

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

Natallia Bokun,
Belarus State Economic University (BSEU),
Minsk, Belarus



Backgrounds of conducting

Current account of labour market in Belarus:

- differs from LFS methodology (survey object – job, not person);
- doesn't allow to regularly assess the actual level of unemployment (6-7 times higher than officially recorded);
- doesn't allow to regularly assess the employment by age groups, to assess underemployment, informal employment;
- lacks a number of indicators KILM (employment status, underemployment)

General

National Statistical Committee of the Republic of Belarus 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.

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**.

Sample size

Calculation of *sample size* takes into account factors:

- 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

Sample size

Key indicator – one of the most important indicators for the survey.

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

- Selection of two or three target populations that comprise small percentages of the total population (1-year, 2-year, 5-year age groups)
- Review of important indicator based on these groups, ignoring indicators that have very low (**less than 5%**) or very high (**more than 50%**) prevalence.
- Maximal indicator value, calculated for target group (**10-15% of the population**) is **15-20%**.
- Do not pick from desirable low coverage indicators an indicator that is already acceptably low.

Key indicator is the real unemployment rate (by the Census results).

Target groups are economically active populations (rural, urban, by regions, 5-year groups).

Sample size

The sample size formula is used:

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

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; p – proportion of the total population upon which the indicator (r) is based; n_h – average household size.

Sample 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

In the **first variant** a small surveyed group is the economically active population, **second** – 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 (w). In this case – to determine the sample size for each area the classic formula of the sampling theory is used, adjusted for deff, non-response and the number of persons aged 15-74 years per one HH in average.

$$n = \frac{t^2 \cdot w(1-w) \cdot f \cdot k_i}{\Delta_w^2 \cdot n_h'}$$

$t = 2$ ($p = 0.95$); Δ_w – defined relative factual error; n_h' – number of persons aged 15-74 years, falling by an average of one HH

Table I – 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
Economically 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 $\mu = 0.06$, relative limited error $\Delta = 0.12$, (without <i>deff</i>)	Relative standard error $\mu = 0.075$, relative limited error $\Delta = 0.15$, (with <i>deff</i>)
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, <i>N</i> , persons		Proportion unemployed in the population aged 15-74 years, <i>w</i>		Sample size, <i>n</i> , number of households	
	urban	rural	urban	rural	Relative standard error $\mu = 0.06$, relative limited error $\Delta = 0.12$, (without <i>deff</i>)	Relative standard error $\mu = 0.06$, relative limited error $\Delta = 0.12$, (with <i>deff</i>)
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

7,000 HH is planned to examine quarterly; **28,000 HH** – annually



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.

Sampling frame is based on the Census 2009 and includes:

- 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.

Sample design

Variables used for the stratification are: administrative districts, urban/rural.

The first stage. Towns and rural councils are selected. 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: $S_i = 40 \cdot (1/0.006) \cdot 2.43 = 16,200$ (towns with a population more than 17 thousand are included).

Other settlements 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) – 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



Sample design

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.

Statistical weighting

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

where B_i - HH weight; 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.

Weighting procedure

Individual survey person weights are based on iterative weighting: 1) simplified method; 2) iterative weighting

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

$$k_{ij} = \frac{S_{ij}}{S_{bij}},$$

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

Weighting procedure

Iterative weighting (IW) involves:

Iteration I

- a) weights are calculated by sex, five-year groups design
- b) the first correct coefficient (k_1) is calculated; variables of weighting are region, sex, rural/urban
- c) the second correct coefficient (k_2) is calculated; variables are region, sex, 12 five-years groups

Individual weights are equal within each region, five-year groups, settlements of one kind

Weighting procedure

Iteration II: follow adjustment of weights

Final individual weight for each five-year groups:

$$K_i = B_b \cdot k_1 \cdot k_2 \cdot k_3,$$

where:

$$B_b = \frac{S_j}{S_j}; k_1 = \frac{S_t}{S_E}; k_2 = \frac{S_{jt}}{S_{E2}}; S_j, s_j - \text{population size in } j\text{-th sex-}$$

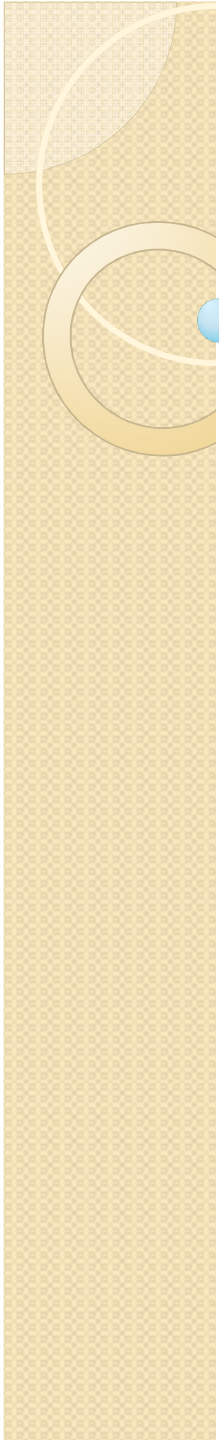
age group based on the result of the Census and survey; S_t – population size in t -th group by rural (urban), sex (on the Census data); S_E – extrapolated population size in t -th group (by B_b); S_{jt} – population size in jt -th sex-age rural (urban) group; S_{E2} – extrapolated population size in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

Table 5 – Indicators of sample representativeness . Mogilev region

Indicators	Characteristic value			Error			
	extrapolated, \mathcal{E}_x		in the general population, x	in absolute terms, $\Delta a = x - \mathcal{E}_x $		in % $\Delta = \frac{ x - \mathcal{E}_x }{x}$	
	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 experience of LFS in Belarus, has shown:

- the most optimal type of selection – a three-stage probability territorial sampling
- iterative weighting is used (HH weights and individual weights)
- the main survey problems: localization of sampling; non-responses; justification of the algorithm and the number of iterations
- 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)
- 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



**THANK YOU FOR
ATTENTION!**