Approved by the order of the Chairman of the Statistics Committee of the Ministry of National Economy of the Republic of Kazakhstan dated February 22, 2018

no. 36

**Methodology for the formation of a sampling frame for conducting a sample survey in agriculture**

**Chapter1. General Provisions**

1. Methodology for the formation of a sampling frame for conducting a sample survey in agriculture (hereinafter - Methodology) refers to the statistical methodology, formed and approved in accordance with [the Law](http://adilet.zan.kz/rus/docs/Z100000257_#z0) of the Republic of Kazakhstan dated March19, 2010 "On State Statistics" (hereinafter - Law).

2. The methodology establishes the main aspects and specific methods for analyzing the characteristics of the sample population, as well as their selection.

3. The methodology 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) in the field of sampling, in organizing and conducting a sample survey in the statistical activities of agriculture.

4. This Methodology uses concepts in the meanings defined in the Law, as well as in the Code of the Republic of Kazakhstan dated October 29, 2015 "Entrepreneurial Code of the Republic of Kazakhstan":

1) agricultural enterprise - a legal entity or its structural subdivision engaged in the production, storage and processing of agricultural products, the provision of services in the field of agriculture;

2) the general population - a complete group of all units of analysis, whose characteristics are subject to evaluation;

3) sample population (sample) - a set of cases (subjects, objects, events, samples), using a certain procedure, selected from the general population for participation in the study;

4) sample design - a probabilistic scheme for the formation of a list of a sample population;

5) standard error of the sample - the standard deviation of the value of the sample parameter from the sample mean value of this parameter.

**Chapter 2. Designing a sampling frame for an agricultural survey**

5. The sample set is formed on the basis of the sample design order, provided in an arbitrary form by the structural divisions of the Committee to the responsible department for the formation of the Committee's sample sets.

The sample design order specifies:

1) the name of the structural unit;

2) name of the survey;

3) the purpose of the survey;

4) description of the population;

5) the main feature for calculating descriptive statistics;

6) requirement for representativeness;

7) examination period;

8) the deadline for submitting the sample to the customer.

6. The main stages in the formation of a sample population are:

1) definition of the general population;

2) methods for selecting sampling units;

3) calculation of the main characteristics of the general and sample population and determination of the sample size;

**Paragraph1. Definition of population**

7. Within the framework of agricultural statistics, the main objects of observation of a sample survey are individual entrepreneurs, peasant or farm enterprises and households. To determine the population for the relevant category of holdings, the Agricultural Statistical Register (hereinafter – ACR) is used, which includes includes information on entities producing agricultural products in the Republic of Kazakhstan.

With the help of ACR, the consistency of statistical data is ensured, and the population of the survey is distinguished.

ACR is an integral part of the official statistics of the Republic of Kazakhstan.

According to the content of the ACR, an organized and systematized list of entities: legal entities, branches and representative offices, individual entrepreneurs, peasant or farm enterprises and households engaged in the production of agricultural products.

8. According to the Methodology for maintaining and updating the Agricultural Statistical Register, 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 2, 2016 No. 295 (registered in the Register of State Registration of Regulatory Legal Acts No.14581), ACR consists of four databases , containing information on the types of accounting units and producers of agricultural products:

1) agricultural enterprises based on legal entities;

2) individual entrepreneurs, peasant or farm enterprises;

3) households;

4) horticultural and dacha cooperatives.

**Paragraph 2. Methods for selecting sampling units**

9. Three methods are used for sampling in agriculture:

1) one-dimensional sampling method with probabilities proportional to size (hereinafter - USPPS);

2) multivariate sampling method with probabilities proportional to size (hereinafter - MSPPS);

 3) method of systematic selection (step by step selection).

10. When using the USPPS method, the probability of being included in the sample is determined for each stratum of population members. The probability is calculated for each element of the general population according to the selection criterion contained in the register.

11. The USPPS method in sampling in agriculture is used to survey the yield of wheat and rice based on information provided by the territorial bodies of the Committee, according to Statistical work plan for the corresponding year. The USPPS method consists of two steps:

1) in the first stage, farms are selected as primary sampling units (hereinafter – PSU) in proportion to their size;

2) in the second stage - as secondary sampling units (hereinafter - SSU) crop areas are selected from each selected farm in the first stage of PSU.

12. The probability of being included in the sample for a peasant or farm holding, where the criterion is one indicator, is calculated by the following formula:

, (1)

Where:

*Pi \_*- the probability of being included in the sample of the i-th farm;

*n* - the given sample size;

*x i*- the value of the sampling criterion (sown area) of the i-th farm;

*0.75* - exponent values;

- the sum of the sampling criterion (sown areas) over the entire general population.

When using the USPPS method, each unit, depending on criterion value is selected more than once, the inclusion probability is adjusted more than "1".

Probability adjustment is done by data expansion. Expanding limits the influence of variable values on the probability of sampling. The exponent value is 0.75.

13. At the first stage, using a systematic selection of agricultural producers, PSU is selected as follows:

1) expanded factor *EXPFACTOR =1/* ;

2) expanded data (seeding) *EPF = posev × EXPFACTOR;*

3) calculation of the accumulative sum *AKSUM = M1, M1+ M 2, M1+ M 2+ M 3, …;*

4) the step (interval) of selection *I = M / n is calculated;*

5) selection of a random number *R 0* between *0* and1*;*

6) calculation of a random start *R = R 0 × I;*

7) calculation of the series *R , R + I , R +2 I , R +3 I.*

14. At the second stage, the surveyed hectare is determined from the farms that fell into the sample at the first stage. The surveyed hectare ( *ga* ) is defined as follows:

 , (2)

Where:

*ga*– surveyed hectare;

*EXPFACTOR* - expanded factor;

*R - AKSUM* - calculation of the accumulative sum.

A one-dimensional sampling method with probabilities proportional to size is given in Appendix1 to this Methodology. Of the141 surveyed farms, 45 farms are selected (about 30%).

The calculation of the sample of holdings is made using a systematic selection of units:

1. accumulative amounts:1*05, 252, 321, 416...;*
2. selection step = *24542 / 45 = 545;*
3. random number = *0.969293;*
4. random start or first holding to be sampled

*R = 0.969293 × 545 = 529;*

1. included in the sample: *529, 529+545 =1074,1074+545=1619,1619+545= 2164* or farms numbered: *5,9,12,14,…;*

6) determination of the selected hectare on the farm *(545-416)/1=129* ha, *(1074-1058)/1=16* ha *, (1619-1452)/1=167* ha, *(2164-1917)/0.2305=1071* ha.

15. The MSPPS method is carried out with a multi-purpose survey of objects of heterogeneous indicators. The probability of hitting the i-th economy is determined by the formula:

(3)

Where:

*Pi \_*- the probability of being included in the sample of the i-th farm;

- the sum of the sampling criterion (sown areas) for the entire general population;

*x1, i x k ,i –* values of sampling criteria for *the i* -th farm;

*n1n k* - the sample sizes adopted for each sampling criterion.

16. When using MSPPS, the maximum hit probability is selected from the calculation for each criterion, if the hit probability is greater than1, the hit probability is assumed to be1*.* The holding probabilities are adjusted to1to ensure that large holdings are not sampled more than once.

17. To obtain a sample, the population is sorted in descending order of the probability of being included in the sample, a random start from *0* to1 *is chosen* , and farms are selected with a step of1 *on a probability scale.*

area under crops is calculated *(1 -1226.8;2 - 209.3* and *3 - 87.9).* Next, the sample size for each type of crop is determined. A multivariate sampling method with probabilities proportional to size is given in Appendix and 2 to this Methodology. The sample size for crop #1 is 8 units, for crop #2 is 8 units, for crop #3 is 7 units. Next, the inclusion probability for each crop, each holding (MSPPS in the above formula) is calculated. The highest probability ( *P max )* is selected from the calculation for each criterion , but if the probability is greater than1, the probability is taken equal to1*. In this example, the* first farm has the highest probability of inclusion in culture No. in culture No.1 ( *0.53* ). Next, using a random number between the interval *0-1* (in the example it is *0* ), we calculate the cumulative amount and determine the farms that fall into the sample.

18. The method of systematic selection involves the determination of the sampling step (“step”) and the starting number of the holding from which to start the sample. The sampling step is defined as follows: step = *N / n.* Next, the number of the farm is randomly determined , and the sampling (random start) begins from it, not exceeding the sampling step.

**Paragraph 3. Calculations of the main characteristics of the general and sample population and determination of the sample size**

19. To calculate the main characteristics of the general population, the following are mainly used:

Variance and standard deviation:

, (4)

Where:

– dispersion;

*–* values of sampling criteria for *the i* -th farm;

– positive mean value;

*N* - the volume of the general population.

= , (5)

Where:

– dispersion;

- standard deviation.

The dispersion shows how close the estimated indicators are to the mathematical expectation of the estimated function, and the standard deviation characterizes the degree of their closeness to the parameter. The evaluation function is a mathematical function, with the help of which the estimated indicator for a particular parameter is calculated. The term "statistical accuracy" refers to the magnitude of the variance, and the term "precision" to the magnitude of the standard deviation.

20. Standard (Standard error) - standard error sampling:

, (6)

Where:

*n* – sample size;

- standard deviation;

- sample standard error.

21. When forming and determining the sample size, it is necessary to pay attention to the calculated analysis of the resulting sample, to the calculations of descriptive statistics generated for all the most important indicators. Descriptive statistics are calculated for each crop type separately.

Descriptive statistics (this dataset) contains the following indicators:

1. *Requsted* - population sizes and required sample sizes;
2. *% of pop* – required sample sizes as a percentage of the populations;
3. *Obteind* - the sizes of sampling sets based on the selection results;
4. *% pop* - effective sample sizes as a percentage of the population;
5. *Expanded sum* is the sum of the register values for the population, and the sum of the register values for the sample, weighted by the probability of the sample.

Calculation formula:

 *Exp sum* = , (7)

Where:

*Exp sum* – the sum of the register criterion for the population;

*P i* - the probability of getting into the sample of the i-th farm;

*Χ i* -sampling probability.

1. ratio adjustment ( Ra ) - the sample bias factor used in data dissemination, in the case of completed questionnaires. Ra is an important indicator in determining the sample size. The sample is considered representative with a coefficient of deviation between 0.7 and1.3. The calculation formula is the following:

 *Ra* = , (8)

Where:

Exp sum *N* - the sum of the register values for the general population;

Exp sum *n* - the sum of the register values for the sample population;

Ra (ratio adjustment) – smoothing factor.

1. *2×CV % – 2×* approximate coefficient of variation of the common sum for the sample.

Calculation formula:

*×100%* (9)

Where:

*w mean* – the average (not weighted) value;

*CV* - coefficient of variation of the sample population;

**** - standard deviation.

The value of the coefficient of variation estimates the percentage of deviation of the results of the formed sample from the true value of the general population;

1. *Sum* - the sum (not weighted) of the register values for the population and the sample;
2. *% pop* - the sum of the register values as a percentage of the sum of the general population of the sample;
3. *mean of positive* - the average (not weighted) value, from the positive for the general population and for the sample, is calculated using the following formula:

 = / *N ,* (10)

Where:

– positive mean value;

*x i*- the value of the sampling criterion of the i-th farm;

*N* - the number of elements of the general population.

1. Minimum - the minimum value of the register for the general population and for the sample;
2. Maximum - the maximum value of the register for the general population and for the sample.
3. spread factor:

 (11)

where:

– distribution factor;

*k* - the number of register indicators taken as selection criteria;

*x1,i x k ,i –* values of sampling criteria for *the i* -th farm;

*n1n k* - the sample sizes adopted for each sampling criterion.

# Appendix1

to the Methodology for the formation of a sampling frame for conducting a sample survey in agriculture

Univariate sampling method with probabilities proportional to size

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Surveyed farms | Sown area |  | EXPFACTOR | EPF | AKSUM | Sample number | PSU | SSU (Selected ha on the holding) |
| 1 | 105 | 1 | 1 | 105 | 105 |   |   |  |
| 2 | 147 | 1 | 1 | 147 | 252 |   |   |  |
| 3 | 69 | 1 | 1 | 69 | 321 |   |   |  |
| 4 | 95 | 1 | 1 | 95 | 416 |   |   |  |
| 5 | 142 | 1 | 1 | 142 | 558 | 1 | 529 | 129 |
| 6 | 400 | 1.3556 | 0.7377 | 295 | 853 |   |   |  |
| 7 | 160 | 1 | 1 | 160 | 1013 |   |   |  |
| 8 | 45 | 1 | 1 | 45 | 1058 |   |   |  |
| 9 | 84 | 1 | 1 | 84 | 1142 | 2 | 1074 | 16 |
| 10 | 120 | 1 | 1 | 120 | 1262 |   |   |  |
| 11 | 190 | 1 | 1 | 190 | 1452 |   |   |  |
| 12 | 170 | 1 | 1 | 170 | 1622 | 3 | 1619 | 167 |
| 13 | 400 | 1.3556 | 0.7377 | 295 | 1917 |   |   |  |
| 14 | 1887 | 4.3392 | 0.2305 | 435 | 2352 | 4 | 2164 | 1071 |
| 15 | 380 | 1.3044 | 0.7666 | 291 | 2643 |   |   |  |
| 16 | 123 | 1 | 1 | 123 | 2766 | 5 | 2709 |  |
| … | … | … | … | … | … | … | … |  |
| 141 | 100 | 1 | 1 | 100 | 24542 | 45 | 24525 |  |

# Appendix 2

to the Methodology for the formation of a sampling frame for conducting a sample survey in agriculture

Multivariate sampling method with probabilities proportional to size

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Surveyed farms | Cultures | Probability of inclusion | Accumulative amount | Getting into the sample |
| 1 | 2 | 3 | P1 | P2 | P3 | Pmax | plast | Random number 0 |
| 1 | 118.8 | 3.6 | 0 | 0.23 | 0.10 | 0.00 | 0.23 | 0.23 | 0.23 |  |
| 2 | 348.2 | 8.3 | 0 | 0.53 | 0.19 | 0.00 | 0.53 | 0.53 | 0.76 |  |
| 3 | 3.9 | 3.9 | 3.9 | 0.02 | 0.11 | 0.22 | 0.22 | 0.22 | 0.98 |  |
| 4 | 10 | 10 | 6 | 0.04 | 0.21 | 0.31 | 0.31 | 0.31 | 1.29 | 1 |
| 5 | 7 | 7 | 7 | 0.03 | 0.16 | 0.34 | 0.34 | 0.34 | 1.63 |  |
| 6 | 5.5 | 5.5 | 5 | 0.02 | 0.14 | 0.27 | 0.27 | 0.27 | 1.90 |  |
| 7 | 71 | 3.4 | 3.4 | 0.16 | 0.10 | 0.20 | 0.20 | 0.20 | 2.09 | 1 |
| 8 | 10 | 10 | 2 | 0.04 | 0.21 | 0.13 | 0.21 | 0.21 | 2.31 |  |
| 9 | 121 | 12 | 0 | 0.24 | 0.25 | 0.00 | 0.25 | 0.25 | 2.56 |  |
| 10 | 4.9 | 4.9 | 4.9 | 0.02 | 0.13 | 0.26 | 0.26 | 0.26 | 2.82 |  |
| 11 | 16.3 | 5.7 | 2.7 | 0.05 | 0.14 | 0.17 | 0.17 | 0.17 | 2.99 |  |
| 12 | 20 | 20 | 8 | 0.06 | 0.36 | 0.38 | 0.38 | 0.38 | 3.37 | 1 |
| 13 | 200 | 50 | 0 | 0.35 | 0.72 | 0.00 | 0.72 | 0.72 | 4.08 | 1 |
| 14 | 135.8 | 7.9 | 3.5 | 0.26 | 0.18 | 0.20 | 0.26 | 0.26 | 4.34 |  |
| 15 | 11.8 | 4.7 | 4.7 | 0.04 | 0.12 | 0.25 | 0.25 | 0.25 | 4.60 |  |
| 16 | 20 | 20 | 14 | 0.06 | 0.36 | 0.58 | 0.58 | 0.58 | 5.17 | 1 |
| 17 | 2.3 | 2.3 | 2.3 | 0.01 | 0.07 | 0.15 | 0.15 | 0.15 | 5.32 |  |
| 18 | 95.8 | 5.6 | 4 | 0.20 | 0.14 | 0.23 | 0.23 | 0.23 | 5.55 |  |
| 19 | 16.5 | 16.5 | 16.5 | 0.05 | 0.31 | 0.65 | 0.65 | 0.65 | 6.20 | 1 |
| 20 | 8 | 8 | 0 | 0.03 | 0.18 | 0.00 | 0.18 | 0.18 | 6.38 |  |
| Total | 1226.8 | 209.3 | 87.9 |   |   |   |   |   |   | 6 |