sum of values variables X and Y in stratum h for the first and second sample respectively. Next for *r* value calculated are balanced replicates r_α (α = 1, 2, …, A; A =128), according to the following formulas:

$$x_\alpha = 2 \cdot \sum_{h=1}^{L} [P_{\alpha h} \cdot x_{1h} + (1 - P_{\alpha h}) \cdot x_{2h}] \tag{12}$$

$$y_\alpha = 2 \cdot \sum_{h-1}^{L} [P_{\alpha h} \cdot y_{1h} + (1 - P_{\alpha h}) \cdot y_{2h}] \tag{13}$$

where: $P_{\alpha h}$ – element of **P** Hadamard matrix, in which elements on values –1 were replaced zeros. Elements of matrix **P** are 1 or 0 (are orthogonal); L – the number of strata (L = 96).

Using (12) and (13) r_a is calculated:

$$r_\alpha = \frac{x_\alpha}{y_\alpha} \tag{14}$$

Next variance V(r) and coefficient of variation CV(r) are estimated:

$$V(r) = \frac{1}{A} \sum_{a=1}^{A} (r_a - \hat{r})^2 \tag{15}$$

$$CV(r) = \frac{\sqrt{V(r)}}{r} \tag{16}$$

Each year for a large number of items are published their estimates, standard errors and relative standard errors (e.g. GUS, 2010b).

2.5. The bootstrap methods

The bootstrap is a method which has been increasingly used to estimate the standard error of estimates obtained from complex survey designs since the publication of the first research article (Efron, 1979). This method has been shown to work well for a wide range of estimators, including medians and quantiles, as well as smooth functions based on totals. In addition, the bootstrap can be less computer intensive than the jackknife method for surveys with a very large number of primary sampling units (PSUs), (Faucher, et al., 2003).

Bootstrap Method for Standard Error Estimation

Sampling statisticians started to study the use of bootstrapping for variance estimation in the mid eighties. A direct extension to surveys samples of the standard bootstrap method developed for i.i.d. samples is to apply the standard bootstrap independently in each stratum. This methodology is often referred to as the ***naïve bootstrap***. Because the naïve bootstrap variance estimator is inconsistent in the case of bounded stratum sample sizes, several ***modified bootstrap*** methods were proposed. The following bootstrap methods that were modified for survey samples are discussed in Shao and Tu (1996):

(i) The with-replacement bootstrap (McCarthy and Snowden, 1985),
(ii) the rescaling bootstrap (Rao and Wu, 1988, Rao, Wu and Yue, 1992),
(iii) the mirror-match bootstrap (Sitter, 1992), and
(iv) the without-replacement bootstrap (Gross, 1980, Chao and Lo, 1985, Bickel and Freedman, 1984, Sitter, 1992).

Modified Bootstrap

Rao and Wu (1988) proposed a bootstrap method for stratified multi-stage designs with WR sampling of PSUs that applied a scale adjustment directly to the survey data values. Rao, Wu and Yue (1992) presented a modification of the 1988 method where the scale adjustment is applied to the survey weights rather than to the data values. This modification increases the applicability of the method, from variance estimation for smooth statistics to the inclusion of non-smooth statistics as well. Here we describe the modified rescaling bootstrap method proposed by Rao, Wu and Yue (1992), and its connection with the McCarthy and Snowden's :

To estimate the variance of the estimator $\hat{\theta}$, the following steps (i) to (iv) are independently replicated *B* times, where B is quite large (typically, *B*=500).

(i) Independently in each stratum *h*, select a bootstrap sample by drawing a simple random sample of $n_h^{(b)}$ primary sampling units (PSUs) with replacement

from the n_h sample PSUs. Let $t_{hi}^{(b)}$ be the number of times that PSU *hi* is selected in the bootstrap sample *b*, *b*=1,2, …, *B*.

(ii) For each secondary sampling unit (SSU) *k* in PSU *hi*, calculate the initial bootstrap weight by rescaling its initial sampling weight:

$$w_{hik}^{(b)} = w_{hik}\{(1-\sqrt{\frac{n_h^{(b)}}{n_h-1}}) + \sqrt{\frac{n_h^{(b)}}{n_h-1}} \cdot \frac{n_h}{n_h^{(b)}} \cdot t_{hi}^{(b)}\}\ , \tag{17}$$

where w_{hik} is the initial sampling weight of the SSU hik, equal to the inverse of its selection probability, i.e. $w_{hik} = \frac{1}{\pi_{hik}}$.

(iii) To obtain the final bootstrap weight $fw_{hik}^{(b)}$, adjust the initial bootstrap weight $w_{hik}^{(b)}$ by *f* using all the same weight adjustments (e.g. non-response and calibration) that were applied to the initial sampling weight w_{hik} to produce the final survey weight fw_{hik}.

(iv) Calculate $\hat{\theta}^{(b)}$, the bootstrap replicate of estimator $\hat{\theta}$ by replacing the final survey weights fw_{hik} in the formula for $\hat{\theta}$.

The bootstrap variance estimator of $\hat{\theta}$ is then given by

$$v_{BS}(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}^{(b)} - \hat{\theta})^2, \text{ or} \tag{18a}$$

$$v^*_{BS}(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}^{(b)} - \hat{\theta}^*_{BS})^2, \text{ with } \hat{\theta}^*_{\mathrm{BS}} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{(b)}\ . \tag{18b}$$

The estimators (18a) and (18b) are Monte Carlo approximations of the bootstrap estimator of $V(\hat{\theta})$ given by $\hat{V}_{BS}(\hat{\theta}) = E_{BS}[\hat{\theta}^{(b)} - E_{BS}(\hat{\theta}^{(b)})]^2$, where E_{BS} denotes the expectation with respect to bootstrap sampling.

Both (18a) and (18b) are used in practice and they usually produce very similar values. However, the variance estimate (18a) is always larger than (18b), i.e. $v_{BS}(\hat{\theta}) \geq v^*_{BS}(\hat{\theta})$:

$$v_{BS}(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}^{(b)} - \hat{\theta}^*_{BS} + \hat{\theta}^*_{BS} - \hat{\theta})^2 = \frac{1}{B},\sum_{b=1}^{B}(\hat{\theta}^{(b)} - \hat{\theta}^*_{BS})^2 + (\hat{\theta}^*_{BS} - \hat{\theta})^2$$

$$= v^*_{BS}(\hat{\theta}) + (\hat{\theta}^*_{BS} - \hat{\theta})^2 \tag{19}$$

We see that the estimator (18a) includes a positive quantity $(\hat{\theta}^*_{BS} - \hat{\theta})^2$ which converges to zero for all consistent estimators $\hat{\theta}$.

Size of the bootstrap sample

If $n_h^{(b)} \leq n_h - 1$, then the bootstrap weights are never negative. When $n^{(b)}{}_h = n_h - 1$, the rescaling bootstrap reduces to McCarthy and Snowden's . Usually surveys use $n_h^{(b)} = n_h - 1$, mainly because it greatly simplifies the calculation of the bootstrap weights; the rescaling formula given in (17) becomes:

$$w_{hik}^{(b)} = w_{hik}\{\cdot\frac{n_h}{n_h - 1}\cdot t_{hi}^{(b)}\} \tag{20}$$

Note that if $t_{hi}^{(b)} = 0$, i.e. PSU hi is not selected to the bootstrap replicate b, its bootstrap weight (20) is zero. This formula is used for the McCarthy and Snowden's estimator.

At the Central Statistical Office of Poland (GUS), we use the McCarthy-Snowden bootstrap in the Labour Force Survey since 2003 (Popiński, 2006), and in the EU- SILC since 2005 till present time (GUS, 2008).

Bootstrap percentile confidence intervals

The percentile method does not assume that the sampling distribution is normal. Instead, the percentile method assumes that the distribution of the bootstrapped statistics approximates the true sampling distribution.

To estimate confidence intervals, we simply generate a large number of bootstrapped statistics and sort them in ascending order. The 95% confidence interval then can be estimated simply by selecting the bootstrapped statistics at the 2.5-th and 97.5-th percentiles.

1. Compute the statistic θ from your sample of values X.
2. Generate B bootstrapped samples, X_b^*, by randomly sampling X with replacement. Typically, B is a large number (e.g., > 1000).
3. For each bootstrapped sample X_b^*, compute θ_b^*.
4. Sort θ_b^* from the smallest to the largest value.
5. Letting $[\theta_1^* \leq \theta_2^* \leq \theta_3^* \leq,..., \theta_B^*]$, represent the ordered θ_b^* values, then the (100 ×α)% confidence interval for θ is $(\theta^*_{(l_o+1)}, \theta^*_{hi})$.
6. In a two-tailed test, H_0 (i.e., that $\hat{\theta} = \mu$) is rejected if $\mu \leq \theta^*_{(l_o+1)}$ or $\mu \geq \theta^*_{hi}$.

It means that one way to estimate confidence intervals from bootstrap samples is to take the α and 1 − α quantiles of the estimated values. These are called *bootstrap percentile intervals*. These can be contrasted with the asymptotic intervals derived from the maximum likelihood estimates plus or minus 1.96 standard errors. The intervals from the asymptotic theory are apparently too narrow (as well as being symmetric).

This simple bootstrap method is not the only way of making improved inferences over the asymptotic approach. Other bootstrap schemes are available, as are approaches based on likelihood or Bayesian considerations.

The interval between the 2.5% and 97.5% percentiles of the bootstrap distribution of a statistic is a 95% bootstrap percentile confidence interval for the corresponding parameter.

More accurate bootstrap confidence intervals: BCa and tilting

Any method for obtaining confidence intervals requires some conditions in order to produce exactly the intended confidence level. These conditions (for example, normality) are never exactly met in practice. So a 95% confidence Bootstrap Confidence Intervals in practice will not capture the true parameter value exactly 95% of the time. In addition to "hitting" the parameter 95% of the time, a good confidence interval should divide its 5% of "misses" equally between high misses and low misses. We will say that a method for obtaining 95% confidence intervals is accurate in a particular setting if 95% of the time it produces intervals that accurate capture the parameter and if the 5% misses are equally shared between high and low misses. Perfect accuracy isn't available in practice, but some methods are more accurate than others.

One advantage of the bootstrap is that we can to some extent check the accuracy of the bootstrap *t* and percentile confidence intervals by examining the bootstrap distribution for bias and skewness and by comparing the two intervals with each other. In general, the t and percentile intervals may not be sufficiently accurate when

- the statistic is strongly biased, as indicated by the bootstrap estimate of bias;
- the sampling distribution of the statistic is clearly skewed, as indicated by the bootstrap distribution and by comparing the t and percentile intervals; or
- we require high accuracy because the stakes are high

Percentile-T Method

Another method for generating confidence intervals that has been useful in some situations is the *percentile-t method*. The t statistic is the basis of the well-known t test, which assumes that the sample statistic of interest (typically the mean) is distributed normally. If this assumption is not valid, then the *t* calculated will not follow the theoretical t distribution, and statistical inferences based on that distribution will be incorrect. The *percentile-t method* addresses this issue by

using the bootstrap to estimate the true distribution of the *t* statistic, and then use the estimated distribution to create confidence intervals and make statistical inferences. The *percentile-t method* is similar to the percentile method, except the bootstrap is used to estimate the distribution of *t*, rather than $\hat{\theta}$. Confidence intervals calculated using the *percentile-t method* sometimes are referred to as "*studentized" i*ntervals.

Here is the algorithm:

1. Compute the statistic θ and the standard deviation of that statistic, $\hat{s}$, from the sample of values X. Then, compute T using the formula T = $\frac{\hat{\theta}-\mu}{\hat{s}/\sqrt{n}}$, where n is the sample size.
2. Generate B bootstrapped samples, X_b^* by randomly sampling X with replacement. Typically, B is a large number (e.g., > 1000).
3. For each X_b^*, compute θ_b^*, $\hat{s}_b^*$, and $T_b^* = \frac{\theta_b^* - \hat{\theta}}{s_b^{*/}/\sqrt{n}}$, where $\hat{s}$ is the standard deviation and n is sample size.
4. Sort T_b^* from smallest to the largest value.
5. Define l_o= ROUND(αB/2)and hi = B − lo.
6. Letting $[T_1^* \le T_2^* \le T_3^* \le ...T_B^*]$ represent the ordered T_b^*values, then the (100 ×α)% confidence interval for T is $((T_{lo}^*, T_{hi}^*)$.
7. In a two-tailed test, Ho, (that $\hat{\theta}$ = μ) is rejected if T«T_{lo} or »T_{hi} and the (100 ×α)% confidence interval for μ is $[\hat{\theta} - T_{hi}^* \frac{\hat{s}}{\sqrt{n}}, \hat{\theta} - T_{lo}^* \frac{\hat{s}}{\sqrt{n}}]$.

Notice that the percentile-t method requires an estimate of standard deviation of the sampling distribution, $s\sqrt{n}$, for each bootstrapped sample. In cases where there is no analytical formula for s, then it can be calculated by performing a bootstrap on the bootstrapped sample. In other words, the calculation of s may require a bootstrap within a bootstrap, or a double bootstrap. Obviously, such a procedure can be costly in terms of computing power.

We have presented these two bootstrap methods of precision estimation for complex sample surveys, and some research in this field will be conducted.

3. The practice of presentation and use of sampling errors in Polish household surveys

Sampling error is often the only error source presented in Polish publications when reporting survey estimates. Sampling errors are communicated in a number

of different forms, e.g., standard errors, relative standard error (coefficient of variation), and confidence intervals.

Reports from the Polish household sample surveys, such as the Household Budget Survey (HBS), the Labour Force Surve,0y (LFS) and EU-Statistics of Income and Living Conditions survey (EU_SILC), contain presentation of sampling errors, estimated according to the methods described above. Results of the HBS are published yearly in a special publication "*Household Budget Survey in .....".* Such publication contains selected items of estimated value with standard error and relative standard error (e.g. GUS, 2010a). Results of the LFS are published quarterly by the Central Statistical Office in a special publication "*Labour Force Survey in Poland*" in the Information and Statistical Papers series. The mentioned publication issues consist of methodological section, analytic chapter and statistical tables. Statistical tables are composed of review tables with selected results supplied *with precision indexes*, review tables from the years 1992-2004 and detailed tables presenting results of survey conducted in the most recently surveyed quarter. The LFS reports include the assessment of the precision of the most important items expressed as relative of a half of 95% confidence interval (e.g. GUS, 2010b). Reports of the EU-SILC results present in the annexes sampling error and relative sampling error for different income items by size and type of households (absolute error of estimates and relative error of estimates as they are called) (e.g. GUS, 2009). These households sample survey results are available also via the CSO internet site http://www.stat.gov.pl (link English/basic data/.

However, we have noticed that some of users of complex household survey results treated them as simple random samples for statistical analysis. For these reasons, several publications were devoted to study the *design effects* for several characteristics (e.g. Kordos et al., 2002).

In our opinion, presentation of sampling errors in reports from the complex household sample surveys in Poland is quite comprehensive, but formal. Users of these results should be informed how to interpret sampling errors for estimated parameters, and that for statistical analysis it is necessary to take into account that these results were obtained by complex sample surveys.

4. Concluding remarks

We presented here only general development of standard error estimation methods in complex household sample surveys in Poland during last forty years. Some of our experiments in this field have been only mentioned. Based on empirical data, we have compared three methods of sampling errors estimation: linearization, jackknife (JRR) and bootstrap (McCarthy & Snowden) for five income poverty indicators using data from European Statistics on Income and Living Conditions (EU-SILC) carried out by Central Statistical Office of Poland in year 2007 (Zięba & Kordos, 2010). We have studied how many bootstrap

replicates are needed for efficient standard error estimation. Our special attention has been paid to bootstrap estimation methods, and we have also tried to compare two bootstrap methods for precision estimation, i.e. the percentile method, and the McCarthy & Snowden method. Our study in this field is still continued to find efficient methods of precision estimation in complex household sample surveys.

Although the bootstrap may seem to perform better than other methods of standard error estimation of complex household sample surveys, in the sense that it permits statistical inference in very general circumstances, it is not substitute for good quality data. The performance of the bootstrap depends of the sample size. It is not possible to recommend minimum sample sizes, because each problem is different. However, increasing the number of bootstrap replicates or using a more sophisticated bootstrap procedure does not compensate for insufficient data. All the bootstrap can do is quantifying the uncertainty in the conclusion. It means that for social data which are usually biased for different reasons, bootstrap may only estimate precision of the estimates.

5. Acknowledgment

The authors would like to thank the Central Statistical Office of Poland for the production and provision of the survey data used, Prof. Vijey Verma and Dr. Waldemar Popinski, the referees, for their very constructive comments, Dr. Robert Wieczorkowski for consultation and computation, and Mr. Bronisław Lednicki for sampling consultation.

REFERENCES

BICKEL, P.J., and FREEDMANN, D.A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *Ann. Statist., 12,* pp. 470–482.

BINDER, D.A., KOVACEVIC, M.S., and ROBERTS, G. (2004). Design-based methods for survey data: Alternative uses of estimating functions. *Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods.* 3301–3312

BRICK, M.J., MORGANSTEIN, D., VALLIANT, R. (2000), Analysis of Complex Sample Data Using Replication. *WESTAT*, 1–10.

CHAN, M.T., and LO, S.H. (1985). A bootstrap method for final populations.. *Sankhya. A,* 47, 399–405.

EFRON, B., (1979) Bootstrap methods: another look at the jackknife, *Annals of Statistics* 7, 1–26

EUROSTAT, (2009) *Methodological studies and quality assessment of EU-SILC*, Report SILC.04 15 April 2009, SAS programs for variance estimation of the measures required for Intermediate Quality Report

FAUCHER D., LANGLET E. R., LESAGE E., (2003), An application of the bootstrap variance estimation method to the participation and activity limitation survey, *Assemblée annuelle de la SSC*, *Recueil de la Section des méthodes d'enquête.*

FAY, R.E. (1989). Theory and Application of Replicate Weighting for Variance Calculations. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 212–217.

GROSS, S. (1980). Median estimation in sample surveys. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*., pp.181–184.

GUPTA, V.K., and NIGAM, A.K. (1987). Mixed arthogonal arrays for variance estimation with unequal numbers of primary selections per stratum. *Biometrika,* 74, 735–742.

GURNEY, M. nad JEWETT, R.S. (1975). Constructing orthogonal replications for standard errors. *J. Amer .Statist. Assoc.*70, 819–821.

GUS (1986), *Methods and Organisation of Household Budget Survey*. Methodological Publication, No. 62, Warszawa (in Polish)..

GUS (2009), *Incomes and Living Conditions of the Population in Poland* (report from the EU-SILC survey of 2007 and 2008), Statistical Information and Elaborations, Warsaw.

GUS (2010a), *Household Budget Surveys in 2009*, Statistical Information and Elaborations, Warsaw.

GUS (2010b), *Labour Force Survey in Poland – I Quarter 2010*), Statistical Information and Elaborations, Warsaw..

KISH L., (1965) *Survey Sampling*, New York, Wiley

KISH L., FRANKEL M. R., (1974) *Inference from Complex Samples*, "Journal of the Royal Statistical Society", Series B. (Methodological), Vol. 36, No. 1. 1–37

KORDOS, J. (1985), Towards an Integrated System of Household Surveys in Poland, *Bulletin of the International Statistical Institute,* Vol. 51, Amsterdam, Book 2 , pp. 13–18.

KORDOS, J. (1996), Forty Years of the Household Budget Surveys in Poland. *Statistics in Transition*, Vol. 2, No. 7, pp.1119–1138.

KORDOS, J. (2001), Some Data Quality Issues in Statistical Publications in Poland, *Statistics in Transition,* vol. 5, No. 3, December 2001, pp. 475–488.

KORDOS ,J., LEDNICKI, B. and ZYRA, M.(2002), The Household Sample Surveys in Poland, *Statistics in Transition,* Vol. 5, No.4, pp. 555–589.

KORDOS ,J., SZARKOWSKI, A. (1974)**,** Method of Precision Estimation for Basic Parameters in Household Budget Surveys .*Wiadomości Statystyczne***,** 1974, nr 5. (in Polish).

KOVAR, J.G., RAO, J.N.K., and WU, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *Canadian Journal of Statistics, **16***, 25–46.

LEDNICKI, B. (1982), Sampling Design and Estimation Method in Household Budget Rotation Survey. *Wiadomosci Statystyczne*, No. 9. (in Polish).

LEDNICKI, B. (2011), Application of the balanced repeated replication method in Polish Household Budget Survey. In: *Methods and Organisation of Household Budget Survey*. Methodological Publication, Warszawa (in Polish) (in preparation).

LOHR, S. (1999), *Sampling: Design and Analysis,* Duxbury Press.

MAHALANOBIS, P.C.(1946), Recent experience in statistical sampling in the Indian Statistical Institute*, JRSS(A),* 108, 326–378.

MCCARTHY P. J., SNOWDEN C. B., (1985), The Bootstrap and Finite Population Sampling, *Vital and Health Statistic*s, Series 2, no. 95, Public Health Service Publication 85–1369, U.S. Government Printing Office, Washington DC

NIEMIRO W., WIECZORKOWSKI R., (2005) Approximating Variance for Measures of Income Inequality: An Approach Based on the Theory of M-Estimators and U-Stat*i*stic*s, Eurostat, 2005*

OSIER G., (2006) *Variance estimation: the linearization approach applied by Eurostat to the 2004 SILC operation*, Technical report, Eurostat and Statistics Finland Methodological Workshop on EU-SILC, Helsinki, 7–8 November 2006 (http://www.stat.fi/eusilc/workshop_en.html)

QUENOUILLE, M. H. (1949). Problems in plane sampling. *Annals of Mathematical Statistics* 20, 255–375.

POPIŃSKI W., (2006), Development of the Polish Labour Force Survey, *Statistics in Transition,* Vol. 7, No. 5, pp.. 1009–1030.

RAO, J.N.K., and WU, C.F.J., (1988). Resampling Inferences with complex survey data. JASA, 83, 321–241.

RAO, J.N.K., and WU, C.F.J., and YUE, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology,* 18, 209–217.

RAO, J.N.K., and SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometricka,* 79, 811–822.

RAO, J.N.K. and SHAO, J. (1996) On Balanced Half-Sample Variance Estimation in Stratified Random Sampling. *Journal of the American Statistical Association,* 91, pp. 343 –348.

RAO, J.N.K. and SHAO, J. (1999). Modified Balanced Repeated Replication for Complex Survey Data. *Biometrika*, 86, pp. 403–415.

SÄRNDAL, C.E., SWENSSON, B., WRETMAN, J.H., (1992) *Model Assisted Survey Sampling*, New York: Springer-Verlag.

SHAO J., (2003), Impact of the Bootstrap on Sample Surveys, "*Statististical Science*" Vol. 18, No 2, 191–198.

SHAO J. and TU, D., (1996) *The Jackknife and Bootstrap*, Springer-Verlag, New York.

SHAO, J. and WU, C. F.(1989). A general theory for jackknife variance estimation. *Ann. Statist., 15,*1563–1579.

SITTER R. R., (1992) Comparing Three Bootstrap Methods for Survey Data, *The Canadian Journal of Statistic*s, Vol. 20, No. 2, 135–154.

SITTER R. R., (1993). Balanced repeated replications based on orthogonal multi-arrays. *Biometrika,* 80, 211–221.

SOME, R. K. (1996), *Practical Sampling Techniques*, Marcel Dekker, Inc., New York.

SUKHATME, P.V. and SUKHATME, B.V. (1970), *Sampling Theory of Surveys with Applications*, FAO, Rome.

SZARKOWSKI, A., WITKOWSKI, J. (1994), The Polish Labour Force Survey*, Statistics in Transition*, Vol. 1, No. 4, pp. 467–483.

TIBSHIRANI, R., (1985) *How many bootstraps?*, Technical report no 362, Department of Statistics, Stanford University

TUKEY, J.W.(1958). Bias and confidence in not-quite large samples. . *Annals of Mathematical Statistics.* 29:614.

VERMA V. (1993). *Sampling Errors in Household Surveys: A Technical Study*. New York: United Nations Department for Economic and Social Information and Analysis. (Statistical Division INT-92-P80-15E).

VERMA V., BETTI G. (2005), Sampling errors and design effects, *Working Papers*, no. 53, Dipartimento di Metodi Quantitativi, Università di Siena.

VERMA, V., BETTI, G. (2010): Taylor linearization sampling errors and design effects for poverty measures and other complex statistics, *Journal of Applied Statistics*; iFirst.

WOLTER, K.M. (1985), *Introduction to Variance Estimation*, Springer-Verlag, New York.

YUNG, W. and RAO, J.N.K. (1996). Jackknife Linearization Variance Estimators Under Stratified Multi-Stage Sampling. *Survey Methodology,* **22**, pp. 23–31.

ZASĘPA, R. (1972), *Sampling Methods*, PWE, Warsaw (in Polish).

ZASEPA, R. (1993). Precision of family budget survey results*, Wiadomosci Statystyczne,* No. 3. (in Polish).

ZIĘBA, A. and KORDOS, J. (2010), Comparing Three Methods of Standard Error Estimation for Poverty Measures. In: J. Wywiał, W. Gamrot (Eds), *Research in Social and Economic Surveys,* Katowice, University of Economics.