Approved by the order of the Chairman of the Committee on Statistics of the Ministry of National Economy of the Republic of Kazakhstan

dated December 14, 2015

no. 202

**Household sampling methodology for the Living Standards Survey**

### General provisions

1. The methodology for constructing a sample of households for the survey of living standards (hereinafter - Methodology) was developed on the basis of the recommendations of the World Bank experts in the framework of the Joint Economic Research Project for 2010 "Development of a new sample for household surveys to assess living standards."
2. This Methodology is intended for use by the structural subdivisions of the Committee on Statistics of the Ministry of National Economy of the Republic of Kazakhstan in the formation of a sample of households for a survey of living standards .
3. The purpose of this Methodology is to construct a sample for a sample survey of households to assess the standard of living (hereinafter - the Survey), including a description of the stages of formation, stratification and calculation of sample weights.
4. This Methodology complies with the sampling methods accepted in the world practice and ensures the receipt of high-quality statistical data.
5. The survey is multipurpose in nature. Its main tasks are to obtain a wide range of data on the standard of living of the population, weights for the calculation of the consumer price index and data for the compilation of household sector accounts in the system of national accounts.
6. The following terms and abbreviations are used in this Methodology:
7. household - an economic entity consisting of one or more individuals living together, pooling all or part of their income and property, and jointly consuming goods and services;
8. interviewer - a person authorized to conduct a survey;
9. sample standard error ( SE ) is the standard deviation of the sample parameter value from the sample mean value of this parameter;
10. relative standard error ( RSE ) is the ratio of the estimated statistic to its mean.

**2. Definition of the study population (sample frame)**

1. The main source for the formation of a sample of households is the information system "Statistical register of housing stock", a component of the integrated information system "e-Statistics" (hereinafter - SRHF ).
2. The use of this database is due to a number of advantages, which include the following:

availability of a ready-made sampling frame, which eliminates one of the significant cost items associated with the compilation of the sampling frame;

the presence of a constantly updated database, which is necessary for the implementation of the rotation of households;

availability of information on households in the territorial context.

1. The general population includes households living in all types of dwellings, with the exception of those living in shared communal apartments, dormitories, nursing homes for the elderly and disabled, orphanages, prisons, hotels, religious communities and other similar dwellings.
2. The final sampling unit is the household, which is also the survey unit.

**3. Sample size**

1. The sample size is determined based on the principle of the optimal combination of costs and specified criteria for the accuracy of the results.
2. As indicators of the accuracy of statistical estimation, the standard error of the sample and the standard relative error of the sample are used.
3. SE is determined by the following formula:

(1)

where,

SE - the standard error of the sample;

δ - the dispersion;

Deff - the impact of the sample design (for urban and rural areas, they are 1.0 and 2.0, respectively);

n - the sample size.

1. About CO is determined by the following formula:

(2)

where,

RSE - the relative standard error of the sample;

SE - the sample standard error ;

 - the mean value of the variable used to estimate the value of the relative standard error.

The sample design impact (Deff) for urban and rural areas is 1.0 and 2.0, respectively. Based on this assumption, RSD are calculated. The assumption of Deff = 1.0 for urban areas is explained by the fact that the urban clusters of the survey are close to random sampling, since large cities are not subdivided into smaller territorial units. The assumption of Deff = 2.0 in rural areas is based solely on the experience of other countries.

1. The sample size ensures obtaining results with an error of no more than 4% at the national level and no more than 7% at the regional level, according to Table 1, given in the appendix to this Methodology. Thus, the sample size is set at 12,000 households (0.3% of the total population).

**4 . Algorithm for the formation of a sample population**

1. The sample set of households is formed by the method of two-stage probabilistic (random) sampling using stratification and random selection procedures at each of the sampling stages. The stratification procedure is aimed at forming a representative sample of households that adequately reflects the territorial features of the population stratification.
2. The sampling process is carried out in 2 stages.

At the first stage, the general population is stratified according to the territorial basis, including the distribution into urban and rural areas. Thus, 30 strata are formed - these are selected urban and rural areas in sixteen regions of the country (a total of 30 strata, taking into account that there are no rural areas in the cities of Astana and Almaty).

1. 400 territorial units are selected as primary sampling units (hereinafter - PSU), which are urban and rural areas and are defined by the first six (6) characters of the code according to the Classifier of Administrative Territorial Objects (hereinafter - CATO ).
2. The volume of work of each of the interviewers is 30 households during one quarter. Each group of 30 households is called a cluster and corresponds to the workload of one interviewer during one quarter.
3. The first stage of sampling includes two consecutive procedures. The initial procedure involves determining the PSU within each stratum with a probability proportional to size (hereinafter – PPS), that is, the number of PSUs in the stratum is formed depending on the number of households present in the stratum.
4. The probability of selection ( P hi) in the choice of primary sampling units (hi) in the stratum (h) is given by the following formula:

(3)

where,

P hi - the probability of selection;

sh - the number of PSUs selected in stratum h;

nhi - the number of households in PSU hi, according to SRHF;

Nh - the total number of households in the stratum, according to SRHF data.

1. The following procedure aims to achieve the optimal value of the relative standard error for each of the 30 strata (400 PSU).
2. To estimate the magnitude of the relative standard error, indicators considered to be the most important for the survey are used as a variable. According to the survey of living standards, such indicators include “average per capita income used for consumption”, “average per capita cash income”, “average per capita cash expenditure”, etc.
3. The characteristic of assessing the accuracy of the indicator "average per capita income used for consumption" is given in Table 2, according to the appendix to this Methodology. According to the calculations, the RSD of the sample for the country does not exceed 0.3%, for the regions - no more than 2.2%.
4. The number of selected PSUs and the number of households in each PSU, as well as their distribution by region, are shown in Table 3, according to the appendix to this Methodology.
5. When conducting the actual selection of PSUs in cities, it must be taken into account that the SRHF does not divide large cities into small territorial units. For example: the city of Kokshetau with a population of over 100,000 is represented in the SRHF as a single line, without further disaggregation. Given the relatively large size of these PSUs, many of them are selected more than once as a result of the standard PPS procedure .
6. The first and last lines of the list of selected primary sample units are presented in Table 4 according to the appendix to this Methodology. The 'Selected' column indicates PSUs in which the standard PPS procedure was performed more than once. The first line shows that the city of Kokshetau was indeed chosen 5 times. Thus, in the city of Kokshetau it is necessary to select 5 clusters (150 households).
7. At the second sampling stage, 30 households are randomly selected from each PSU. The sampling frame for the second step is the list of individual dwellings in the PSU. The cluster (or clusters) of dwellings to be visited in the survey is selected with equal probability from among the eligible dwellings in the PSU.
8. The probability (phij) of choosing a household (hij) in the PSU (hi) of a stratum (h) is given by the following formula:

(4)

where,

phi - the probability of choosing PSU hi is given by equation (1);

mhi - the number of required dwellings in PSU hi (normally always 30);

n'hi - the total number of eligible dwellings in the PSU.

1. In principle, each selected dwelling is expected to accommodate one household. If it is found that there is more than one household in a given dwelling, then one of them is randomly selected. If the dwelling turned out to be empty, then you need to apply the method, which is discussed later in the section “Sampling bias”.
2. To test the performance of interviewers in urban areas, it is recommended to use the method of interpenetrating sampling.
3. In large cities, where it is necessary to have k number of clusters (for k > 1), first 40k number of dwellings will be selected by systematic equiprobability sampling (hereinafter – SEPS), and then a random permutation of digits from 1 to k will be repeated 40 times to divide ( city) for the number of clusters k – one for each of the k number of interviewers in the city.
4. In cities, the households in each cluster will not be located in a specific area, which makes it possible to check the quality of the work of the interviewers (for example: you can check the number of survey refusals in the same area of the city by different interviewers).

**5 . Sample segmentation**

1. For the practical application of the sample, an important factor is the geographical location of the selected survey units from each other within the same cluster, since their remoteness from each other creates additional costs and inconvenience during the survey. Some of the selected PSUs are too large in population and survey units are too geographically dispersed from each other.

For example: in the rural area Ontustik Kazakhstan region, one of the PSUs ( CATO - 515437) consists of 14 settlements with a total number of households equal to 3183 and the distance between them reaches about 25 km.

1. If such cases arise, it is possible to adjust the sample through a process of segmentation (splitting or disaggregation).
2. First, in PSU, three localities are selected by SEPS, taking into account the number of SRHF households inhabited in them, and then 10 households are selected in each of the three localities, also by SEPS. 10 reserve households (for replacement) are distributed among three selected localities.
3. This procedure is an additional sampling step and allows for a more clustered sample, yet it will not change the selection probabilities or change the weights of the sampled households. Therefore, formulas (4) and (5) apply unchanged in disaggregated (segmented) PSUs.

**6 . Sample bias**

1. There are cases during the survey when it is not possible to interview households due to the fact that the dwelling is not found, or not occupied, or the household refuses to participate in the survey. Opt-out is undesirable because it reduces the sample size and mainly because it is a source of potential sample bias and leads to skewed statistical results.
2. An effective way to overcome the problem of missing data is to carefully document each case.
3. If it becomes difficult to comply with the principle of “no replacement”, a list of reserve households should be provided in the amount of 1/3 of the number of households in each cluster (10 households per interviewer).
4. A total of 40 dwellings are selected in each cluster: 30 in the nominal sample and 10 substitutes. These 40 dwellings are selected from among all eligible dwellings in the PSU by SEPS, and then a randomly selected cyclic permutation of the numbers 0, 1, 2, and 3 is repeated ten (10) times to divide the cluster into four groups of 10 dwellings. Dwellings in groups 1, 2 and 3 will constitute the nominal sample, while dwellings in group 0 will be held in reserve for possible replacement.
5. The selection of reserve households from the additional list is carried out on a bottom-up basis. That is, out of 10 households, households are sequentially selected for replacement: from the beginning - the tenth in a row (the last in the list), then - the ninth in a row, and so on.

**7 . Weighing**

1. In order to obtain general population data, statistical weighting of survey results is performed. The implementation of this method is carried out by assigning to each surveyed household a statistical weight that characterizes the total number of households represented by the part that fell into the sample. The weights for indicators of the standard of living of the population are calculated quarterly.
2. The weights are calculated using SRHF data on the distribution of surveyed households separately by urban and rural population in a regional context.
3. The probability weight (whij) of a household in PSU (hi) of stratum hi (h) is the reciprocal of its selection probability phij and is given by the following formula:

(5)

where,

W hij – household weight ;

P hij - the reciprocal value of the selection probability;

Nh - the total number of households in the stratum, according to SRHF data;

n'hi - the total number of suitable dwellings in PSU;

S h - the number of PSUs selected in stratum h;

mhi - the number of required dwellings in PSU hi (normally always 30).

Because n'hi ≈ n'hi and mhi ≈ 30 across all PSUs, the sample will be roughly self-weighted within each stratum.

1. The sum of the "base" weights provides an estimate of the number of all households in a given region and the country as a whole. However, their use makes it possible to maintain the compliance of the sampling population with the original principles of sampling, while in practice there are cases that violate this compliance (refusals to participate in the survey and other cases of non-receipt of data).
2. Under-response from sampled households is a serious shortcoming that skews survey results. Under these conditions, it becomes necessary to compensate for the missing data. This procedure is designed to carry out the calculation of indicators, i.e. compensate for the lack of sample survey data.
3. Underreporting occurs when the sample unit does not participate in the survey at all, or takes part in it. To compensate for total non-response, a simple weight adjustment scheme is applied by assigning large weights to all responding households in a given locality. The weights of all responding households in a given locality are increased by the same factor. For example: if 90 percent of households in PSU answered all questions, the weights for all respondents are increased by a factor of 1/0.9=1.11. All non-responding households are excluded from the sample by assigning an actual weight of zero for each household.
4. In the absence of data only on certain issues, the method of imputation (imputation) of values is applied. This method consists of replacing missing data for an individual item with a value that is predicted based on other information available for the item or other items in the survey.
5. The calculated personal weights are recorded as additional variables in the database for each household and are used in the formation of summary results for the corresponding quarter at the republican and regional levels. Disseminated data on average for the year are obtained by combining quarterly data.

**8 . Sample rotation**

1. In order to avoid the effect of household fatigue from participation in the survey, the sample of households is periodically rotated. The process of systematic sample rotation does not involve the replacement of households that have refused to participate in the survey.
2. In order to track the impact of seasonal fluctuations and other phenomena on the same households during the reporting year, it is advisable to rotate at the beginning of the year, that is, every year in December it is necessary to update the sample.
3. The rotation of households (updating the sample) must be carried out annually in the amount of 1/3 of the total number of surveyed households. That is, every year 1/3 of households are withdrawn from the sample and replaced by others. Thus, after 3 years, the sample will be completely updated.
4. At the end of the year (December), 10 households in each cluster are randomly selected and removed from the sample. Another 10 of the initially selected households are randomly selected and removed from the sample at the end of the following year, and the remaining 10 households at the end of the following year. Each time, 10 new households are randomly selected from the updated SRHF database to take the place of the withdrawn ones.
5. When forming a sample in the list of selected households, an additional field with a sign of rotation is provided. A "1" in this field indicates that the household is subject to rotation after the first year, "2" after the second year, and "3" after the third year. Thus, systematic rotation is carried out by removing groups 1, 2 and 3 from the sample at the end of the first, second and third years.

Appendix

to the Household Sampling Method for the Living Standards Survey

**Table 1 . Determining the sample size**

|  |  |  |  |
| --- | --- | --- | --- |
| Error value | | Sample size:number of households,thousand units | Expenses for the examination  million tenge |
| by republic | by region |
| Option number 1 | | 24 | 800 |
| < 1 % | < 3 % |
| Option number 2 | | 18 | 600 |
| < 2 % | < 5 % |
| Option number 3 | | 12 | 400 |
| < 4 % | < 7% |

**Table 2 . Characteristics of the assessment of the accuracy of the indicator “income used for consumption” based on the results of a household survey on the assessment of living standards**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Income used for consumption  per year on average per capita, tenge | Sampling Standard Error  (Se) | Confidence interval value (Р=0.95), ± tenge | | Relative standard error  (RSE), % |
| lower limit | upper limit |
| Republic of Kazakhstan | 458743 | 1191 | 456409 | 461077 | 0.26% |
| including: |  |  |  |  |  |
| urban area | 552761 | 3613 | 545681 | 559841 | 0.65% |
| rural area | 360228 | 2558 | 355214 | 365242 | 0.71% |
|  |  |  |  |  |  |
| Akmola | 470134 | 9124 | 452252 | 488016 | 1.94% |
| Aktobe | 472494 | 8087 | 456645 | 488343 | 1.71% |
| Almaty | 535466 | 8797 | 518224 | 552707 | 1.64% |
| Atyrau | 415869 | 6907 | 402330 | 429407 | 1.66% |
| Batys Kazakhstan | 416992 | 8043 | 401228 | 432756 | 1.93% |
| Zhambyl | 355963 | 5464 | 345254 | 366671 | 1.53% |
| Karaganda | 562665 | 10796 | 541505 | 583824 | 1.92% |
| Kostanai | 462741 | 9229 | 444652 | 480829 | 1.99% |
| Kyzylorda | 393944 | 6492 | 381221 | 406668 | 1.65% |
| Mangystau | 449044 | 8410 | 432560 | 465528 | 1.87% |
| Ontustik Kazakhstan | 307032 | 3564 | 300047 | 314016 | 1.16% |
| Pavlodar | 465175 | 7636 | 450208 | 480141 | 1.64% |
| Soltustik Kazakhstan | 461012 | 10119 | 441179 | 480844 | 2.19% |
| Shygys Kazakhstan region region | 492494 | 9188 | 474487 | 510502 | 1.87% |
| Astana city | 567388 | 9060 | 549630 | 585145 | 1.60% |
| Almaty city | 701398 | 11007 | 679825 | 722971 | 1.57% |

**Table 3 . Distribution of primary sampling units by strata (urban and rural)**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Code | Region name | Number of households | | | Number of PSUs | | |
| City | Village | Total | City | Village | Total |
| 11 | Akmola | 115 888 | 79 089 | 194 977 | 12 | 16 | 28 |
| 15 | Aktobe | 133 540 | 32 803 | 166 343 | 12 | 16 | 28 |
| 19 | Almaty | 110 045 | 260 502 | 370 547 | 8 | 16 | 24 |
| 23 | Atyrau | 56 823 | 31 931 | 88 754 | 10 | 8 | 18 |
| 27 | Batys Kazakhstan | 100 630 | 76 727 | 177 357 | 8 | 14 | 22 |
| 31 | Zhambyl | 123 593 | 117 878 | 241 471 | 9 | 14 | 23 |
| 35 | Karaganda | 378 012 | 66 854 | 444 866 | 20 | 12 | 32 |
| 39 | Kostanai | 179 666 | 127 047 | 306 713 | 12 | 15 | 27 |
| 43 | Kyzylorda | 55 226 | 69 545 | 124 771 | 8 | 12 | 20 |
| 47 | Mangystau | 73 270 | 16 828 | 90 098 | 12 | 8 | 20 |
| 51 | Ontustik Kazakhstan | 232 170 | 260 099 | 492 269 | 10 | 16 | 26 |
| 55 | Pavlodar | 190 793 | 63 953 | 254 746 | 12 | 16 | 28 |
| 59 | Soltustik Kazakhstan | 97 757 | 114 127 | 211 884 | 9 | 13 | 22 |
| 63 | Shygys Kazakhstan | 299 061 | 171 035 | 470 096 | 14 | 16 | 30 |
| 71 | Astana city | 148 587 | - | 148 587 | 22 | - | 22 |
| 75 | Almaty city | 386 251 | - | 386 251 | 30 | - | 30 |
|  | Total | 2 681 312 | 1 488 418 | 4 169 730 | 208 | 192 | 400 |

**Table 4 . First and last lines of the list of selected primary sampling units (PSUs)**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Code | Region name | District name | Locality name | Urban/Rural | CATO | Apartments | D/H | People | Probability | Selected |
| 111010 | Akmola | Kokshetau | Kokshetau | 1 | 1 | 45 645 | 45 003 | 109 117 | 4.65998 | 5 |
| 111810 | Akmola | Stepnogorsk | Stepnogorsk | 1 | 1 | 21 597 | 21 084 | 46 928 | 2.18321 | 2 |
| 113220 | Akmola | Akkol | Akkolskaya g.a. | 1 | 1 | 4403 | 4400 | 13 231 | 0.45561 | 1 |
| 113820 | Akmola | Atbassar | Atbassar | 1 | 1 | 9 502 | 9 428 | 27 121 | 0.97625 | 1 |
| 114620 | Akmola | Yereymentau | Yereymentau | 1 | 1 | 4034 | 4029 | 10 841 | 0.41720 | 1 |
| 117020 | Akmola | Burabai | Shchuchinskaya g.a. | 1 | 1 | 13 875 | 13 674 | 33 903 | 1.41592 | 2 |
| 113239 | Akmola | Akkol | Ivanovsky | 2 | 1 | 374 | 360 | 1 249 | 0.07283 | 1 |
| 113433 | Akmola | Arshalynsky | a.o. Zhibek Zholy | 2 | 4 | 559 | 558 | 2050 | 0.11289 | 1 |
| 636473 | Shygys Kazakhstan | Urjar | a.o. Makanshy | 2 | 1 | 3 117 | 2773 | 11 648 | 0.25941 | 1 |
| 711110 | Astana city | Astana city | Almaty | 1 | 1 | 79 627 | 81 274 | 256 464 | 12.03354 | 12 |
| 751110 | Almaty city | Almaty city | Almaly city | 1 | 1 | 65 939 | 66 434 | 168 138 | 5.15991 | 5 |
| 751910 | Almaty city | Medeusky | Turksibsky | 1 | 1 | 49 333 | 49 735 | 151 982 | 3.86290 | 4 |