Sampling Methodology of the Survey of Agricultural Holdings in Georgia

Author: Giorgi Balakhadze
Geostat, Agricultural and Environment Statistics Department
30 Tsotne Dadiani Str.
Tbilisi, Georgia
gbalakhadze@geostat.ge

Abstract

The paper will overview details of the sampling methodology practices for the survey of agricultural holdings – the important source of official agricultural statistics of Georgia conducted permanently by the National Statistics of Georgia (Geostat).

A good elaborated sampling methodology is very important within statistical surveys and it reflects on the whole statistical processes, from data collection to processing. Therefore, a sample planned on a wrong manner directly cause errors in terms of final statistical data production. Thus, errors in sampling methodology can decrease entire trust towards the statistical information throughout data stakeholders.

The sampling design crucial regarding various statistical assessments and estimations, calculation statistical errors. This last one is significantly dependent on the standard error of estimation, whose calculation formula essentially provides sample design.

A general source of sampling of the survey of agricultural holdings of Georgia is the dataset of the agricultural census conducted in 2014. The sampling size for the survey of agricultural holdings covers about 12,000 holding. Noteworthy, that total number of holdings in sampling frame amounts 640,000.

A two stage stratified sample is used for the survey of agricultural holdings, where the strata is defining by three following factors:

1. **A geographical region –** it implies a region where the holding operates. Noteworthy, if the holding operates in more than one region, as the "statistical region" of a holding would be considered accordingly to FAO's methodology.
2. **Legal status** – this is another strata actively used in sampling of the survey of agricultural holdings of Georgia and allows to distinguish family and non-family holding (enterprise).
3. **Size of holding** – in order to identify size of the holding, an aggregated indicator agricultural index is used, which is calculated throughout all agricultural assets of the holding (agricultural land (sown/arable land, permanent yards, meadows, and pastures, greenhouses) and livestock in one unit. convention carried out on the basis of the economical comparison and for benchmark unit was considered one hectare sown of the maize (the conventional factor of land that doesn't give economical input in direct form (such as uncultivated arable land, meadows and pastures, etc.) evaluated by minimal potential or some other methods).

It should be noted, that size of the holding is divided into 4 groups - (small, medium, large, and extra-large holdings) in each region. The borders of stratums are determined individually by the method of square roots. In small and medium holdings there wasn't revealed a necessity of separating family holdings and enterprises due to the nonexistence/few numbers of such size enterprises, the similarity of scale, or non-significant statistical difference between probabilistic distributions; there was no necessity in extra-large holdings also because this type of holdings appearing in the sample with probability equals to 1. Therefore, only the stratum of large holdings is split by family holdings and enterprises stratum. As a result, we have 55 stratums (11 regions X 5 size/legal status). For redistributing of sample size by the stratums was used von Neumann's allocation principle (redistributing of whole sample size by the proportion of multiplication of the number of holdings in stratum and standard deviation in stratum (by the agricultural index)).

The paper will be in touch with the issues related to stratification and sampling. the paper will help young experts and researchers to get additional information on experience adopted in Georgia, through cooperation with international partners.

Introduction

National Statistics Office of Georgia, the legal entity of public law, carries out its activities independently. It is an institution established to produce the statistics and disseminate the statistical information according to the Georgian legislation. Statistical system in Georgia is centralized. Geostat is a coordinating institution of statistical system of Georgia. According to the law, it provides coordinated work with official statistics producing bodies, issuing recommendations on statistical standards and methodology required for statistics, coordinates the exchange of available information in administrative bodies for the purpose of producing statistics and promotes the approved statistical standards and methodology. Production of official statistics is based on the 10 basic UN principles of official statistics. Geostat is the agency responsible for conducting censuses on population, housing and agriculture.

Utilization of Georgian agricultural potential is vitally important for the Georgian economy. For this purpose, the government places great emphasis on the need for investment in increasing output and productivity of Agricultural Sector. Demand for agricultural data is increasing daily and Geostat has to respond to user needs.

Despite the fact that share of agriculture, hunting, forestry and fishing in GDP was only 8.0% in 2017, agriculture has always been one of the important sectors of Georgian economy. 43.2% of employees in Georgia were employed in this sector. The share of rural population is 42.8% according to the Population Census 2014.

Almost every household living in villages is an agricultural holding. According to the Census of Agriculture 2014, total number of agricultural holdings is around 642 thousand, out of them only 0.3% is legal entity while other is household. Majority of agricultural holdings are small and they produce agricultural products mainly for own consumption.

Georgia has favorable geographical location and climate, which allows to produce more than 25 kinds of permanent and more than 20 kinds of annual crops. Also, animal husbandry is quite common in the agriculture of Georgia. In Georgia the agricultural sector is not well-specialized and majority of holdings produce many different kinds of agricultural products.

Sample Design

The role of sampling in statistical surveys is extremely important. Sampling design significantly determines statistical surveys at every stage, including data calculation and verification of reliability. And reliability, obviously, is the central issue for any statistical survey, because of the stochastic nature of the results. This issue is essentially related to the confidence interval of the estimation, that in turn depends on the standard error (the dispersion of the estimation, as of random variable (defined at the probability space of all possible samples)) and this variance is the function both for the data, as for a sampling design which has non-constant relation with it. The complex economical surveys are based on complicated sampling design, precisely for this reason, non-parametric and to some extent, universal methods of variance estimating become more important in modern statistics, despite that in this case the estimation of the stochastic variable is no more accommodates into the frame of constancy. However, it should be noted that the non-parametric

methods are mainly based on sub-sampling techniques, and the design of it also depends on the initial design.

Generally, sampling design is defined as follows: Let G represents the population and

$$|G| = \{\{s_n\}_n : s \subseteq G \text{ where s is a set of the members of sequence } \{s_n\}_n\}$$

is the set of all subsequences from it. Then the real-valued p function, defined on $|G|$ calls sampling design if it satisfies two conditions: (i) $\forall s \in |G| : p(s) \geq 0$ ლა (ii) $\sum_{s\in|G|} p(s) = 1$. Thus, p satisfies Kolmogorov's Axioms, so sampling design can be conceptually understood as the probabilities of realizations of a particular sample. For clarification, let us note that $|G|$ we defined as a set of subsequences, but not subsets. The reason for this is that theoretically, and in principle in many practical situations also, it is possible for one point (an element of the population) to occur more than once in the sample. By this criterion, distinguish from each other samples with (case when there exists at least one sample with at least one element occurring at least two, that have a strictly positive probability of selection) and samples without replacement. Let us note that in order to make a probabilistic sample, certain information about the population is needed, which is the data from the agricultural census for the survey of agricultural farms in Georgia.

The elementary and intuitively most understandable sampling design is simple random sampling. At this time, by using certain statistical methods the sample size is formed, and then all possible samples have an equal probability of selection (in this case, all the selected points have the same statistical weight and is the ratio of the number of the population to the number of the sample). To estimate a parameter or parameters that have more or less uniform distribution, this sampling design is effective and fully justified but, for estimation of the parameters that empirically are characterized by a clearly defined asymmetrical behavior and/or contain sharp outliers from the distribution it is dramatically undesirable since it is possible to get huge underestimation and overestimation. The characteristics of agriculture in Georgia, probably, like in many other countries, require an assessment of just such parameters. In addition, evaluation is needed at both the country and regional levels, which further complicates the situation.

In this situation, the standard solution of this problem in the theory of statistical sampling is stratification, i.e., dividing the entire population into more or less homogeneous groups and conducting a simple random selection or some procedures of any other design in each of them (in the case of one-stage selection is usually used a simple random or systematic sampling). It is possible to make stratification on categorical variable(s) as well as on continuous variables. Since our goal is to present the data at a regional level, the region itself was used as one of the variables that define the stratum. As the main reason for stratification is to divide the population into groups of several homogeneous holdings, it is necessary to use some variable or set of variables that determine the size of the agricultural holding. Theoretically, all indicators on which the survey is based (of whose description is the goal of the survey) can be used, but in this case, the procedure becomes unnecessarily complicated. In this case, it is advisable to use a small number of the most general variables, and at best a variable. It is possible to use as a variable the following categories: the total area of land operated by agricultural holdings, the total area of agricultural land, production of any main crops, sown area of any annual crops, number of any main species of livestock, the total incomes of holdings, or the total expenses of holdings, etc. However, none of the above indicators can completely independently rank holdings by size (total area may include non-agricultural land; agricultural land includes both arable land and permanent crops, which have

different importance; the holding may be large, but in the reference year, this should not be reflected in income (in addition, in censuses, on which the samples are based on, there is rarely collecting information about income and expenditure, or about agricultural production. It is the same in the case of Georgia, agricultural census covers only land areas by land categories and crops, and numbers of livestock), etc.

Agricultural Index

In order to avoid stratification with more than one continuous variable, which can unnecessarily increase the number of strata, it is necessary to choose some other, more general indicator, in the role of which the Geostat uses the so-called agricultural index, with which agricultural assets like agricultural land and livestock are standardized i.e. converted into one comparable unit (Nadareishvili, 2016).

To achieve comparability, the average expected income per hectare/unit (in the case of land/animal livestock, respectively) based on past agricultural surveys was compared to a baseline that was taken as a maize crop per hectare. Accordingly, for each i-th asset, a conversion factor was calculated $C_i = \frac{U_i}{U_s}$, after which a new variable was obtained - the agricultural index, calculated as follows:

$$\mathrm{Ind}_j = \sum_i C_i A_{ij} \quad (1)$$

where Ind_j is an agricultural index of the farm j, and A_{ij} is the amount (ha/unit) of the asset i that is operated by the holding j.

We note here that C_i ratios are random variables with its probabilistic nature, and we use only its average value, which for random variables with high variance can give deep pointwise errors (errors at the holding level), especially since some assets (i) are quite general and combine rather heterogeneous cases, for example, a greenhouse, regardless of what is grown there; an orchard from which only the nuts area are singled out as a completely different case, and the rest of the plants are combined under one variable. Even if this is not the case, homogeneity (ensuring only slight deviations from the mean) is still not guaranteed, because even if we decompose the orchards variable into different ones for each type of planting, we are still not immune to the presence of, for example, non-distributed apple varieties in the region, which are characterized by significantly higher profitability or a higher price in the market. This is the reason why the initial sample weights resulting from stratification may not be the final weights, but may require some adjustments (post-stratification).

It should also be noted that for 2014 census-based surveys of agricultural holdings, it does not take into account agricultural assets like scattered trees, agricultural machinery, operated by the holding, etc. In general, it is important to say that stratification, whatever it may be, usually does not distort the quality of the estimate, but on the contrary, reduces the variance of the estimate compared to a simple random sample, i.e., gives a more efficient estimate. Thus, for calculating the stratifying variable, we may not use such variables which can acceptably estimate even by simple random sampling. In the example of Georgia, scattered trees and the agricultural machinery used by holding are such variables. As proof, the table below compares the variable number of trees in orchards with the variable of scattered trees by several statistical parameters:

Table 1

	Coefficient of variance (%)	Coefficient of variance (without hazelnut trees) (%)	Coefficient of variance (without hazelnut trees, below the 95th percentile) (%)	Skewness (without hazelnut trees)		
				Tradicional (Fisher's)	Pearson's second	Kelly's (based on quantilles)
Trees in permanent orchards	269	302	174	24	0.87	0.95
Scattered trees	144	129	67	12	0.67	0.67

Stratification

One of the most important part of stratification by a continuous variable is to determine the boundaries of the stratum, which in turn, mostly depends on the form of allocation that will use later, i.e. depends on the selection of the proportion of sample size between stratum. However, for the consistency of the discussion, we will first consider the methods of determining the boundaries of the stratum (we mean that the Neymann allocation is selected, which we will discuss further).

However, before determining the boundaries of strata, first, it should be to select the quantity of stratum. About this, it is interesting that it is proved that in the case of uniform distribution, if the y variable is divided into L strata with equal width (difference between upper and lower boundaries of the strata), then

$$V(\bar{y}_{str}) = \frac{V(\bar{y})}{L^2} (2)$$

where $V(\bar{y})$ is the variance of the average value of the stratifying variable in the case of simple random sampling, $V(\bar{y}_{str})$ is the same variance in the case of a stratified sampling (when $n_h = n/L$, where n is the sampling size in simple random sampling, and n_h sampling size in strata k. It is obviously, that $\sum_k n_k = n$).

However, the mentioned fact would not be interesting enough if it ended here; Cochran empirically has shown that this attitude approximately occurs in other cases, even for asymmetric distributions (in the condition of Neymann allocation). Thus, under stratification conditions, the variance of the stratifying variable decreases by the square of the number of strata. However, it is clear that an infinite increment in the quantity of stratum is not the right decision, since in this case the cost increases inefficiently and logistics become more difficult. Depending on empirical conditions, it considers that the optimal number of strata (determined by one continuous variable) in surveys is from 2 to 6. In Georgia, in the survey of agriculture holdings, 4 strata are used for each region.

The determination of stratum boundaries is based on the idea that the boundaries are chosen in such a way as to minimize the variance of the mean of the stratifying variable. Accordingly, $y_1, \dots, y_K$ strata boundaries must satisfy the following system of equations.

$$\frac{\partial V(\bar{y}_{str})}{\partial y_i} = 0, \ i = \overline{1,K} (3)$$

In the case of Neyman's allocation, this system of equations takes the following form:

$$\sigma_i + \frac{(y_i - \mu_i)^2}{\sigma_i} = \sigma_{i+1} + \frac{(y_i - \mu_{i+1})^2}{\sigma_{i+1}}, \ i = \overline{1,K} \ (4)$$

where μ_i is mean and σ_i is the standard deviation of stratifying variable in i-th strata. Usually, these parameters are unknown, and therefore it is very difficult to determine the boundaries of the

strata. That is why there are many approximate methods that are easier to implement in practice due to certain assumptions.

One such classical method, used by Geostat in the agricultural survey, is the Dalenius-Hodge method, or, as it is otherwise called, the method of cumulative arithmetic roots. The method is based on the assumption that stratifying variable is uniformly distributed in potential stratums, so the specified variable is piecewise uniformly distributed. It is obvious that in practice such distributions appear rare, however, if we assume that the domain is divided into a sufficient number of subintervals, the variable is approximately uniform on the obtained subintervals (in practice, it is often not possible to confirm this fact with statistical tests, although practitioners consider it a minor problem if the variable does not contain sharp outliers or from large strata all the holdings will include in the sample with a probability of 1 - this is exactly the approach Geostat uses in the mentioned survey). It is not hard to prove that, with a given condition, $y_1, \dots, y_K$ boundaries of the stratum must satisfy the conditions:

$$\int_{y_{i-1}}^{y_i} \sqrt{f(y)}dy = \frac{1}{K}\int_{y_1}^{y_K} \sqrt{f(y)}dy\,, \quad i = \overline{2,K} \text{ (5)}$$

where $f(y)$ is the distribution density function of the random variable from which y variable is realized. Note that the right side of the equation is constant, that is, it does not depend on i. Since in our case we are dealing with the data, the function of relative frequencies of the grouped data is used instead of the $f(y)$ theoretical function (theoretically, it is possible to use the Rosenblatt kernel density estimator). Accordingly, this method implies choosing the points as strata boundaries in such a way that the sums of the roots from the frequencies of the subintervals, that fell into the strata, are equal.

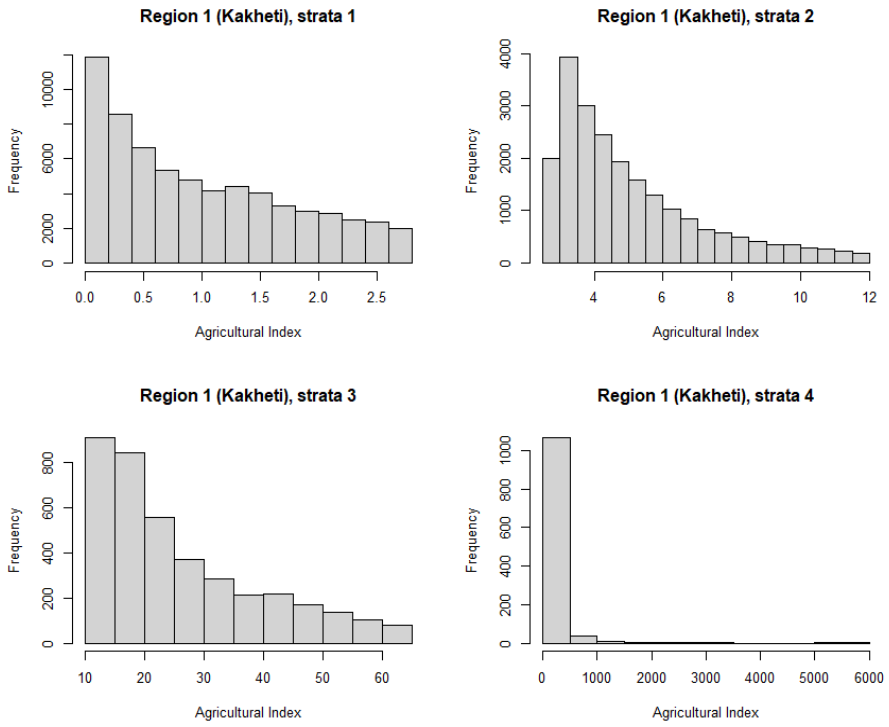
One of the problems refers when using this method is that the results obtained significantly depend on the grouping of the data, that is, on the length of the grouping interval. To illustrate this, let's take the example of Georgia for three large geographical regions when in the first case for the length of the group/bin (while binning the data) there were taken agricultural index equal to 0.2 (from an economic point of view - 0.2 ha of sown area of maize), and in second case - 0.1:

Table 2

Region	Length of bins: 0.2 AI				Length of bins: 0.1 AI			
	y_0	y_1	y_2	y_3	y_0	y_1	y_2	y_3
Region 1 (Kakheti)	0.0	2.6	11.0	57.8	0.0	2.4	9.3	46.5
Region 2 (Samegrelo)	0.0	2.4	5.4	11.0	0.0	2.3	5.2	10.4
Region 3 (Shida Kartli)	0.0	1.2	3.2	8.8	0.0	1.1	3.0	7.6

It should also be noted that there is no uniform distributions in the stratun (the histograms below clearly show that, despite stratification, all farms from large strata should be included in the survey surely):

Figure 1



There are also other well-studied methods for determining strata boundaries. The existence of a uniform distribution within a stratum is also based on Gunning and Horgan's method, according to which the boundaries of the stratum should choose so that the coefficients of variation be the same within the stratum (the idea belongs to Cochran):

$$CV_1(y) = \cdots = CV_K(y) \quad (6)$$

So,

$$\frac{S_1(y)}{\overline{y_1}} = \cdots = \frac{S_K(y)}{\overline{y_K}},$$

where $CV_i(y)$ is the coefficient of variation, $S_i(y)$ is the standard deviation and $\overline{y_i}$ is the arithmetic mean of the y variable within the boundaries of the corresponding stratum. If we consider the fact that y the variable is uniformly distributed in the stratum and use the well-known formulas of the mathematical expectation and variation of a uniformly distributed random variable, we get that the condition about equality of the variation coefficients is equivalent to the following condition:

$$\frac{y_{i+1} - y_i}{y_{i+1} + y_i} = \frac{y_i - y_{i-1}}{y_i + y_{i-1}}$$

for any i, whence simple algebraic transformations yield the following:

$$y_i = y_0 \left(\frac{y_K}{y_0}\right)^{\frac{i}{K}}, \quad i = \overline{0, K}$$

As we can see, in this case, the boundaries of the strata establish a geometric progression, therefore this method is also called geometric stratification. It should be noted that with this method,

stratification is performed only for the variable that does not contain zero in data, since in this case $y_0 \neq 0$. Due to the geometric nature of the stratification, the use of this method is not recommended in the case of symmetric distributions, it is relevant for deeply asymmetric, Pareto-type distributed data, which often appears in agricultural statistics, usually, especially in developing countries, where there are a large number of small and a small number of big producers. In addition, this method is avoided for very small y_0, since in this case, we will have many strata with small lengths.

Another well-known (iterative) method is proposed by Lavallee and Hidiroglou, which is based on the assumption (rather than a recommendation) that the largest strata, resulting from stratification, will be completely sampled, i.e., the variance of the estimate into extremely large strata will be zero; as a result, the variance of the mean value of the variable y will be:

$$V(\bar{y}_{str}) = \sum_{i=1}^{K-1} \left(\frac{N_i}{N}\right)^2 \frac{S_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right) \quad (7)$$

Because, as we mentioned above

$$S_K^2 = 0 \quad \Rightarrow \quad \left(\frac{N_K}{N}\right)^2 \frac{S_K^2}{n_K}\left(1 - \frac{n_K}{N_K}\right) = 0.$$

For n sample size the equation can be solved as follows:

$$n = n\left(y_1, \dots, y_{K-1}, CV(\bar{y}_{str})\right) = N_K + \frac{\sum_{i=1}^{K-1} \left(\frac{N_i}{N}\right)^2 \frac{S_i^2}{n_i/(n - N_K)}}{\left(\bar{y}CV(\bar{y}_{str}) + \sum_{i=1}^{K-1} \frac{N_i}{N}\frac{S_i^2}{N}\right)^2}$$

The idea of Lavallee and Hidiroglou is to choose the $y_1, \dots, y_{K-1}$ boundaries of the stratum in such a way as to minimize the $n(y_1, \dots, y_{K-1})$ sample size for fixed coefficient of variation, which means to solve the following system of equations:

$$\frac{\partial n}{\partial y_i} = 0, \qquad i = \overline{1, K-1}$$

However, the analytical (functional) solution of the mentioned system is associated with great technical difficulties, in addition, as known from classical mathematical analysis, the satisfying of the mentioned system of equations by the $K-1$ dimensional Euclidean vector is a necessary, but not sufficient condition for reaching a minimum value of n on it, since the mentioned point can only be the local, but not global extremum point. That is why at different times a number of iterative and numerical methods have been proposed that approximate the solution, among which are the algorithms of himself Hidiroglou, Sati, Kozak, and others.

The table 3 below compares the results obtained by three methods; Results seem quite different from each other:

Table 3

Region	Dalenius-Hodge			Gunning - Horgan			Lavallee-Hidiroglou (Kozak algorithm)		
	y_1	y_2	y_3	y_1	y_2	y_3	y_1	y_2	y_3
Region 1 (Kakheti)	2.4	9.3	46.5	0.01	1.1	80.4	3.7	22.5	160.7
Region 2 (Samegrelo)	2.3	5.2	10.4	0.08	5.6	424.0	2.5	7.0	33.0
Region 3 (Shida Kartli)	1.1	3.0	7.6	0.01	0.5	15.5	1.4	4.6	22.8

It should be noted that although the stratification procedure described above is based on an agricultural index and therefore serves as statistical reliability for its statistical estimations, statistical offices usually do not produce information about this one, but about the main agricultural indicators for individual crops, livestock, or other agricultural assets. Taking this into account, it must be emphasized that often the stratification by an objective collective indicator (in our case, such indicator is an agricultural index) can significantly coincide with the stratifications by the individual important compiler (crops/livestock/assets); however, if the collective indicator consists of several equally important and often empirically incompatible compilers, then, in this case, the difference can even be very noticeable (and in agriculture, we often have such situation, because usually, in many cases, holding is mainly specialized and, accordingly, is focused on one (or a small number of) activity). For example, let's compare the results for different stratifying variables while using the cumulative square root method for Kakheti, one of the largest regions in Georgia:

In the table 4, the issue of changing the stratum was discussed in each case only for those holdings that own the specified asset.

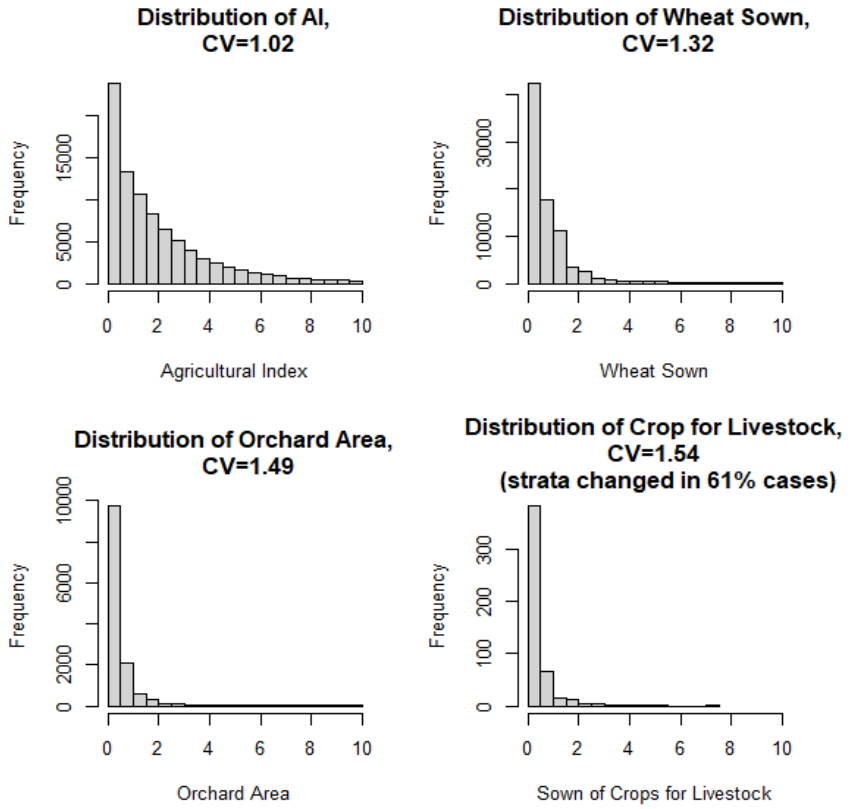
Table 4

Alternative stratifying variable	Strata has not changed (%)	Strata has changed (%)
Arable land	39	61
Wheat sown area	54	46
Maize sown area	56	44
Barley sown area	53	47
Potate sown area	56	44
Orchard area	58	42
Wineyard area	54	46
Area of meadows and pastures	52	48
Quantity of cattle	45	55
Quantity of sheep	61	39

In the table, the issue of changing strata was discussed in each case only for those farms that own the apropriate asset.

As we can see, the change of strata was the rarest for the most general variable (for example, arable land), which is due to the fact that the agricultural index is "filled" more fully with the general variable; at least, we can say that in this case the distance between the maximas of the stratifying variables is smaller. However, the small distance itself cannot lead to such a result unless there is also some correspondence in the distributions. (We are speaking of a theoretical or relative distribution, otherwise, the scales will be different):

Figure 2



The graphs show histograms of the agricultural index and area of various crops in the indicated region in the intervals from 0 to 10 of the corresponding variable. Pay attention that the further the distribution moves from the origin (which, as can be seen from the diagram, within a certain margin of error can well modeled by the exponential distribution, especially, since it satisfies the well-known property of that distribution: $\mu \approx \sigma$, so $CV \approx 1$; however, neither the Kolmogorov-Smirnov criterion, nor any well-known Monte-Carlo method not show a statistically significant relationship between the empirical distribution and any well-known right-skewed distribution, such as itself exponential, gamma, chi-squared, or the Weibull distribution), the share of holding that change the strata while changing the stratifying variable, is more. It is interesting that mainly between which strata the change takes place; consider the latter for the same region when re-stratifying by wheat sown:

Table 4

		Stratification with wheat sown				Total
		Strata 1	Strata 2	Strata 3	Strata 4	
Stratification with agricultural index	Strata 1	1451	1273	77	0	2801
	Strata 2	1075	1871	1180	51	4177
	Strata 3	151	327	374	375	1227
	Strata 4	27	36	64	289	416
Total		2704	3507	1695	715	8621

As we can see, and this was to be expected, the change occurs in both directions and mainly between the first and second strata. As a result of the above discussion, there emerges one "defect" of stratification with the agricultural index (and the collective variable in general) when representatives of the highest stratum participate in the survey with probability 1 and the survey is used to estimate parameters of individual agricultural crops and livestock: a holding may fall into the specified stratum, while there is no "objective necessity": The holding owns a large number of agricultural assets (land/animals) in general, but not any specific one (However, on the other hand, if we also estimate, for example, the average incomes and/or expenses of the farms within the scope of this survey, then this will only positively affect the accuracy of the estimation).

Sampling Methodology

This question of stratification could be finished here, but there is also another factor that is particularly important, as it is at least desirable, and in many cases, it is necessary to analyze or calculate a number of indicators in terms of this factor, and this factor is the legal status of the holding, more precisely, whether the holding is a household or a legal entity. To determine whether the latter (legal status) should be chosen as another stratifying variable, from the 44 strata already identified above (11 regions x 4 sizes), 33 (excluding the large stratum, since there we have apriori a complete sample) were tested with the question of heterogeneity and found that although the ratios of the average values of the agricultural index by legal status are close to one and there is no significant statistical difference between the strata, the difference between the absolute values (and this is what we are interested in) was significant in the third stratum and insignificant in the first and second almost in all major regions. Mann-Whitney test does not show a statistically significant difference between the averages in the first and second stratum almost in all regions, and it was significant in the third stratum; the Levene test for variation, the Kolmogorov-Smirnov test for distributions, etc., gave similar results. As a result, the third stratum was divided into two strata according to legal status. Thus, the sample frame of the agricultural survey now consists of 55 permanent strata.

After defining and fixing the stratum and determining the sample size in the regions, which depends, on the one hand, on the financial possibilities and, on the other hand, on a certain predefined reliability (and on the third hand, on this issue we have to talk about now), it is necessary to distribute the mentioned sample size among the stratum. There are several approaches to this in the theory of sampling. The first and simplest approach is that the selection should be greater in the strata in which there are more points (objects) and should be as many times as the number of mentioned points. In other words, the sample size should be distributed proportionally to the size of the strata:

$$n_h \propto N_h \quad \Leftrightarrow \quad n_h = n\frac{N_h}{\sum_i N_i}$$

where h is the defining indicator of the stratum, n_h is the sample size in the stratum, N_h is the population size in the stratum, and n is the total sample size (in the region in our case). The latter is called allocation by the stratum size.

It is clear that this approach is somewhat naive, because no matter how large the strata is, if additionally, I know that the variable takes only one value, then it is quite sufficient if I choose one point/holding (i.e., as small as possible), and if the strata are usually non-homogenous, I should take a large part of strata in the sample to achieve some accuracy, no matter how small the stratum

is. Thus, in this approach, important is the standard deviation, not the size of the stratum. So allocation by the standard deviation implies the following:

$$n_h \propto \sigma_h \quad \Leftrightarrow \quad n_h = n\frac{\sigma_h}{\sum_i \sigma_i}$$

where this means that σ_h the standard deviation (or its estimate) is known to us from previous rounds of surveys or census data.

There is a compromise option between these two methods, where both the size of the stratum and its variation are taken into account, and the sample size in the stratum is taken in proportion to the product of these two values:

$$n_h \propto \sigma_h N_h \quad \Leftrightarrow \quad n_h = n\frac{\sigma_h N_h}{\sum_i \sigma_i N_i}$$

The latter is known as the Neymann allocation and is also referred to as the optimal allocation in some statistical literature.

All three examples mentioned above are special cases of allocation of the following general scheme:

$$n_h \propto CV_h X_h^m \quad \Leftrightarrow \quad n_h = n\frac{CV_h X_h^m}{\sum_i CV_i X_i^m}$$

where X_h is a function, defend of strata, that can be interpreted as an indicator of the importance of the strata h, m is a constant (in the statistical literature, it is usually in $[0,1]$ interval), that determines the functional form of the association n_h with the given X_h variable. Of course, it is theoretically possible to consider a more general version if we change the polynomial form of the second co-multiplier to the general functional form: $n_h \propto CV_h f(X_h)$, where f is non-negative function. The three types of allocation discussed above are clearly special cases of this one; for example, this posterior transformation turns into a Neymann distribution when $m = 1$ and $X_h = N_h E_h$, where E_h is the mean of the variables in the strata h. Moreover, the case $m = 0$ was transformed into the allocation by the coefficient of variation.

For one of the regions, let's compare the shares of sample size according to the strata in the case of several types of allocations from the above (except for the fourth stratum; for simplicity, the third stratum is not divided into sub-strata by legal status):

Table 5

Strata	Allocation by the size of strata	Allocation by standard deviation	Neymann allocation	Allocation by the coefficient of variation
Strata 1	69	6	34	49
Strata 2	26	14	31	22
Strata 3	5	80	35	29

One of the most popular and frequently used in practice Neyman allocation is used in the Survey of Agricultural Holdings in Georgia.

Table 6

Region Number	Region	Holding size					
		Family holding				Enterprise	
		Small	Medium	Large	Extra Large	Large	
		1	2	3	4	8	სულ
11	Tbilisi	84	64	48	10	4	210
15	Adjara	288	280	236	13	4	821
23	Guria	312	168	324	13	4	821
26	Imereti	540	512	396	18	8	1474
29	Kakheti	624	672	860	141	20	2317
32	Mtskheta	264	264	316	13	4	861
35	Racha	204	160	240	4	4	612
38	Samegrelo-Zemo Svaneti	600	456	420	34	16	1526
41	Samtskhe-Javakheti	276	328	292	38	4	938
44	Kvemo Kartli	432	416	456	71	24	1399
47	Shida Kartli	348	304	340	65	12	1069
	Overall	3972	3624	3928	420	104	12048

Conclusion

Reliable agricultural statistics are essential for monitoring agricultural development, supporting evidence-based policy decisions, and ensuring food security planning at both national and regional levels. The sampling methodology implemented in the Survey of Agricultural Holdings of Georgia demonstrates how classical statistical sampling theory can be successfully adapted and applied within the context of official statistics production in a developing country.

The experience of Georgia illustrates that a carefully designed stratification framework — built upon a well-conceived composite indicator such as the agricultural index — significantly improves the quality, efficiency, and reliability of agricultural statistics. By converting diverse agricultural assets (land categories and livestock) into a single comparable unit benchmarked against maize cultivation per hectare, the agricultural index provides a pragmatic and theoretically grounded basis for stratification, even in the presence of the heterogeneous and highly skewed distributions typical of agricultural populations.

The two-stage stratified sampling design, incorporating geographic region, legal status, and holding size as stratification variables, yields 55 strata that adequately capture the structural diversity of Georgian agriculture. The application of the Dalenius-Hodge method for determining stratum boundaries. The adoption of Neyman (optimal) allocation for distributing the total sample size across strata represents a sound methodological decision, as it balances both stratum size and within-stratum variability. Compared to simpler approaches such as proportional allocation or allocation by standard deviation alone, Neyman allocation yields more statistically efficient estimates — a consideration of particular importance when survey resources are limited and precision requirements are high.

References

1. Nadareishvili M. New design of sample for the survey of agricultural holdings in Georgia. Rome: FAO; 2016.
2. Cochran WG. Sampling Techniques. 3rd ed. New York: John Wiley & Sons; 1977.
3. Dalenius T, Hodges JL. Minimum variance stratification. J Am Stat Assoc. 1959;54(285):88–101.
4. Lavallée P, Hidiroglou M. On the stratification of skewed populations. Survey Methodol. 1988;14:13–25.
5. FAO. World Programme for the Census of Agriculture 2020. Rome: FAO; 2017.