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

dated June 30, 2016

no. 132

**Methodology for constructing a sample of households for a population employment survey**

**Chapter 1. General provisions**

1. Methodology for constructing a sample of households for the survey of employment (hereinafter – Methodology) refers to the statistical methodology, formed in accordance with the [Law](http://adilet.zan.kz/rus/docs/Z100000257_#z0) of the Republic of Kazakhstan "On State Statistics".

2. This Methodology is applied by the Bureau of National Statistics of the Agency for Strategic Planning and Reforms of the Republic of Kazakhstan to form a sample of households for the population employment survey.

3. The purpose of this Methodology is to construct a sample of households for the sample survey of employment of the population (hereinafter – SSEP). The methodology defines the main aspects and methods of sampling taking into account the district level, including a description of the stages of formation, stratification and calculation of sample weights.

 4. The Methodology uses concepts in the meanings defined in the Laws of the Republic of Kazakhstan "On State Statistics", "On Employment", as well as definitions adopted by the International Conferences of Labor Statisticians (hereinafter – ICLS) and recommendations of the International Labor Organization (hereinafter – ILO).

**Chapter 2. Population (Sampling Frame)**

5. The information system "Statistical register of housing stock", a component of the integrated information system "e–Statistics" (hereinafter – SRHF ) is used as the general population for the formation of a sample population of SSEP households, taking into account the district level.

6. This database has a number of advantages:

the presence of a ready–made sample frame;

availability of a constantly updated database;

availability of information on households in the territorial context.

7. SRHF accounting units are all residential buildings and residential premises located on the territory of Kazakhstan. These include apartments, single–family (individual) houses, two–family houses, three or more apartment buildings.

8. The SRHF database consists of two tables, house – the main table, contains information on the whole residential building, apartment – additional table, contains information about each apartment of the residential building.

 9. Features of the formation of the general population:

1. the general population includes all types of residential premises, with the exception of common communal apartments, hostels, medical and social institutions (organizations) for the elderly, persons with disabilities, children with disabilities in a hospital, orphanages, prisons, hotels, religious communities and other similar residential premises;
2. in order to reduce the burden on interviewers when collecting information, settlements with less than 100 households are excluded from the general population.

**Chapter 3. Sample design**

10. SSEP, taking into account the district level, is carried out in all regions of Kazakhstan on the basis of a sampling method of observation, with subsequent distribution of the results to the entire population of the country of the surveyed age (15 years and older).

11. The SSEP organizational model at the district level is based on an overlay sample, generated once a year and evenly spread over twelve months to reduce the burden on interviewers.

12. A household sampled in the current year has an equal probability of being selected again in the next year's sample. The imposition is carried out in the amount of the part calculated annually from the total number of households.

13. For the purpose of smoothing out jumps in monthly employment survey data, a monthly overlay of the annually calculated portion of households is envisaged. Every month, starting from February, an annually calculated part of households move from the previous month to the next month (surveyed two months in a row), and the rest are new every month.

14. Monthly formation of the catalog for overlay is carried out automatically at the territorial level in the software package based on the catalog of entered reports for the reporting month and the specified number of households Z (calculated by the Department of Statistical Registers and Classifications when forming the sample population for the current year) by using the Nn step , where n is the month in which the selected households are used. Step Nn = Notc/Z, where Z is the number of households defined for the overlay for a given region, constant for all months of the reporting year, Notc is the number of households that actually reported in the current month.

15. This model provides:

1. representative annual totals based on the pooling of all twelve monthly samples:

for the Republic of Kazakhstan;

by regions;

by regions.

1. representative quarterly totals based on the combination of three monthly samples included in the quarter:

for the Republic of Kazakhstan;

by regions.

1. representative monthly totals based on a monthly sample:

for the Republic of Kazakhstan.

1. smoothing jumps in monthly data.

16. The primary unit of the survey is residential buildings (individual houses, apartment buildings), all apartments in the house that are included in the sample are included in the survey.

The secondary unit of the survey is households (persons aged 15 and over who live in them).

The sample set of households is formed by the method of two–stage stratified cluster sampling.

The formation of the sample population is carried out centrally at the republican level in the context of the regions of Kazakhstan a.

**Chapter 4 Stratification of the general population**

17. Stratification is a powerful technique for improving the efficiency of the sampling method. The advantages of stratification are:

greater accuracy compared to other selection options for the same data;

greater accuracy when using a smaller sample size, which saves money;

ensuring the representativeness of the sample.

18. The need for stratification of the general population is due to the heterogeneity of the units of observation in terms of characteristics.

19. The following stratification procedure is used to sample the SSEP at the district level:

the general population is stratified by districts, by type of locality (urban, rural), by classes of building dimensions (1 apartment buildings, 2–20 apartment buildings, 21–80 apartment buildings, more than 80 apartment buildings);

the size of the general population in each stratum is determined by summing up the number of houses.

20. Direct selection of houses of different types is carried out with the help of specialized software using the random selection method. The use of this software product reduces the amount of work and gives a guaranteed high–quality result.

If several households actually live in the selected addresses, then one of them is randomly selected.

After the selection, lists of households included in the sample are formed for each region.

**Chapter 5 Determining the Sample Size**

21. Sample size is the total number of observation units in the sample. The sample size depends on the size of the relative sample error, which with a certain probability provides a given accuracy of the observation results.

22. To determine the optimal sample size for SSEP needed to estimate the population to a given accuracy, the following formula is used:

 **** (1)

where,

k – the argument of the Laplace function ( k = 1.96 for 95% confidence level);

N – the volume of the general population;

 – dispersion;

E – the absolute error.

23. Population variance is defined as the mean of the squared deviations of all individual observations from their mean.

Population variance:

 (2)

###### If the error is expressed as standard error ( μ), you can use the following formula:

 **(3)

where, RSE – the relative standard error of the sample.

If we do not take into account the correction for the finite population, the formula will look like this:

 **(4)

24. In determining the sample size for the current year for the SSEP, the calculation of the relative sampling error of the previous year's SSEP results is used. Relative sampling error is inversely proportional to sample size, with a fourfold increase in sample size, errors are halved.

25. The main requirement for the results of the monthly survey is to obtain representative data at the district level with a standard sampling error of 5–7% for the "unemployment rate" indicator.

Taking into account the requirements, the optimal sample size of 5% was calculated and adopted.

**Chapter 6. Compensation for non–responses**

 26. Under–response from sampled households is a serious shortcoming that skews survey results. Under these conditions, SSEP hiccups the need to compensate for missing data. This procedure is designed to carry out the calculation of indicators, to compensate for the lack of data from a sample survey.

27. During the survey, there are cases where it is not possible to interview households. There are two types of reasons objective and subjective.

Objective reasons for refusal :

all household members aged 73 and over;

house is destroyed;

left , empty apartment (house);

association of apartments;

house for demolition;

house (address) not found;

another reason (change of purpose of the premises).

Subjective reasons for refusal :

refusal of the household to participate in the survey.

28. Substitution of addresses of residential premises in the absence of household members or refusal to participate in the survey is not carried out.

The interviewer interviews household members living at the addresses only according to the lists provided in the sample.

28–1. If there are objective reasons for the non–response during the household survey, reserve lists are compiled. A back–up frame is drawn up by sampling specialists for possible replacement of households due to objective reasons for non–response. The reserve sample is compiled in the same way as the main sample.

Substitution of addresses of dwellings is allowed only from the reserve list with the household that is specifically intended to serve as a substitute or replacement for a non–responding household for objective reasons. This is done to avoid substitution for a "convenient" household, which increases sampling error.

29. When processing cases of non–provision of data, non–received answers for objective reasons are not compensated. Only cases of non–received answers due to subjective reasons are subject to compensation.

For these processing purposes, a reweighting method is applied, which consists in adjusting the sample weights.

30. In order to obtain data that are disseminated to the general population, statistical weighting of the survey results is carried out.

Weighing the results of a sample survey is carried out by assigning an appropriate weight to each individual unit of observation – a person.

31. The final individual weight K is the product of the base weight F and the compensation and extrapolation factors.

The individual weight calculated for each respondent is entered into the individual database as a spread multiplier and is used to generate spread data for the entire population of the surveyed age, gender, and location type for any survey program indicators.

32. Weights for employment indicators are calculated monthly. To calculate the weights, SRHF data on the distribution of surveyed households separately by urban and rural population in the regional context are used.

33. When calculating individual weights, the method of iterative weighting of the sample in relation to the total population in the period closest to the reference period (the critical week of the survey) is applied.

The procedure consists in comparing the sample population, divided into groups taking into account gender, age and regional characteristics, with the entire population, distributed according to the same characteristics.

34. When processing the results of a monthly survey, the calculation of individual weighing weights (extrapolation coefficients) is carried out sequentially in several stages.

35. In order to equalize random and non–random systematic errors that are inevitable in sampling:

 alignment of known cases of failed surveys – compensation;

 distribution to the general population up to SRHF indicators is carried out – adaptation or extrapolation.

For this, the corresponding factors, the compensation factor and the extrapolation factor are calculated.

**Paragraph 1. Compensation**

36. With the help of this procedure, the additional calculation of indicators for households, with subjective reasons for the refusal of the survey , is carried out, and the missing data from the sample survey are compensated.

37. Before carrying out this procedure, sum up the number of households to be surveyed (S 1 ), the number of households actually interviewed (S 2 ) and the number of households not surveyed with an indication of the reason (Sp).

The number of non–respondents is divided into groups – for objective reasons (Sp 1 ) and for subjective reasons (Sp 2 ):

objective reasons are related to the impossibility of conducting an examination, due to the destruction (demolition) or merger of the dwelling itself, death, absence for a long time or departure to a new place of a member of the household to be examined and other unforeseen circumstances;

subjective reasons include the absence of residents at the time of the survey, the refusal of an individual member or the entire household from the survey.

38. With the help of the compensation procedure, the additional calculation of indicators for households falling under the subjective reasons for refusing to survey is carried out, and the missing data of the sample survey are compensated. For this, an intermediate compensating factor (weight) K1 is calculated

 *K 1 = 1 + Sp 2 / S 2*(5)

where,

K1 – compensating factor (adjustment factor);

Sp 2 – the number of households not surveyed for subjective reasons for not surveying;

S 2 – the number of households actually surveyed.

39. An intermediate compensating factor is calculated with an accuracy of five decimal places for each district separately and is assigned to each respondent in this district, regardless of his gender and age.

f ) – adjusted for the compensating factor ( K 1 ).

Adjusted for the base weight, the compensating factor is used to determine the size resulting from bringing the number of households to the survey (S 1 )

 *S 1 = S 2 \* K 1* (6)

40. To compensate for total non–response, a weight adjustment scheme is applied by assigning large weights to all responding households in a given locality. The weights of all households that answered questions in a given locality are increased by the same factor. All non–responding households are excluded from the sample by assigning an actual weight of zero to each.

**Section 2. Extrapolation**

41. The dissemination of survey sample data is based on assigning an appropriate individual weight to each individual unit of observation, the household.

To do this, a comparison of survey data on a sample population (the number of surveyed citizens), stratified according to gender, age and regional characteristics, is carried out with the general population according to current demographic calculations, stratified according to the same characteristics.

42. In general, the formula for calculating the adaptation factor (weight) is

 *Vg=Wg\*N/Ng* (7)

where,

Vg – weight on the attribute g;

Wg – the proportion of the population in the general population, with the characteristic g;

N – the total number of respondents;

Ng – the number of respondents, with characteristic g.

43. The weighting procedure is carried out on the basis of data on the structure of the population according to the existing population, used as a general population, only within regional strata (strata) by gender and age. For each respondent, a system of adaptation factors (weights) is calculated according to the following criteria:

territorial structure (district);

terrain type;

gender (men and women);

6 age groups (6–10 years old; 11–14 years old; 15–34 years old; 35–54 years old; 55–71 years old; 72 years old and above).

44. To calculate the basic individual weight, the following formula for calculating the adaptation factor is used:

 *K 2 k = S k / S 1 k* (8)

where,

K 2 – adaptation factor;

S – the population in the general population;

S 1 – the population of actually surveyed households;

K – a distinguishing feature depending on the total characteristics of the person for whom the factor is calculated.

45. The final individual weight (or extrapolation factor) is the product of the compensation factor and the adaptation factor.

 *K = K 1 \* K 2 k ,* (9)

where,

K – individual weight (extrapolation coefficient);

K*1* – compensating factor (adjustment factor);

K*2 k* – an adaptation factor.

At the same time, if the compensation factor for all members of a given household has the same value, then the adaptation factors, as well as the final extrapolation factor (individual weight), respectively, are different.

46. The calculated individual weights are recorded as additional variables in the individual data base for each respondent and are used in the formation of summary results for the corresponding month, quarter, year at the republican, regional, district levels.

The disseminated data for the current year average is obtained by pooling the monthly samples for the year, and the disseminated quarterly data is obtained by combining the monthly samples for the quarter.

**Chapter 7 Sample rotation**

47. Rotation of households (updating the sample) is carried out annually at a rate of 100%, every year the entire sample is replaced by new households.

**Chapter 8 Assessment of the accuracy of indicators**

48. Sampling standard error and standard relative sampling error are used as measures of statistical estimation accuracy.

49. Sampling standard error is the standard deviation of the value of a sample parameter from the sample mean of that parameter.

The relative standard error is the ratio of the estimated statistic to its mean.

50. The accuracy of the survey by district is calculated once a year in terms of the number of employed, the number of unemployed, the labor force, the number of persons not included in the labor force, the share of the labor force in the population, the unemployment rate.

51. Sampling standard error and standard relative sampling error are used as measures of statistical estimation accuracy.

52. The standard error of the sample ( μ) determines the possible discrepancies between the characteristics of the sample and the general population. The value of the standard error of the sample is determined by the formula:

$μ=\sqrt{\frac{σ^{2}}{n}\left(1-^{n}/\_{N}\right)}$(1 0 )

where,

– general dispersion;

– sample size;

N – the volume of the general population.

53. Based on the results of the SSEP, the status of the respondent is determined from the answers of the respondents to questions (for example : yes or no ). Variables determined by the choice of one of two alternative answers are binary. According to the probability theory for a binary variable, the variance is calculated. To calculate the relative standard error of SSEP, the proportion of the responses (yes or no) is used. The variance of the proportion is determined, then the standard relative error is calculated.

The standard error of the sample fraction ( μg) is determined by the formula:

 (11)

where:

w – the share of answers (the share of "success") in percent;

– sample size;

N – the volume of the general population.»;

add paragraph 54 with the following content:

“54. Often it is not the absolute values of the standard error that need to be considered, but its value in relation to the statistic being estimated. The relative standard error (hereinafter – RSE) is calculated as the ratio of the estimated statistic to its mean.

RSE – determined by the following formula:

*RSE = SE / *(12)

where,

RSE – the relative standard error of the sample;

SE – the standard error of the sample;

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

 *= /* $N\_{i}$(13)

where,

$N\_{i}$– the number of elements of the general population i–stratum;

$x\_{i}$– the index of the i–stratum.