Labour Force Survey in Belarus: determination of sample size, sample design, statistical weighting

Natallia Bokun[1]

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

Abstract

The first experience of forming sampling frames in Belarus for the Labour Force Survey (LFS) is analyzed. Various options for determining the sample size are shown. Some issues of sample design and estimation are considered.
Keywords: sample size, selection, three-stage sample, iterative weighting.

1 Introduction

In Belarus, until recently, data on the size and structure of the labor force have been formed once a year, when calculating the balance of labour resources. The major sources of information on the labour market were as follows: continuous reporting of organizations, administrative sources and census. Despite a rather adequate and detailed measurement of indicators of the economically active population, the existing system of the current account had no possibility of monthly and annual estimates of the actual level of unemployment, which according to the Census 2009, 6-7 times higher than its' recorded amount; did not allow to estimate employment by age, professional groups, to determine the status of employment, underemployment, etc. These factors have caused need for a specialized 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 is 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 years.

2 Sample size

Calculation of sampling frame, on which representativeness, duration and cost of the survey largely depends, is the most important stage of sampling.

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

- the need to select a key indicator by which the sample size is calculated;
- 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.

As a *key indicator* it is recommended to select one of the most important indicators for the survey and on its basis to estimate the maximum size of sampling frame, yielding an estimate for the minimal (not less than 2.5%) stratum of the population.

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%.
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).

Design-effect (deff) describes the influence of sample structure on the value of selection bias, it is defined as 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). 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.

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):

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

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:

- random selection for rural and urban population for each region;
- random selection for Belarus (for target groups);
- random selection for each region;
- stratified sampling for each region.

In the first variant a small surveyed group is the economically active population, according to the second it is the number of economically active population in a particular age range (15-20, 20-24 or 15-74 years). In the third and fourth variants a key indicator is an unemployment rate for the unit of a total population: the proportion of unemployed in the population aged 15-74 years. In this case, there is no need to use the surveyed small groups in the calculation – to determine the sample size for each area the classic formula of the sampling theory is used used (Bokun, N., Chernysheva, T (1997), p. 27-50, 44-53), adjusted for deff, non-response and the number of persons aged 15-74 years per one HH in average.

The examples of sample size determination are given in Tables 1-3.

Table 1 – Sample size for LFS. Variant 2

Target group	Real unemployment rate		Target group size		Average household size, n_h	Number of persons of age 15-74 on average, falling to one HH, n'_h	Predicted sample size	
	persons	%, r	to total population, p	to 15-74-year 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 of age 20-24 (565833 persons)	60627	10.7	5.95	7.5	2.43	1.94	28860	28860
Economicalyl active population of age 15-74 in rural area (1051627 persons)	69346	6.6	11.06	14.0	2.43	1.94	26328	26052

Table 2 – Sample size for LFS. Variant 3

Regions	Population of age 15-74, *N*, persons	Number of unemployment, persons	Proportion unemployed in the population aged 15-74 years, *w*	Number of persons of age 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

Table 3 – Sample size for LFS. Variant 4

Regions	Population of age 15-74, *N*, persons		Proportion unemployed in the population aged 15-74 years, *w*		Sample size, *n* , number of households	
	urban	rural	urban	rural	Relative standard error μ =0.06, relative limited error Δ =0.12, (without *deff*)	Relative standard error μ =0.06, relative limited error Δ =0.12, (with *deff*)
Brest region	728125	345102	0,048	0,043	1987	2981
Vitebsk region	727698	252147	0,039	0,035	2828	4242
Gomel region	844646	288282	0,040	0,044	2525	3788
Grodno region	589695	239568	0,041	0,032	2773	4160
Minsk	1513844		0,037		4211	6317
Minsk region	631161	482710	0,034	0,033	2570	3855
Mogilev region	670561	198346	0,044	0,046	2209	3314
Total	5705730	1806155	0,040	0,038	19103	28657

Calculation results by different variants have shown that required annual sample size is 26-29 thousands of households, or in average – 28 thousands. Without taking into account non-responses sample size is 22 thousands. Therefore predicted sample fraction is 0.6%, or 22 000 HH. It is planned to examine 7000 HH on a quarterly basis.

3 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.

As a sample frame for each stage of the selection the data sets are used which are built by the Census 2009:

- set of cities in each region (the first stage);
- set of village councils in each region (the first stage);
- census enumeration districts in each selected city (the second stage);

- villages (settlements) in each selected village council (the second stage);
- the household totality in each census enumeration district and village (the third stage).

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

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.

The first stage. Towns, including urban settlements and rural councils are selected. At first the towns, which necessarily have to get into the survey, are defined. A criterion of population size for their selection is calculated from the peak value of the interviewer (40 HH), the coefficient of the sample (k = n / N) and the average household size (according to Census 2009 – 2.43): $S_{\tilde{a}} = 40 \cdot \frac{1}{0,006} \cdot 2,43 = 16200$. Thus, the sample includes all the "large" cities with a population 17 thousand people or more. Urban settlements with a population less than 17 thousand people are selected systematically or randomly within each region. Their number depends on the pre-planned number of interviewers and the proportions of the population in small and medium-sized towns (table 4). There are 78 cities to be surveyed (43 large, 35 small and medium-sized), which represent over 38% of the total number of cities in Belarus.

Table 4 – The composition of sampling frame for LFS

Region	Number of cities	Number of village councils	Number of selected		Number of selected households		
			enumeration areas	settlements in the village councils	urban	rural	total
Brest region	13	13	32	22	2560	1560	4120
Vitebsk region	14	10	34	47	2720	1200	3920
Gomel region	14	10	38	17	3040	1200	4240
Grodno region	11	11	28	36	2240	1320	3560
Minsk	1	-	56	-	4480	-	4480
Minsk region	13	16	28	33	2240	1920	4160
Mogilev region	12	8	32	25	2560	968	3528
Total	78	68	248	180	19840	8160	28000

The second stage. In urban areas, enumeration areas according to census are selected, in rural – settlements according to census or village councils accounting. They are selected either according to a predetermined loading and the number of interviewers, or by a combination of random and systematic selection with probability proportional to population size.

The third stage. In the selected sites in urban areas and settlements in rural areas the lists of residential apartments and housing estates are compiled. From an actualized inventory of housing units HHs in urban and rural areas are randomly selected.

4 Statistical weighting

The methodology of weighting is based on the assignment for each individual unit corresponding statistical weight.

HH weights are calculated as reciprocal of overall sample probabilities:

$$B_i = \frac{1}{p_1 \cdot p_2 \cdot p_3}, \quad (2)$$

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 HH 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)). It is possible to use one of two ways: a) the a simplified method; b) iterative weighting (2 or more iterations).

A simplified method (SM) assumes the calculation of individual weights based on the size of age groups, separately for rural and urban areas:

$$k_{uij} = \frac{S_{ij}}{S_{bij}}, \quad (3)$$

where S_{ij} - individual weight *i*-th gender-are group in urban (rural) area of *j*-th region; S_{ij} - the size of *i*-th gender-are group in urban (rural) area in total population; S_{bij} - the size of *i*-th gender-are group in urban (rural) area, that has been selected within the region.

Iterative weighting (IW) involves several iterations:

Iteration I:

a) weights are calculated separately by sex, urban and rural areas;

b) the first correction coefficient (k1) is calculated; weighted variables are: region, sex, rural/urban;

c) the second correction coefficient (k2) is calculated; variables are: region, sex, 12 five-years groups.

Individual weights are equal within each region, five-year 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 five-year group:

$$K_i = B_b \cdot k_1 \cdot k_2 \cdot k_3, \quad (4)$$

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 Bb); S_{jt} – population size in *jt*-th sex-age rural (urban) group; S_{E2} – extrapolated population size in *jt*-th group (by Bb and k1); 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 don't exceed 7-8%: for the number of unemployed – 6%, number of employed – 1.8%, unemployment rate – 6.6%.

Table 5 – Indicators of sample representativeness . Mogilev region

Indicators	Characteristic value			Error			
	extrapolated, $Э_x$		in the general population, x	in absolute terms, $\Delta a = \lvert x - Э_x \rvert$		in % $\Delta = \frac{\lvert x - Э_x \rvert}{x}$	
	SM	IW		SM	IW	SM	IW
Number of employed, persons	50516	506231	515876	9360	9645	1.81	1.87
Urban area	400763	402333	412962	12199	10629	2.95	2.57
- Male	192868	194658	205508	12640	10850	6.15	5.28
- Female	207894	207675	207454	440	221	0.21	0.11
Rural area	105754	103898	102914	2840	984	2.76	0.96
- Male	57064	55228	55228	1836	0.3	3.32	0.0006
- Female	48690	48670	47686	1003	984	2.10	2.06
Total number of employed, persons							
- Male	249933	249885	260736	10804	10851	4.14	4.16
- Female	256584	256346	255140	1444	1206	0.57	0.47
Number of unemployed, persons	40624	40510	38511	2113	1899	5.49	4.19
Urban area	31995	32094	29332	2663	2762	9.08	9.42
- Male	19876	20046	18381	1495	1665	8.13	9.06
- Female	12120	12049	10951	1169	998	10.67	9.10
Rural area	8629	8416	9179	550	763	5.99	8.31
- Male	6065	5932	6572	507	640	7.72	9.75
- Female	2564	2485	2607	43	122	1.63	4.69
Number of unemployed (persons) among							
- Male	25940	25977	24953	987	1024	3.96	4.10
- Female	14684	14533	13558	1126	975	8.31	7.19

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).

5 Concluding remarks

The use of three-stage territorial sampling and iterative weighting provides very reliable information over larger number variables of LFS, conducted in Belarus. However, standard errors, calculated by the level of unemployment, the unemployed, in the context of gender-age groups at regional level are rather high (10-12%). Moreover, under a given load and a limited number of interviewers (200), it is not possible on a quarterly basis to question the estimated number of HH - 28000. On the basis of the selected annual array of HH (28000), built by regions, for each quarter, randomly generated four sub-samples are formed (each includes 7000 HH). If the annual array of information makes it possible to obtain a 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, the 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. It is possible to increase the number of iterations, to use alternative weighting schemes.

References

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

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

Multiple Indicator Cluster Survey Manual. Eurostat, 2009.

Metodika provedenia bazovyh obsledovanij naselenija. Kiev, 2008.