Approved by the order of the Chairman of the Committee on Statistics of the Ministry of National Economy of the Republic of Kazakhstan dated November 11, 2016

no. 266

**Methodology for conducting a sample survey of households**

**Chapter 1. General provisions**

1. Methodology for conducting a sample survey of households (hereinafter - Methodology) refers to a statistical methodology formed in accordance with international standards and approved in accordance with the Law of the Republic of Kazakhstan dated March 19, 2010 " [On](http://adilet.zan.kz/rus/docs/Z100000257_#z0) State Statistics" (hereinafter - Law).

2. Method sets the main aspects and methods of analysis of the sample and general population of households and is intended for use by the structural divisions of the Committee on Statistics of the Ministry of National Economy of the Republic of Kazakhstan (hereinafter - the Committee).

3. A subset of households is selected for sample surveys, and observations are made or data are collected within this subset. The results obtained are extrapolated (distributed) to the entire population as a whole.

4. Main application advantages sampling method in modern statistics are:

1) reducing the time for statistical observations (surveys) ;

2) reducing the information load on respondents;

3) significant savings in labor costs, material and financial resources for conducting the survey;

4 ) greatly accelerated receiving research results compared to solid examination.

5. This Methodology uses concepts in the meanings defined in the Law , as well as the following definitions:

1. **panel method of observation - a method of collecting information in which** a certain group of units of analysis is periodically polled for a relatively long time , and the subject of the study remains constant;
2. general population - a complete group of all units of analysis, whose characteristics are to be assessed ;
3. representativeness - the correspondence of the characteristics [of the sample](http://ru.wikipedia.org/wiki/%D0%92%D1%8B%D0%B1%D0%BE%D1%80%D0%BA%D0%B0) to the characteristics [of the population](http://ru.wikipedia.org/wiki/%D0%9F%D0%BE%D0%BF%D1%83%D0%BB%D1%8F%D1%86%D0%B8%D1%8F) or [the general](http://ru.wikipedia.org/wiki/%D0%93%D0%B5%D0%BD%D0%B5%D1%80%D0%B0%D0%BB%D1%8C%D0%BD%D0%B0%D1%8F_%D1%81%D0%BE%D0%B2%D0%BE%D0%BA%D1%83%D0%BF%D0%BD%D0%BE%D1%81%D1%82%D1%8C) population ;
4. mathematical expectation - the average value of a particular characteristic in all possible samples, as well as the weighted average of all possible results with a weight of probabilities reflecting the possibility of occurrence in each result ;
5. parameter - a quantity calculated from all values in the population set, that is, a descriptive measurement of the population ;
6. stratum - division into special layers of units (respondents) with the same or similar indicators;
7. sampling plan - a set of specifications that define the general population and units of the sample, as well as the degree of probability of possible samples;
8. selective population (sample) - a set of cases (subjects, objects, events, samples), using a certain procedure, selected from the general population for participation in the study;
9. sample size - the total number of observation units in the sample.

**Chapter 2. Process of planning and sampling**

6. When planning a survey, it is necessary to determine the geographic areas to be covered and the population to be surveyed.

7. When determining the statistical population, it is necessary to identify the population group from which the sample is formed. Outlying areas with few households or residents are removed from the sampling frame because their coverage is too expensive. They represent only a small fraction of the population, their impact on population figures is very small. The report on the results of such a survey clearly indicates the exclusion of these areas.

8. The process of forming a sample for conducting a survey consists of several stages:

definition of the general population ;

establishing a sampling frame;

choice between probabilistic and improbable methods of selection;

definition of sampling plan;

determination of the sample size;

direct sampling according to the plan.

**Chapter 3. Population definition and sampling frame**

Population census data or information system The Statistical Housing Register (hereinafter - IS SHR) is the main source for the sampling frame for household surveys in the Republic of Kazakhstan. Census data serve as a means of providing information on the size, composition and geographical distribution of the population, in addition to socioeconomic and demographic characteristics. The population census collects information on each individual in the household and on each set of dwellings throughout the territory. To avoid cases If data is not received from respondents, IS SHR is used to form the sample. IS SHR was created to generate and collect data on dwelling houses and dwellings for housing stock statistics and sampling for household surveys.

10. Accounting units in IS SHR are all residential buildings and residential premises (apartments) located on the territory of the Republic of Kazakhstan.

These include:

residential premises (apartment);

single-family (individual) house;

semi-detached house;

three or more apartment buildings.

Each house and apartment has an identification number (hereinafter - ID ).

In addition, IS SHR contains the following data: apartment ID, classifier of administrative territorial objects (CATO), street, house number, apartment number, total area, living area. The data contained in IS SHR is updated daily.

**Chapter 4. Sampling strategy and methods**

11. To select elements from the general population, the following methods of probabilistic selection are used :

simple random sampling (hereinafter - SRS);

systematic random sampling (step sampling) ;

sample with probability proportional to size (hereinafter - SPPS).

12. Simple random selection (sample) provides an equal probability of being selected for each element of the general population. There are the following varieties of this method:

repeated random selection;

repetitive random selection.

13. Non-repeated random selection gives more accurate results of sample observation compared to repeated, since with the same sample size, observation covers more units of the general population. In cases where non-repetitive sampling is not possible, resampling is used.

1 4. The essence of systematic random sampling is the selection from the basis of the element, starting with the first element, which is selected at random.

For example, when forming a systematic sample with a size of 500 elements from a general population of 15,000 employees of an organization.

First, a random start is determined, then a selection step. (15,000/ 500=30, selection step is 30).

1 5. The SPPS selection method improves the estimation accuracy if the auxiliary size variable used to determine the probabilities is approximately proportional to the features being studied. When using the SPPS method, there is a greater likelihood that units with large features will fall into the sample. The SPPS sampling method is often used in household surveys to select areas where the likelihood of items being included in the sample is proportional to the size of the resident population in the sampling areas.

**Section 1. Stratified sampling**

1 6. In planning a household survey, a widely used method is stratification for the survey population prior to sampling. It serves the purpose of classifying a population into subpopulations based on additional information that is known about the population. For example, territorial characteristics or gender and age categories, type of area, number of residents, type or type of structure, building. The main principle of formation of stratum (stratification) is heterogeneity between stratum and homogeneity within stratum. Urban and rural areas are formed as two separate stratum for the household survey. Urban and rural populations differ from each other in many aspects (type of employment, source and size of income, average household size, birth rate) while persons belonging to one of these subgroups have similar characteristics. The probability of selection with a starified sample using non-repetitive random selection is calculated by the following formula:



where:

- the size of the sample in the stratum;

- the size of the general population in the stratum.

17. The benefits of a stratified sample are:

1) decrease in variation (scatter of variables) in the sample;

2) ensuring proportional representation of the selected units in each layer;

3) increasing the number of elements from subsets in the final sample will increase the reliability of the statistical analysis.

**Section 2. Cluster (nested) sampling**

18. "Cluster sampling" refers to sampling designs that include all members of the group. The groups themselves are referred to as clusters or nests. Cluster sampling reduces the reliability of the sample due to the high probability that individuals living in the same cluster tend to be homogeneous or to have more or less similar characteristics (the more homogeneous the sample, the lower its reliability). In order to avoid the degree of homogeneity of clusters, the number of clusters is increased. For example, for a survey of the standard of living of the population, the cluster sampling method is used. The sample is 12,000 households and 400 clusters of 30 households each are formed than 200 clusters of 60 households each.

19. In the first case, the design effect of sampling (hereinafter - ) is much smaller. Value decreases if households are selected randomly from all households in the same cluster, and not from sub-segments that are geographically adjacent to each other.



where:

-dispersion of a cluster or stratum;

- the variance for a simple random sample.

This calculation gives an estimate of the "full" measure , including the effects of stratification, as well as the variance in cluster sizes associated with nesting grouping. If greater than 1, then this indicates that the sampling design is less efficient (gives a larger error for a sample of the same size) than a simple random sample. This is compensated in terms of the sample by a commensurate increase in sample size. less than 1 indicates otherwise.

**Paragraph 3. Two-stage sampling**

20. For household surveys, one of the commonly used methods is two-stage sampling.

The aim of this two-stage approach is to save costs by conducting a short screening survey at the initial large sample stage. For example, at the first stage, the selection of settlements is carried out, and at the second stage, the selection of households.

21. To calculate the overall probability of selection ( *P* ) into the household sample for cluster, staged and two-stage sampling, a standard ratio is used, taking into account the staged nature of sampling. The formula for calculating the selection probability *P* for a two-stage sample is given below :



where:

- the probability of including primary sampling units (hereinafter - PSU) in the sample;

– the probability of including a household in the sample at the second stage is calculated by the formula.

The probability of including PSU in the sample ( ) at the first stage is calculated by the formula:



where:

- the number of PSUs to be selected in *the g* -th separately for urban and rural population ( *h* );

- the number of households in *the i* -th PSU included in the sample at the first stage (in the *g* -th separately for urban and rural population ( *h* ));

- total number of households for the entire population of PSU (in *g* -th separately for urban and rural population ( *h* )).

 - the probability of including a household in the sample at the second stage is calculated by the formula:

*,*

where:

- the number of households to be selected within *the i* -th PSU (in *the g* -th separately for urban and rural population ( *h* ));

- the number of households in *the i* -th PSU included in the sample at the first stage (in the *g* -th separately for urban and rural population ( *h* )).

**Chapter 5 Determining and Distributing the Sample Size**

2 2. To determine the optimal sample size required to estimate the population with a given accuracy, the following formula is used:

****

where:

*t* - argument of the Laplace function ( *t* = 1.96 for 95% confidence level);

*N* - the volume of the general population;

- dispersion;

Δ - the marginal sampling error.

2 3. To determine the sample size, the following parameters of the population are estimated:

1) Medium an arithmetic attribute (for example, household income and expenses, number of people living in households) is calculated for all units of the general population and is called the general average ( ) and is calculated using the following formula.

,

where:

- the number of elements of the general population i -strate ;

 - the sum of the index of the i -stratum.

2) Population variance - defined as the mean of the squared deviations of all individual observations from their mean.

Population variance is calculated using the following formula:

**

The square root of the variance is called the standard deviation or standard deviation and is calculated using the formula:



###### 3) If the error is expressed as standard error ( *μ*), then the following formula is used to determine the sample size:

**

where:

*RSE* - the relative sample standard error.

If the final population adjustment is not taken into account, the formula for determining the sample size will be as follows:

**

24. Once the sample size has been determined, the sample should be allocated to stratum if it is a stratified sample, or to clusters if it is a cluster sample. The sample distribution is made by the same sample size in each stratum (uniform distribution), or distributed in other ways. In order to determine the distribution of the sample to different stratum, there are two important criteria that affect how the sample size in the stratum is determined:

The first criterion is convenience: a method of proportional distribution is chosen, in which the sample size in *the i* -th stratum is calculated by the formula:



where:

*n i* - sample size of *i* - stratum ;

*i* = 1,2,…, *h* ;

*N i* - the number of households in *the i* -th stratum, while *i* = 1,2…., *h.*

The second criterion is accuracy: the method of optimal distribution is chosen, which gives the smallest mean square error (standard error) of the sample.

2 5. Where the costs of sampling from different stratum are the same, the optimal distribution formula is called the Neumann distribution. In this case, the sample size in *the i* -th stratum is determined by the formula:



where:

*h* - the number of stratum in the aggregate;

*i* = 1,2,…, *h* ;

*S i* - standard deviation *of the i* -th stratum.

The use of optimally distributed stratification (determining the sample size in *the i* -th stratum) is possible only if there is a standard deviation.

**Chapter 6 Sample rotation and overlays**

26. In a panel sample survey, in order to prevent the effect of household fatigue from participation in the survey and to smooth out jumps in periodic survey data, rotation (replacement) of households in the sample is periodically performed. Households are rotated once a year. For example, if it is decided that the size of the rotation will be equal to 1/3 of the total number of surveyed households, that is, annually 1/3 of the households will be removed from the sample and replaced by others, in this case, after 3 years the sample population will be completely updated. In this case, the replacement of households is carried out by the same method that was used initially to form the sample. In this case, households are replaced by those households that belong to the same stratum, segment or cluster.

**Chapter 7. Compensation for non-response**

27. In surveys, two types of non-response of respondents are distinguished - complete, when information about the object of observation is completely absent, and partial, when there are no answers to certain questions of the survey program.

28. There are two types of reasons for complete non- response : objective and subjective reasons.

Objective reasons for non-response include: if all members of the household do not fit the target group of the survey (by gender and age category), the house is destroyed, an empty apartment (house), merging of apartments, another reason (change of purpose of the premises), the house (address) was not found and other reasons.

Subjective reasons for non-response include: household refusal to participate in the survey.

29. 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 frame is constructed in the same way as the main sample frame for a particular survey. Substitution is allowed only from the reserve list with a 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.

In cases of non-response, household surveys do not provide for a substitute for compensation unless there is a good justification for its use for a particular purpose.

3 0. To correct for non-responses due to subjective reasons, weighting adjustments are applied and weighting is performed. Weighing the results of a sample survey is carried out by assigning an appropriate weight to each individual unit of observation - a person.

Weighting data in household surveys involves two steps: calculating basic weights and making adjustments (weight adjustments) for non-response.

3 1. The calculation of the basis weights begins after determining the probabilities of selecting sampling units. The probability of selecting a sampling unit depends on the sampling design used to select that unit. The basis weight of a sample unit is the reciprocal of its probability of being selected for inclusion in the sample. In mathematical terms, if a unit is included in the sample with probability , then its basis weight, denoted as ( spread factor ), is calculated by the formula:



3 2. The problem of non-response from the sampling unit for subjective reasons in household surveys is solved by adjusting the sample weights. The calculation of the adjusted non-response weight for the i-th sample unit is calculated using the following equation:

,

where:

- the number of actual reports.

The calculation of the final adjusted non-response weight for *the ith* sample unit is calculated using the following equation:

,

where:

 - the initial basis weight ;

 - adjusted weight in case of non-responses.

Initial weights are adjusted to compensate for non-responses.

3 3. The process of filling in partially missing values is called imputation. There are two main approaches to imputation:

methods based on the use of donor data;

methods for computing estimates of missing values.

3 4. The donor method is based on the use of response data from a similar respondent.In the methods of the donor, the imputed value for the record - the recipient is taken from another record donor, selected according to certain rules. There are two methods of donor imputation:

1. Random selection of a donor. The random selection of records used as donors is usually done in some specially defined classes. Classes are fairly homogeneous groups of observations according to the given parameters.
2. Nearest neighbor method. Imputation by the nearest neighbor method involves choosing such a record from among the donor ones, which has the smallest distance from the imputed record.

Methods for calculating missing value estimates are to substitute for the missing mean or to substitute historical data from previous surveys.

**Chapter 8 Estimation of sample parameters based on survey results**

3 5. Based on the results of surveys, the parameters of the sample are evaluated. To assess the reliability of the sample results, some statistical characteristics are calculated.

Estimate of the standard error of the sample. Possible discrepancies between the characteristics of the sample and the general population are measured by the standard error (mean error) of the sample. The sample standard error is determined by the following formula:



where:

μ- standard error ;

- general dispersion;

- the size of the sample.

36. The sample standard error shows the absolute values of the error. To determine the estimated value in shares, the relative standard error (coefficient of variation) is used. This coefficient is expressed as a percentage and is calculated by the formula:



37. The relative standard error is calculated after sampling surveys. The greater the value of the relative standard error, the relatively greater the scatter and the lower the evenness of the values under study. If the relative standard error for each stratum is less than 10%, then the variability of the variation series is considered insignificant, from 10% to 20% refers to the average, more than 20% and less than 33% to significant, more than 33% the sample is unrepresentative and a decision is made to increase the volume samples. To reduce the relative sample standard error by a factor of two, the sample size is quadrupled.

38. To establish the marginal sampling error, the standard error *( μ)* of the sample is used. Marginal sampling error by stratum* = t ⋅ μ.* The marginal error is used to calculate the confidence interval. The confidence interval shows in what range the results of sample observations will be located.

39. The coefficient *t is determined by the probability P* given by the researcher (0≤ *P≤1* ). For *P* values approaching unity, the possibility of the general mean differing from the calculated sample mean by more than ∆, guaranteed by the given confidence level of the probability *P , is practically excluded.* In this case, the higher the level of confidence (for example, the values 0.90; 0.95; 0.99, etc. are used), the higher the coefficient t, and, consequently, the value of the marginal error Δ.

*t* =1.28 for 80% confidence level;

*t* =1.64 for 90% confidence level;

*t* =1.96 for 95% confidence level;

*t* =2.58 for 99% confidence level;

Like the sampling characteristic itself, the sampling error is a random variable. Lyapunov's theorem, indicates the probability that the sampling error will not exceed some given value Δ, i.e. that |-| ≤ Δ.

The confidence interval of the general mean is determined based on the inequalities |-| ≤ Δ, from which it follows that x -Δ ≤  ≤x + Δ.

**Chapter 9 Spreading the results of the sample to the population**

40. \_ ultimate purpose selective observations is characteristic general aggregates on basis data received \_ By sample. The distribution of sample observation data to the general population is carried out according to the following formula:



or:



where:

- the sum of the indicator of the sample *i* -strate;

 - the average value of the indicator of the sample population *i* -strate.