STATISTICS IN TRANSITION-new series, Spring 2013
Vol. 14, No. 1, pp. 89–106

SAMPLE SURVEYS OF HOUSEHOLDS IN BELARUS: STATE AND PERSPECTIVES

Natalia Bokun[1]

ABSTRACT

The main principles, characteristics and problems of three sample surveys of households (HH), conducted by the State Statistics of Belarus are considered: 1) The Household Sample Surveys (on expenses and incomes), 2) Private Subsidiary Plots in rural areas (PSP) and 3) Labour Force Survey (LFS). For each of them the purpose, sampling plan, sample design, data collection mode, the methods of estimation and possible ways to improve the surveys are discussed.

Key words: sample fraction, territorial probabilistic multistage sampling, weighting, non-responses, private subsidiary plots, Labour Force Survey.

1. Introduction

Over 70% of Belarus's population of 9.49 million resides in urban areas. According to the Census (2009) there were 2.5 million households in rural areas and 1.1 million in urban areas. There is a big income inequality. About 20% of the population have incomes below the minimum consumer budget which is set to 1171.6 thousands of Belarusian rubles or 144.6$ for a single person. The biggest part of the household expenditures is spent on purchasing foodstuffs (37–40%). Expenditures on clothing, footwear, textiles, furniture, and household goods make up 17–18%, housing and utility are about 7–8%, and costs for education, health, culture, recreation and sport amount to 7–9%. Almost all rural residents have personal subsidiary plots. Thus, households produce about 30–35% of all agricultural products, about 89–90% of all potatoes, more than 80% of vegetables, 32–33% of eggs and 13–19% of all milk. The main information source about the household status is the census but it is complemented by three nation-wide sample surveys: the Household Sample Survey, the sampling of subsidiary plots and the Labour Force Survey. They will be described in three separate sections below, which are followed by a discussion of the future development of sample surveys and statistics in Belarus.

[1] Belarus State Economic University, Minsk. E-mail: nataliabokun@rambler.ru.

In addition, in Belarus a number of experimental surveys of health care are held (Institute of Statistics, 2005), the living standards of certain categories of workers (Institute of Statistics, 2006), consumption of alcoholic beverages (Institute of Statistics, 1999, Belarus National Academy of Sciences, 2009-2011) and public opinion polls. They are of small size and they are held irregularly. In 2005 and 2012 Multiple Indicator Cluster Surveys were held (MICS 3 and MICS 4). These surveys were conducted under the auspices of UNICEF. Despite the extensive program, the questions about illness and health are not detailed enough. For the information in the field of small businesses development, retail trade, wages in the context of professions and positions can only be obtained on the basis of industry enterprises sample surveys.

The implementation process of sampling methods in practical statistics is extremely slow. The survey of reproductive health and marketing surveys are not conducted; sample surveys of enterprises cover a limited range of issues. The priority is given to the continuous reporting.

2. Household Sample Survey

Until 1995 a survey of family budgets of working people was conducted in Belarus. The sample size was 3.5 thousand persons. Two-stage sample design was used: at the first stage the enterprises were selected within branches and then employees were selected. This principle of selection ensured representativeness of data about employees' incomes, but due to development of market relations and liberalization of labor activity the statistics of family budgets has ceased to provide objective information about amounts and sources of income. In this regard a new model for Household Surveys was developed and implemented in the statistical practice. It was based on the international standards in sample design, development tools, data processing (Metodicheskie ukazania, 1997). In accordance with the proposed methodology Household Sample Survey (HSS) has been conducted since January, 1995.

HSS is the only information basis for studying living standards. Its main purpose is to get information about the welfare of all population and particular demographic groups, detailed income and expenditure data.

The information obtained is actively used by the government, research institutes and other users. The data are used for analysis and publication to assess living standards, development of the social policy, billing the household sector SNA, in the CPI and other economical and statistical calculations.

The survey is carried out in all regions and separately in Minsk. Private households are sampled. The participation in the survey is voluntary.

The household (HH) is a group of people living together and maintaining a joint unit. Persons not belonging to any HH and living and managing a household are considered as single person HHs.

Sampling plan. The sample size is approximately 0.2% or 6000 HHs. The survey covers 49 cities and 53 rural soviets.

The sampling frame is based on the Census data. The sample design is multistage sampling. Territorial three-stage probability sampling is used:

1) at the first stage sample units are cities and rural soviets (village councils);
2) at the second stage sample units are local polling districts in cities and settlements (villages and hamlets) listed in the registers of the rural soviets (village councils);
3) at the third stage sample units are households.

At the first stage large cities are fully observed (over 72 thousands of people); small cities are selected through the sampling step, which is proportional to the population of each region. At the second and third stage systematic sampling is also used. The first unit is determined randomly.

The procedure of cities and rural soviets selection is repeated once in ten years, selection of polling districts and HHs is carried out annually.

Weighting procedure. The methodology of weighing and extrapolation data on a general population is based on assignment to each unit (HH) the corresponding weight (B_i):

$$B_i = \frac{1}{p_1 \cdot p_2 \cdot p_3}, \tag{2.1}$$

where p_1 - the probability of selecting a city or a rural soviet; p_2 - the probability of a polling district in a city, zone or rural soviet; p_3 - the probability of selecting a household within a polling district or zone.

Base HH weights are corrected *for uninhabited apartments* and *non-responses* by using overweighting procedures. Weighted cells are constructed with the usage of the following characteristics: region, type of a settlement, type of housing, size of HH. Each cell includes at least 20 HHs. On the basis of the modified cells new weights are calculated:

$$V_{ki}^{*} = \frac{\sum_{j=1}^{M_k} V_{kj} + \sum_{j=1}^{N_k} V_{kj}}{\sum_{j=1}^{N_k} V_{kj}} \cdot V_{ki}, \tag{2.2}$$

$$V_{li} = \frac{\sum_{j=1}^{L} B_{lj} + \sum_{j=1}^{K} B_{lj}}{\sum_{j=1}^{K} B_{lj}} \cdot B_{li}\,, \qquad (2.3)$$

where V_{ki} *- weight of HH, which fell in the *k*th cell, corrected for non-responses; V_{ki} - weight of HH, which fell in the *k*th cell, corrected for non-residential apartments; M_k - number of non-responses in the *k*th cell; N_k - number of responses in the *k*th cell; B_{li} - base weight for the *i*th HH in the *l*th region; V_{li} - base weight for the *i*th HH in the *l*th region, corrected for non-residential apartments; $l = 1,7$ - number of the region; L – number of non-residential apartments in *l*th region; M - number of HHs left in the *l*th region.

Data collection. Data are collected with the use of face-to-face interviews using paper and pencil (PAPI). The field staff comprises 150 interviewers. Before visiting the HH the interviewer sends a copy of a letter-appeal to each selected address with the request to take part in the survey. The letter briefly describes the procedure of examination and the date of the first visit.

The sample program assumes filling in some questionnaires (living conditions, personal subsidiary plots, education, health, and employment), daily and quarterly questionnaires: expenditures on foodstuffs and nonfood products, payment of services, etc.

The main components of the survey are: the main interview (a questionnaire, which is to be filled at the beginning of the survey); four quarterly interviews (conducted in April, July, October and January); four two-weeks diaries, which the households get once every quarter and in which they have to indicate their expenditures on foodstuffs and non-food products separately for each day. More than 10 000 variables are observed in the survey.

The average response rate is 70-80%. Refusals are 1-2% of the total non-response; remaining 98% are “impossible to contact”, “not at home”, “unable to answer”, “incapacity”, “unreturned questionnaire”. State statistics bodies ensure the confidentiality of the information provided by households. This information is used exclusively for the compilation of summary statistics.

Dissemination of the HH data to users is carried out by the publication of statistical books, bulletins and with the help of the Belstat website. The main publications are: Incomes and expenditures of the population in the Republic of Belarus; Social status and living standards of the population of the Republic of Belarus; Statistical Yearbook of the Republic of Belarus.

Problems. The mechanism of quarterly HH survey samples is sufficiently worked out. Nevertheless, the survey process has such problems as: high non-response rate (up to 30%), the need of building regional and demographic sub-samples, usage of more sophisticated models of imputation, in addition the replacement of non-response data by neighboring units etc. The solutions can be: to increase the number of weighted variables, to increase the number of interviewers, and to use ratio imputation and regression imputation.

3. Sampling of Subsidiary Plots

A Personal Subsidiary Plot (PSP) is a small plot of land around the house that is worked by the holder. In Belarus the sample survey of Personal Subsidiary Plots (PSP) has been conducted since June, 2010. Its main purposes are:

- to obtain data on the output of plant growing and livestock products, the number of livestock and poultry, the size of sown areas, the amount of feed consumption of livestock and poultry, and the amount of sales;
- to calculate the gross output in agriculture;
- to develop food balance sheets and funds for personal food consumption.

Survey objects are HHs, personal subsidiary households of citizens in rural areas. These households are examined for each region. Participation in the survey is voluntary.

PSP data and results are used extensively by Belstat and other government bodies to estimate total agricultural output in Belarus and in each region, to develop regional economic policy taking into account trends in agricultural production, output of PSP.

Sampling plan. The sample size is approximately 3600 PSPs, the sample fraction is 0.32%, and the maximum relative error is 5–10%.

The sampling frame is based on the last Census data and household register. The sampling frames consist of:

- a set of districts in each region;
- a set of village councils (rural soviets) in each selected area;
- villages (settlements) in each selected village council;
- the totality of the households in each village (data household register).

Two indicators are recorded for each unit: the size of the total land area and the number of conventional livestock.

Territorial four-stage probability sampling is used (Bokun, N., (2010); Nauchno-obosnovannoe metodologicheskoe obespechenie, 2010).

At the first stage sample units are districts within the region; at the second stage – village councils within selected districts; at the third – villages within selected rural councils; at the fourth – private household farms in the selected villages.

At each stage the selection of units is based on the probability that is proportional to the sample indicators (total land area, number of conventional livestock).

The first stage. Districts are selected. In Belarus a district is an administrative unit in rural areas, providing statistical data. Therefore, in the frames of HH survey sample their maximal representativeness is desirable. The maximum sampling rate that exceeds the normal or mean can range from 50% to 90%. For the area selection it is reasonable to use the middle of the interval, i.e. sampling rate 70%. 80 districts are selected from the 118 districts ($118 \cdot 0.7 \approx 80$), which are distributed over the regions (Table 1). A systematic sampling algorithm is used: districts are ranked by the number of households. For each district the values of indicators "total area of land" and "amount of conditional livestock" are calculated for the private plots.

Table 1. The composition of district sample

Regions	Number of districts		District sampling rate, d_1	Number of households		
	in region, N_1	in sample, n_1		total, N	in selected districts, N_{11}	selected, (n), distribution by regions
Brest	16	11	0.687	210606	147614	495
Vitebsk	21	14	0.667	161100	117648	630
Gomel	21	14	0.667	174159	140204	630
Grodno	17	12	0.706	166640	127045	540
Minsk	22	15	0.682	273061	209715	675
Mogilev	21	14	0.667	128692	100381	630
TOTAL	**118**	**80**	**0.678**	**1114258**	**842607**	**3600**

The second stage. One village council is randomly chosen in each of the selected districts, which leads to a certain degree of uncertainty in the representativeness of the district. But since summary information is presented only by regions, and not by districts, it is neglected. In addition, every interviewer is assigned to one village council. The interviewer does not need to make long journeys to conduct surveys in other areas. Averages, totals, figures for the oscillation of the analyzed characteristics are estimated. Then village council of medium size with minimal values of deviations from the mean values in a district is selected.

The third stage. The list of settlements ranked by the number of households is composed for each selected village council. The settlements consisting of a small number of households (1, 2, 3 etc. HHs) are excluded. Unit selection is determined by a random number generator or by a table of random numbers.

The fourth stage. The number of subsidiary plots is determined. The disproportionate approach is used. This means that 45 households are sampled for each selected locality settlement. Household selection is done in a mechanical way using the accumulated amount of the parameters "total area of land" and "conditional livestock".

Weighting procedure. The extrapolating (here and in the following you should replace extrapolation/extrapolate by estimation/estimate) of mean and total values of sowing and harvesting areas, and all kinds of cultures, the total land area, the number of cattle, the gross collection of crops, livestock production, feed consumption of livestock and poultry and others are provided.

Extrapolation is carried out by the following methods:

1) by simplified method;
2) by the probability of selection for each of the four stages of selection;
3) by ratio estimation.

Simplified method. The methodology of weighing and estimations is based on assigning to each unit (PSP) the corresponding weight (B_e):

$$B_e = \frac{1}{p_p \cdot p_c \cdot p_n \cdot p_q}, \tag{3.1}$$

where p_p – probability of selection of region in a district; p_c – probability of selection of village council in the selected district; p_n – probability of selecting a point in the selected village council; p_q – the probability of selection of each household within a sampled settlement.

The investigator has to take into account variability of the studied parameters. Therefore, for each HH several basic weights are calculated (for crop, livestock, etc.).

Taking into account the variability of the studied indicators some basic weights are calculated for each HH (for the extrapolation of indicators for crops, livestock, etc.). For the calculation of the selection probability the size of PSP is estimated by land area and livestock. For example, the estimation of the crop base weight is determined by the formula:

$$B_{e'p} = \frac{1}{p'_p \cdot p'_c \cdot p'_n \cdot p'_q}; \tag{3.2}$$

$$p'_p = \frac{S_{pj} \cdot n_{1i}}{S_i};\ p'_c = \frac{S_{cj}}{S_{pj}};\ p'_n = \frac{S_{nj}}{S_{cj}};\ p'_q = \frac{S_{ej} \cdot n_{4j}}{S_{nj}}, \tag{3.3}$$

where p'_p, p'_c, p'_n, p'_q are selection probabilities of district, village council, village, HH, calculated taking into account the land area of PSP in the region, district, village council, village, separate HH respectively. S_{Pj} is land area of PSP in the selected *j*-th district (1st stage); S_i - total area of PSP land in the *i*-th region;

n_{1i} - the number of districts selected in the *i*-th region; n_{4j} - the number of households selected in the *j*-th village council; S_{ej} - land area in the selected *e*-th household (4 stage); S_{cj} - total land area of PSP in the selected village hall *j*-th district (2nd stage); S_{nj} - the total land area of PSP in the selected village *j*-th district (3rd stage).

Data extrapolation on the probability of selection for each of the four stages of sampling. At each stage of the selection the value of the average and total values of a characteristic are extrapolated. The calculation is made separately for livestock and crop.

IV stage. Estimation of each characteristic for the crop is carried out by the following weight:

$$B_{ep_4} = \frac{S_n}{S_e}, \tag{3.4}$$

where Bep_4 is the reciprocal of the probability of selecting PSP from the totality of human settlements on indicators of crop at stage 4; S_n – the area of PSP land in all selected locations; S_e – the land area in the *e*-th HH included in the sample.

III stage. For estimation of crop characteristics the weight of a settlement is calculated as follows:

$$B_{ep_3} = \frac{S_c}{S_{nj}}, \tag{3.5}$$

where Bep_3 – is the reciprocal of the probability of selecting a settlement in a village council selected at the stage 2; S_c – the area of PSP land in selected village council; S_{nj} – the area of PSP village council land in the selected village *j*-th region.

II stage. Weighting of the village council (for plants):

$$B_{ep_2} = \frac{S_p}{S_{cj}}, \tag{3.6}$$

where S_p – the area of household land in all selected districts of a region; S_{cj} – the area of land in the selected council in the *j*-th district.

I stage. The area weights are calculated as follows:

- crop $$B_{ep_1} = \frac{S_i}{S_{pj}}; \tag{3.7}$$

- livestock $$B_{el_1} = \frac{(S_i + Y_i)}{(S_{pj} + Y_{pj})}, \quad (3.8)$$

where S_i – the area of PSP land in the *i*-th region; Y_i – conventional livestock in PSP of *i*-th region; S_{pj} and Y_{pj} – the area of land and conventional livestock in the selected *j*-th district in the *i*-th region respectively.

Extrapolated total value of a characteristic is defined as a product of the average value of the trait and the number of households in the region, or as a sum of weighted values of a variable at the first stage.

Ratio estimation. The sample population for each region is formed. Average and total values are extrapolated using of the raising coefficients (*Kp*):

- crop $$K_{pp_1} = S : \sum_{1}^{n_4} S_e ; \quad (3.9)$$

- livestock $$K_{pl} = \frac{S + Y}{\sum_{1}^{n_4} (S_e + Y_e)}. \quad (3.10)$$

Selection of the optimal extrapolation method depends on the initial data and is determined by the minimal standard sample error. We use the classical formula for calculating the variance for multi-stage sample, as well as the variance of ratio estimators (Bokun, N., Chernysheva, T (1997); Cochran, W (1997)).

Non-response adjustment is based on the donor imputation: selection of values with replacement from the set of respondents.

The results of subsidiary plots sample survey held in Belarus in 2010 are shown in Table 2, where: X_1 is gross harvest of grains and legumes (quintals); X_2 is gross harvest of potatoes (quintals); X_3 is gross harvest of vegetables (quintals); X_4 is the number of cows; X_5 is the number of pigs.

Table 2. Sample survey of subsidiary plots in Belarus, 2010

Indicators	Total value of parameter			Sample error, %
	sample	general	estimated value	
1. Simplified extrapolation method				
X1	409	27209	23858.5	9.8
X2	1020	80306	91629.1	14.1
X3	6231.34	799409.3	945701.2	18.3
X4	33426.31	4604302.6	5460922.4	3.1
X5	6124.32	1104521.2	1129925.1	2.3

Table 2. Sample survey of subsidiary plots in Belarus, 2010 (cont.)

Indicators	Total value of parameter			Sample error, %
	sample	general	estimated value	
2. Data extrapolating on the probability of selection for each of the four stages				
X3	6231.34	799409.3	815397.48	2.0
X4	33426.31	4604302.6	4765453.1	3.5
X5	6124.32	1104521.2	1158893.6	3.9
3. Ratio estimated				
X1	409	27209	32786.8	20.5
X2	1020	80306	81831.8	1.9
X3	6231.34	799409.3	945701.2	18.3
X4	33426.31	4604302.6	4415526.2	4.1
X5	6124.32	1104521.2	976396.8	11.6

Data presented in Table 2 are examples of different estimation methods used in subsidiary plots sample surveys held in Belarus. The most preferred extrapolation methods are based on using base weights, which take into account the sizes of cultivated areas and livestock. In some cases ratio estimators are better. Additional usage of extrapolation over probabilities at each of the four selection stages is also possible (in the case of high error in the first two methods, for example, when evaluating the total yield of vegetables).

Preliminary assessment of the acceptable degree of accuracy shows that the standard relative error for the whole Belarus is 1-2%; for the regions it is 5-6%; for small-size areas it is 8-15%. The standard relative error of the sample for sown area is 5-6%; for land area – 0.1-0.5%; for the number of livestock – 5-10%; for the planted area with potatoes and vegetables – 5-5.6%.

Data collection. Face-to-face interviews are used to survey the items of interest in the questionnaire. According to the national specificities the optimal interviewer load is nearly 45 households. The data are collected by 80 field workers using paper questionnaires. Respondents maintain their accounting records of the volume of crop production, livestock, provide information about the presence and movement of poultry livestock, acreage size of family members, etc.

Five questionnaires are used: basic questionnaire (as of 1 January), questionnaire on the crop area (as of 1 June), on the presence and movement of livestock and poultry (quarterly), diary registration of crop production (5 times a year, June-October), diary of livestock products registration and feeds accounting

(monthly). The collected information is confidential and it is used for the aggregate indicators calculation.

The average response rate is 85-90%. Refusals make 60-65% of the total non-response. Main publications are: Agriculture in the Republic of Belarus; Statistical Yearbook of the Republic of Belarus. Survey results are also presented on the website www.belstat.gov.by.

Problems. The results of surveys in 2010-2011 have shown: 1) real response rate was higher than the planned one (85-90% versus 80%). This fact indicates a positive attitude of respondents to the survey; 2) at the regional level for the investigated variables (land area, crop area, number of pigs, etc.), the discrepancy between the estimates and the data of households recording are within an acceptable range (10-15%); an exception is the number of indicators of cattle, for which estimates are much lower than the continuous data records; this may be due to errors in the sampling frame: in some areas the number of livestock is overestimated, and it needs updating; 3) relative standard errors for most indicators of questionnaires did not exceed the permissible level; 4) it is quite difficult to select any option of extrapolation for various indicators of the questionnaire. Further improvement of the survey methodology may be related, firstly, to updating household recording, secondly, to the development of algorithm of choosing the optimal method for extrapolating the individual indicators (sections) of the questionnaire, and, thirdly, to study the possible application of the iterative weighting.

4. Labour force survey

Nowadays, the National Statistical Committee of the Republic of Belarus together with some foreign and national experts makes the preparatory work on implementation of the Labour Force Survey (LFS). In November 2011 a test sample survey was conducted. Since 2012 LFS has been provided on a regular basis.

The purposes are:

- to obtain empirical statistics on the labour force, economically active population, employed, unemployed;
- to obtain empirical statistics on labour force, employed, unemployed by sex, regions, rural, urban;
- to determine real labour force demand and supply.

Frequency of the results: quarterly and annual.

LFS data will be widely used for the labour market analysis, assess the actual level of unemployment, making optimal management decisions in the field of employment.

The survey covers the whole country: urban and rural areas in each region. Private households are surveyed. Participation in the survey is voluntary.

The target population comprises all residents aged 15-74.

Sampling plan. The size of the sample is perhaps the most important parameter of the sample design, as it affects the precision, cost and duration of the survey more than any other factors.

To calculate the ***sample size***, with the usage of the appropriate formula, recommended strategy for calculating the sample size is to take into account several factors, connected with sample precision, design-effect (*deff*), household size and non-responses. These factors are:

- the precision, needed relative sample error;
- desired confidence level;
- estimated (or known) proportion of the population in the specified target group;
- predicted coverage rate, or prevalence for the specified indicator;
- sample *deff*;
- average household size;
- adjustment for potential loss of sampled households due to non-response.

Design-effect *(deff)* is a ratio of sample variances of the actual stratified cluster sample (σ_a^2) and of a simple random sample of the same overall sample size (σ^2):

$$deff = \sigma_a^2 / \sigma^2 . \qquad (4.1)$$

Two sets of problems arise at this stage. First, the value of *deff* can be easily calculated after the survey, it is not often known before the survey. Second, the value of *deff* is different for each indicator and each target group. Consequently, it is necessary to choose one more important key indicator. International statistical practice has shown that the optimal value of *deff* is 1.5 (Multiple Indicator Cluster Survey Manual (2009), p. 4.3-4.8) (which may be sometimes high). Therefore, the sample size will be large enough to measure all main indicators.

Key indicator is the most important indicator that will yield the largest sample size.

Selection of the target group and key indicator includes the following stages:

1. Selection of two or three target populations that comprise small percentages of the total population (1-year, 2-year, 5-year age groups) (Multiple Indicator Cluster Survey Manual (2009), p. 4.8).
2. Review of important indicator based on these groups, ignoring indicators that have very low (less than 5%) or very high (more than 50%) prevalence.
3. Maximal indicator value, calculated for target group (10-15% of the population) is 15-20% [6; 7].
4. Do not pick from desirable low coverage indicators an indicator that is already acceptably low.

Key indicator, used in Belorussian LFS, is the real unemployment rate (by the Census results). Target groups are economically active populations (rural, urban, by regions, 5-year groups).

The sample size formula is used (Bokun, N., Chernysheva, T (1997), p. 44-53; Multiple Indicator Cluster Survey Manual (2009), p. 4.5-4.8, 4.11):

$$n = \frac{4r(1-r)\cdot f \cdot 1.2}{(0.12r)^2 \cdot p \cdot n_h}, \tag{4.2}$$

where n – required size for the key indicator; 4 – the factor to achieve 95% level of confidence, t-criteria; r – predicted prevalence for the key indicator; 1.2 – essential factor in order to raise the sample size by 20% for non-response; f – the symbol for deff (1.5); 0.12 – recommended relative sample error (95% level of confidence); p – proportion of the total population upon which the indicator (r) is based; n_h – average household size.

Several types of the sample size calculations were executed:
1) random selection for rural and urban population for each region;
2) random selection for Belarus (for target groups);
3) random selection for each region;
4) stratified sampling for each region.

The examples of sample size determination are given in Tables 3 and 4.

Table 3. Sample size for LFS. Variant 2

Target group	Real unemployment rate		Target group size		Average household size, n_h	Number of persons aged 15-74 on average, falling to one HH, n_h'	Predicted sample size	
	persons	%, r	to total population, p	to 15-74 years age group, p'			$n_1 = \frac{4r(1-r)\cdot f \cdot 1.2}{(0.12r)^2 \cdot p \cdot n_h}$	$n_2 = \frac{4r(1-r)\cdot 1.5 \cdot 1.2}{(0.12r)^2 \cdot p' \cdot n_h'}$
Economically active population aged 20-24 (565833 persons)	60627	10.7	5.95	7.5	2.43	1.94	28860	28860
Economically active population aged 15-74 in rural area (1051627 persons)	69346	6.6	11.06	14.0	2.43	1.94	26328	26052

Table 4. Sample size for LFS. Variant 3

Regions	Population aged 15-74, *N*, persons	Number of unemployment, persons	Proportion unemployed in the population aged 15-74, *w*	Number of persons aged 15-74 on average, falling to one HH, n'_h	Sample size, *n* , number of households	
					Relative standard error μ =0,06, relative limited error Δ =0,12, (without *deff*)	Relative standard error μ =0,075, relative limited error Δ =0,15, (with *deff*)
Brest region	1073227	50065	0.047	1.92	3502	3380
Vitebsk region	979845	37108	0.038	1.87	4480	4312
Gomel region	1132928	46840	0.041	1.89	4102	3946
Grodno region	829263	31757	0.038	1.87	4474	4308
Minsk	1513844	56293	0.037	2.06	4191	4043
Minsk region	1113871	37345	0.033	1.94	4997	4811
Mogilev region	868907	38511	0.044	1.97	3651	3513
Total	7511885	297919	0.040	1.94	29397	28313

Calculation results by different variants have shown that required annual sample size is 26-29 thousand of households, or in average – 28 thousand. Without taking into account non-responses the sample size is 22 thousand. Therefore, predicted sample fraction is 0.6%, or 22 000 HHs. It is planned to examine 7 000 HHs on a quarterly basis.

Sample frame is based on the 2009 Census and includes:

- set of cities in each region;
- set of village councils in each region;
- census enumeration districts in each selected city;
- villages (settlements) in each selected village council;
- the household totality in each census enumeration district and village.

Annual updating of the lists of enumeration areas and HHs is assumed.

Sample design. The territorial three-stage sample is used: primary unit – city or village council; secondary unit – census enumeration district or village (zone); final sampling unit – household.

There are 25 census enumeration districts in cities and 16 village councils (zones).

At each stage units are selected with systematic sampling with the probability that is proportional to population size or to the number of households. Variables used for the stratification are: administrative districts, urban/rural.

Weighting procedure is connected with HH weights and individual's weights. HH weights are calculated as reciprocal of overall sample probabilities:

$$B_i = \frac{1}{p_1 \cdot p_2 \cdot p_3}, \tag{4.3}$$

where p_1 - the probability of selecting a city or a rural soviet; p_2 - the probability of selecting each polling district in cities, zones and rural soviets; p_3 - the probability of selecting each household within the Census enumerated district or zone.

For the case of *non-response* an additional array of HHs is reserved within not less than 20% of the total sample ($28000 \cdot 0{,}2 \approx 6000$).

Individual's weights are based on iterative weighting (Multiple Indicator Cluster Survey Manual (2009); Metodika provedenia bazovyh obsledovanij naselenija (1997)):

Iteration I:

a) weights are calculated separately by sex within 5-year age groups;
b) the first correction coefficient (k_1) is calculated; weighted variables are: region, sex, rural/urban;
c) the second correction coefficient (k_2) is calculated; variables are: region, sex, eleven 5-year age groups .

Individual weights are equal within each region, 5-year age groups, one kind of a settlement.

Iteration II:

At the second iteration the operations are implemented on the subsequent adjustment of the basic weight and intermediate extrapolated data on the same criteria as for the first iteration.

Final individual weights for each 5-year age group:

$$K_i = B_b \cdot k_1 \cdot k_2 \cdot k_3, \tag{4.4}$$

where: $B_b = \frac{S_j}{s_j}$; $k_1 = \frac{S_t}{S_E}$; $k_2 = \frac{S_{jt}}{S_{E2}}$; S_j, s_j – population size in *j*-th sex-age group based on the result of the Census and survey; S_t – population size in *t*-th group by rural (urban), sex (on the Census data); S_E – extrapolated population size in *t*-th group (by B_b); S_{jt} – population size in *jt*-th sex-age rural (urban) group; S_{E2} – extrapolated population size in *jt*-th group (by B_b and k_1); k_3 – generic correction coefficient, calculated in the second iteration ($k_3 = k_{31} \cdot k_{32} \cdot \ldots \cdot k_{3n}$).

Preliminary results of iterative weighting for unemployment rate and employment rate, calculated for Mogilev region (Table 5) have shown that received sample population is representative. Relative errors for the region do not

exceed 7-8%: for the number of unemployed – 6%, the number of employed – 1.8%, the unemployment rate – 6.6%.

Table 5. Indicators of sample representativeness. Mogilev region. Iterative weighting

Indicators	Characteristic value		Error	
	extrapolated, $Э_x$	in the general population, x	in absolute terms, $\Delta a = \lvert x - Э_x \rvert$	in % $\Delta = \frac{\lvert x - Э_x \rvert}{x}$
Number of employed, persons	506231.11	515876	9644.89	1.87
Urban area	402333.2	412962	10628.8	2.57
- Male	194657.81	205508	10850.2	5.28
- Female	207675.39	207454	221.39	0.11
Rural area	103897.91	102914	983.91	0.96
- Male	55227.66	55228	0.34	0.0006
- Female	48670.25	47686	984.25	2.06
Total number of employed, persons				
- Male	249885.05	260736	10851	4.16
- Female	256345.64	255140	1205.64	0.47
Number of unemployed, persons	40510.33	38511	1899.33	4.19
Urban area	32094.01	29332	2762.01	9.42
- Male	20045.51	18381	1664.51	9.06
- Female	12048.50	10951	997.5	9.10
Rural area	8416.32	9179	762.68	8.31
- Male	5931.53	6572	640.47	9.75
- Female	2484.79	2607	122.21	4.69
Number of unemployed (persons) among				
- Male	25977.04	24953	1024.04	4.10
- Female	14533.29	13558	975.29	7.19
Unemployment rate, %	7.41	6.95	0.46	6.62
Urban area	7.39	6.63	0.76	10.46
- Male	9.34	8.21	1.13	13.76
- Female	5.48	5.01	0.47	9.38
Rural area	7.49	8.19	0.7	8.55
- Male	9.70	10.63	0.93	8.75
- Female	4.86	5.18	0.32	6.18
Unemployment rate (%) among:				
- Male	9.42	8.73	0.69	7.90
- Female	5.37	5.05	0.32	6.34

The results of trial calculations and testing of the first version of methodological and software sampling have shown that the main difficulties are associated with the use of different weighting schemes, determining the number of iterations steps, evaluation of structural indicators of employment and unemployment, the presence of atypical employment on the level of primary units (cities, districts).

Data collection. The data are collected by 200 field workers with face-to-face interviews using paper questionnaires. The optimal interviewer's load in the cities is 40 HHs, in rural areas – 30 HHs. The predicted response rate is 80%. The reference week is the week before the interview.

The main component of the survey is "The questionnaire on studying employment for the surveyed week". It includes 57 questions, which are combined into seven sections, and includes the details about the respondent, basic and additional paid work, self-employment, unemployment, employment in the PSPs.

Preliminary results of the survey are to be presented on the website of Belstat.

Problems. Under a given load and a limited number of interviewers (200), it is not possible to question the estimated number of HHs (28 000) on a quarterly basis. On the basis of the selected annual array of HHs (28 000), built by regions, for each quarter, randomly generated four sub-samples are formed (each includes 7 000 HHs). If the annual array of information makes it possible to obtain sufficiently representative data at the level of the republic and regions on most indicators (number of employed, unemployed, the economically active population, employment, unemployment, and in the context of all sex-age groups, urban and rural areas), the quarterly array makes it possible to design and evaluate the indicators with an acceptable degree of accuracy (10-12%) only at the level of the country. To improve the representativeness by region the indicators of the survey can be formed on the basis of the three samples – the average for three consecutive quarters. In addition, improving the quality of sample data is possible due to testing and using various schemes of the iterative weighting.

5. Concluding remarks

The household surveys make it possible to get the information on living standards of the population, actual employment and unemployment and products produced in PSPs.

The sample units are HHs and target population groups (for example, persons aged 15-74), Personal Subsidiary Plots of citizens in rural areas. The surveys cover the whole country: the regions and Minsk city. The sample fraction is at the level of 0.2-0.6% of HHs, sample frames are Census and additional databases (household survey for the PSPs). Face-to-face paper assisted interview is used.

The experience of household sample survey construction in Belarus has shown that the most applicable form of HH selection is multi-stage territorial probability sampling. The population can be stratified by the group of indicators: the administrative center, the type of housing, the size and composition of the HH. For the survey of PSPs the additional stratification variables are: the area of land, conventional livestock, and for LFS - gender and age groups of those aged 15-74. Weighting and extrapolation are carried out both on the basis of individual weights that are calculated with the usage of linear functions (e. g., the reciprocal product of the probability of selection units at various stages of the sample), and with the usage of sophisticated estimates (ratio estimators are applied for estimation of some parameters of the PSP population).

The main problems for researchers and practitioners of statistics are: the issues of sample localization, the construction of regional (district) samples, non-sampling errors, non-response (20-30%), presence of atypical units, not appropriate extrapolation, the use of different weighting schemes, the assessment of structural employment and unemployment indicators (for LFS), improving the representativeness of the quarterly data.

Possible directions for improvement of the surveys are connected with using ratio and regression imputation, demographic and territorial sub-samples, usage of combined estimation methods for each indicator, presented in questionnaire (PSP), clarifying the steps and subsequent realization of iterative weighting scheme (LFS). It would be interesting to evaluate the goodness of sample strategy by means of Monte Carlo simulation from the census data (LFS) and household register data (PSP).

REFERENCES

BOKUN, N., (2010). Sampling of Subsidiary Plots in Belarus: methodological problems of population formation and data estimation. *Workshop on Survey sampling theory and methodology*. August, 23-27. Vilnius, Lithuania.

BOKUN, N., CHERNYSHEVA, T., (1997). Metody vyborochnyh obsledovanij, Minsk.

COCHRAN, W., (1997). Sampling techniques. *John, Willey and sons, inc.* New-York.

Metodicheskie ukazania po vyborochnomu obsledovaniu domashnih hoziajstv v Respublike Belarus. (1997). Minsk.

Metodika provedenia bazovyh obsledovanij naselenija. (2008). Kiev.

Metodologichni osnovi formuvannia viborkovyh sukupnostej dlia provedennia organami derzhavnoj statistiki Ukraini bazovyh derzhavnyh viborkovyh obstezhen naselenia (domogospodarstv). (2005). – 156 p., Kiev.

Money incomes and expenditures of population of the Republic of Belarus: statistical book. (2011). National Statistical Committee of Republic of Belarus, Minsk.

Multiple Indicator Cluster Survey Manual. (2009). Eurostat.

Nauchno-obosnovannoe metodologicheskoe obespechenie po formirovaniu vyborochnoj sovokupnosti lichnyh podsobnyh hoziajstv grazhdan, postojanno prozhivajuschih v selskoj mestnosti: otchet o NIR № GR. (2010). Nauchnyj rukovoditel – N. Bokun, BSEU, Minsk.

Social conditions and living standards of population in the Republic of Belarus: statistical book. (2012). National Statistical Committee of Republic of Belarus, Minsk.

Statistical Yearbook: statistical book. (2012). National Statistical Committee of Republic of Belarus, Minsk.