# Sampling techniques for sampling units with different size

*Mónika Galambosné Tiszberger*[1]

Usually it is a not too difficult problem to select efficient sample from a population, which includes units with different size. Stratified sampling might be a proper solution for this matter. The situation is getting more complicated if the statistician has to observe more characteristics of the unit, and these characteristics behave in various ways. Homogeneous strata cannot be created from every point of view. The field of my concrete research is the livestock surveys in Hungary.

If we "only" keep in mind the official requirements of the European Union, data in 6 categories of different livestock have to be provided by given reliability level. It means in practice, that one unique sample has to be worked out to suit these requirements. In Hungary there are more than 600 thousand private holdings, of which more than 50 percent raise usually more kinds of livestock. There are very small households, producing mainly for their own consumptions, and there are also huge units, which produce one company scale. The deviation of the indicators is extremely high in most of the cases.

In my research I attempt to see what would be the optimal solution from every point of view through working out different sampling schemes (simple random, stratified, and concentrated, mixture of these). In the presentation I would like to show the results and the final outcome of the work.

*Keywords: agriculture, sampling techniques, livestock survey*

## 1. Introduction

Statistical data concerning the livestock have been collected in Hungary since more than one hundred and fifty years. At the beginning the livestock was surveyed concurrently with the census of the population. From 1884 individual livestock surveys were carried out, while from 1895 livestock has always been part of the recurring agricultural censuses. (Laczka 2000) Since 1957 representative surveys of private farms have been conducted whereas all agricultural enterprises have always reported on their livestock. At the beginning the frequency of surveys was three monthly; until 2008 they were conducted in every four months (1 April, 1 August, 1 December); currently twice in a year (1 June, 1 December). The surveys cover all kind of animals. Full-scope observation applies in case of the agricultural enterprises; on the other hand there is sample survey for private holdings and households. After 1986 the breakdown of main animal species including the cattle stock by breed was surveyed following the Agricultural Census in 2000 (AC 2000).

[1] Mónika Galambosné Tiszberger, PhD student, University of Pécs, Faculty of Economics, Doctoral School of Regional Policy and Economics (Pécs)

In the international literature (to leave out of consideration the general literature dealing with statistics and sampling theory), in the topic of agriculture livestock is not in the focus of the interest. Crop statistics and area frame sampling in connection with soil and crops are the fields, which are worked out. Livestock data in most of the developed countries are coming from administrative data, so the way of collection and the sampling procedure is out of scope.

Before looking up details about the sampling techniques, it is necessary to summarize, what is the population that needs to be surveyed in case of agricultural statistics. There are two main groups involved in agricultural activity:

Agricultural enterprises: Every enterprise engaged in any agricultural activity, regardless its size. It is a business unit with or without legal entity excluding private entrepreneurs and private holdings.

Private holding: A technically and economically stand-alone production unit involved in agricultural activity, or holdings operated by private entrepreneurs, that used

- productive land (arable land, kitchen garden, orchards, vineyard, grassland, forest, reed, fishpond) of at least 1500 $m^2$ area, or
- orchard or vineyard of at least 500 $m^2$ area, or
- 100 $m^2$ land area under cover, or
- 50 $m^2$ of mushroom area during the reference year, or
- had a livestock consisting at least of
- one large animal (such as cattle, pig, horse, sheep, goat, buffalo), or
- 50 heads of poultry (such as hens, geese, ducks, turkeys, guinea fowls), or
- 25 heads each of rabbits, furred animals, pigeons, or
- 5 bee colonies

on the reference date of the survey.

The share of agricultural enterprises and private holdings in the agricultural value produced by the country is about equal, 50-50 percent. However the distribution of their number is not so balanced. There are about 7 700 active agricultural enterprises and more than 618 000 private holdings according to the latest Farm Structure Survey (FSS) in 2007. It means, that in number the share is 1-99 percent. Obviously the group of private holdings can be observed only through appropriately selected samples. As it is, I will deal only with private holdings throughout this article.

The aim of this paper is to present the beginning of a way, which tries to create an effective sample method to fulfil the EU requirements for the livestock surveys. The whole research is based on the data collected in FSS 2007. I have the opportunity to use the database at the regional office of the Hungarian Central Statistical Office (HCSO) in Pécs. As an important topic, the recent sampling techniques used by the HCSO will be introduced. A chapter will show the nature and characteristics of the Hungarian animal husbandry and the specialties of the different species. Then I have to summarize the requirements for the accuracy of the livestock data, which is determined by EU legislation, and which will be the bottleneck of the re-

search. I will present the analyses concerning the results of using simple random sampling by taking different aspects of the population, and the first trials on stratification. The work is not finished yet, so only partial results are indicated.

## 2. Livestock of Hungary

In Hungary the two main livestock types are cattle and pigs. Poultry is also important, but from the point of view of this paper it is not included in the analyses. In Table 1 I present the main data on livestock from the year 2007. In case of cattle and pigs, most of the livestock is kept by enterprises, which specialized for animal husbandry, and have more heads in average, kept by more efficiency. Sheep and goats are in the hand of private holdings. Sheep shows a very impressive average stock, but later on we will see, that the distribution of the livestock is not too fortunate from the point of view of sampling. Goats are in an even worse situation, as the average stock is quite small. Both of these two latter types show high deviations, as we will see in Chapter 7.

*Table 1.* Main data of livestock in Hungary, 2007

| Livestock | Country livestock total, heads | Livestock of agricultural enterprises, heads (percentage of total) | Livestock of private holdings, heads (percentage of total) | Livestock keeper private holdings, holdings | Average livestock in private holdings, heads |
|---|---|---|---|---|---|
| Cattle | 705 077 | 485 250 (69) | 219 827(31) | 18 907 | 11,63 |
| Pig | 3 871 147 | 2 603 958 (67) | 1 267 189(33) | 281 930 | 4,49 |
| Sheep | 1 232 005 | 172 660 (14) | 1 059 345(86) | 21 468 | 49,35 |
| Goat | 67 271 | 2 872 (4) | 64 399(96) | 15 380 | 4,19 |

*Source:* FSS 2007 and own calculation

I would like to give a picture about the structure of the Hungarian livestock not only by its size, but the value they represent. It is obvious that livestock heads cannot be added up. One might use the national livestock unit – which is an equivalent of the total livestock used for aggregation of various species of different genders, ages, equal to one or more animals of 500 kg live weight – to compare the amount of animals. But as the base of this indicator is the weight of the animals, from economic aspect it is rather meaningless. The best indicator would be the standard gross margin[2] (SGM), as it is an indicator, worked out by Eurostat, and used by every member state, but unfortunately, at present time I have no details about this at

[2] The SGM is equal to the unit production value of products and services net of variable costs.

the level of livestock, only at the level of the holdings. Gross production value[3] of agricultural products is the indicator of HCSO to measure agricultural value. According to the available data of 2000, the composition of the 4 types this article works with shows that pigs have a share of more than 50 percent. Cattle follow them, which represents about 1/3 of the value. Sheep and goat together gives only a bit more than 10 percent of the value produced by these 4 types on animals. These facts underline the original statement at the beginning of this chapter, so the two most import livestock are cattle and pigs.

It is also important from the point of view of sampling to see the distribution of livestock holdings and of the livestock itself of the different animal species.

- 92 percent of the pig keeper holdings have less than 11 pigs, and they give almost half of the livestock. It means that if we would be able to cover the remaining 8 percent – which represents not too many – of the holders, it would give the other half of the livestock. The skewness of the distribution is notable.
- In case of cattle, holdings having 10 or less animals constitute almost 80 percent of the cattle keepers, but have a share of the livestock of only 28 percent. The distribution in the higher sections shows various pictures.
- Sheep stock represents more concentrated production. 15 percent of the biggest holders (having more than 50 heads) give more than 80 percent of sheep. Unfortunately the remaining holdings are very diverse from the aspect of the size of their livestock. The distribution of the livestock is skewed, but the opposite way as we saw in case of pigs.
- Goats are the most special ones. The mode of the held heads is 2 and 3. It means that statistician cannot really gain from the observation of bigger keepers, because there are only a few of them, and they represent just a few percent of the total production of goats.

Altogether the distribution of the different species shows different nature. It would cause problems during stratification and in the situation where I would like to create a combined sampling plan, which works for every species.

## 3. Sampling method today

Regular surveys mean those implemented every year (survey on sown area, annual production, crops, etc.) or even several times within a year (livestock). Surveyors

[3] Value of agricultural products produced in the framework of agricultural production in a certain time period, irrespective of whether those were produced in a unit in agricultural or in other branch. This value includes the value of two main branches of agricultural: crop production, animal husbandry. Gross production value of animal husbandry includes the production value of breeding (live born animals), value of livestock change and weight growth and values of products and by-products from animal husbandry. Gross production value is the sum value of total amount of produced products multiplied by average prices determined for each utilisation types.

visit the selected private holdings to fill in the questionnaires (face-to-face interviews). The sample frame is based on the AC 2000, and it is updated with the Census on vineyards and orchards' (2001) and the farm structure surveys' results. The frame population is divided into two groups according to the size of the holding. A smaller group, declared as large holdings, is selected on a full scope base, according to my initiatives, and they receive the questionnaire by mail. This preferential group is selected through natural figures like the size of livestock or land area. (Altogether about 1500 private holding belong to this group. Out of it 700 are livestock holdings.) The specific thresholds in the selection of this group were adjusted according to their share, paying attention to the financial possibilities. This kind of "take all" philosophy in case of large holdings results in more reliable figures, as a notable part of the production of private holdings is observed without sampling errors. Another advantage of my initiative was to introduce the data collection by mail to the respondents, and start a process to generalize and make this form of surveys acceptable among them.

The rest of the holdings (more than 900 thousand holdings as a frame population) form the base for sampling. A universal sample is used to keep the surveys' budget on a cost-effective level. It means that for all of the above-mentioned surveys only one sample is selected, and there are no special sample population for the different types of surveys.

A two-stage, concentrated, stratified sampling technique is applied. Organisational and financial reasons made it necessary to use concentration, which is theoretically not optimal for minimizing the sampling error with a given sample size. However, the face-to-face interviews are much more economical and faster if the surveyor has to visit holdings within a small district (a part of a settlement), instead of travelling kilometres to find the different data suppliers. It is also easier and more effective to organise and manage a smaller number of surveyors within the regions.

The sample selection implemented through the following steps: In the first stage every 9th survey district is selected randomly, stratified by county (on this NUTS III level there are 19 counties in Hungary). These districts are the primary sampling units (PSU). In these selected districts 2 strata are determined:

- Stratum "A": all holdings exceeding at least one of the following thresholds: 5 cattle, 10 pigs, 26 sheep, 100 chickens, 100 ducks, 100 turkeys, 26 geese, 25 bee colonies, 5 ha arable land, 1 ha vineyard or 1 ha orchard.
- Stratum "B": holdings not exceeding the thresholds mentioned above.

The secondary (of final) sampling units (SSU) are the private holdings within the selected PSUs. Every holding is selected in stratum "A", and randomly every 4th is in the sample in case of stratum "B". It results about 40 thousand holdings in the sample population. It means a sampling rate of about 4 percent. The size can be easily adjusted to financial possibilities or quality requirements by changing the sampling rate of either PSUs or SSUs in stratum "B". (Previously every 8th district was in the sample and the sampling rate was 33 percent in stratum "B".) New sample is

selected after the bigger surveys like censuses of farm structure surveys. It means that one particular private holding will be a data supplier for 3-4 years, and then the sample is refreshed.

The main problems with this sample design, as I see, are the following:

- Counties, or in the future regions as strata are necessary, because the main figures are published at county – in the future at regional – level, but obviously they do not form homogeneous groups. This aspect is an obligation (which is problematic, but it can not be "solved").
- The stratification on the second level does not serve the aim of building homogenous groups as well, because of the large number of variables used. Thus the variances of the different variables within the groups are still high.
- Concentration effects worse sampling errors with the same sample size compared to simple random sampling without concentration. However, I must admit, as long as there are face-to-face interviews, this part unfortunately cannot be modified in practice.
- The universality of the sample is the weakest point as I see. The effectiveness of the sampling cannot be sufficient if it has to cover so many topics and characteristics.

## 4. The requirements

After the coming census in 2010 new sample selection would be necessary. It would be a convenient solution to continue the method used in the previous 10 years. However, there is a new legislation, which declares clearly the allowed maximum relative standard errors for the main livestock types. This EU regulation[4] from the year 2008 requires the following relative standard error by 68 percent probability level for the country totals:

[4] Regulation (EC) No 1165/2008 of the European Parliament and of the Council of 19 November 2008 concerning livestock and meat statistics and repealing Council Directives 93/23/EEC, 93/24/EEC and 93/25/EEC.

*Table 2.* Maximum relative standard errors according to EU legislation

| Livestock | Maximum relative standard error |
|---|---|
| Cattle | 5% |
| Cow | 5% |
| Pig | 2% |
| Sheep | 2% |
| Goat | 5% |

Source: Regulation (EC) No 1165/2008 of the European Parliament and of the Council

As sampling techniques are used only in case of private holdings, the relative standard error (coming from the nature of sampling) can be concerned as 0. It means, that by taking into account the distribution of the livestock totals per holding type, we can define the maximum relative standard errors for the livestock total of the private holdings (which will be obviously higher, less strict at the end). I have collected the necessary information for these calculations from the FSS 2007.

*Table 3.* Data of the main livestock, 2007

| Livestock | Country livestock total | Livestock of agricultural enterprises | Livestock of private holdings | "Healing" weight of the livestock total of enterprises | Maximum relative standard error for the livestock total of private holdings |
|---|---|---|---|---|---|
| Cattle | 705 077 | 485 250 | 219 827 | 0,312 | **16,04%** |
| Cow | 322 369 | 225 477 | 96 892 | 0,301 | **16,64%** |
| Pig | 3 871 147 | 2 603 958 | 1 267 189 | 0,327 | **6,11%** |
| Sheep | 1 232 005 | 172 660 | 1 059 345 | 0,860 | **2,33%** |
| Goat | 67 271 | 2 872 | 64 399 | 0,957 | **5,22%** |

*Source:* FSS 2007 and own calculation

The "healing" weight of the livestock total of agricultural enterprises comes from the following relationship (Ay 1976):

$$V = \sqrt{\frac{X_{ae}^2 \times 0 + X_{ph}^2 \times V_{ph}^2}{X^2}}$$

where:
$V_{ph}$: relative standard error of private holdings
V: required relative standard error
$X_{ae}$: total of agricultural enterprises' livestock
$X_{ph}$: total of private holdings' livestock
X: country total (livestock)

As the required error limits are given to the country totals of the main livestock types, the maximum relative standard error allowed for the livestock total of the private holdings can be calculated by rearranging the above-mentioned equation. (The results are in the last column of Table 3.)

$$V_{eg} = \sqrt{\frac{V^2 \times X^2}{X_{eg}^2}}$$

Apparently the estimations of the private holders' livestock have more latitude in those cases where the livestock is held mainly by the agricultural enterprises. The biggest challenge is stated for the number of sheep and goat. The modest size of the livestock kept by the enterprises does not really ease the strict requirements.

## 5. The base of the research

It would have been really nice to do every calculation and simulation on an up-to-date population of agricultural private holdings, but for the nature of statistical work, it is off course impossible. The latest full scope survey has been carried out in 2000.

Naturally the database of the registered 958 534 private holdings is available at HCSO, but the agricultural sector shows rapid changes in the last 10 years (35 percent of the private holdings had disappeared), this database is not good enough to simulate the current situation. On the ground of these reasons, I decided to be satisfied with a smaller part of the population, but with more recent data. The latest "bigger" survey has been carried out in 2007 (FSS 2007). The sampling rate was about 18 percent. Concentrated, one-stage sampling design had been worked out for this FSS. Every 6$^{th}$ survey district was selected by simple random within the regions (NUTS II level). During the survey, the enumerators looked up the whole district. It means, that they have asked every household about their agricultural activity, and not only those, who were on their list. It gave us the opportunity to collect information about the new agricultural holdings (in the selected districts) as well, and not only about those who were previously surveyed. In my research I use this database of FSS 2007, which includes almost 111 thousand private holdings. The size of the sample and the way of selection ensure that this sample population shows practically the same distributions according to the different variables as the frame population, even on regional level (NUTS II). To justify these statements, I investigated the distribution of the sample population by regions, and compared it to the official results of FSS 2007. The differences of the proportions are under 3.7 percentage points, and

the sample rate is also similar in the regions. According to these results, I believe, that general conclusions can be drawn[5] from the analyses of the sample population.

The final aim of the calculations is to work out an optimal sampling design. By optimal I mean to stay within the resent sample size (30-40 thousand private holdings), and ensure the fulfilment of the EU requirements about the maximum relative standard errors for this 5 specific livestock. As I work according to theoretical possibilities, I would handle the costs and management of the survey as secondary elements. During my work in HCSO I had to work within the financial and organisational frames, so I have never had the opportunity to try the statistical theory in practise. Naturally I hope, that HCSO would be able to build in my results into the sampling designs of the future.

## 6. Special features of livestock

In general the livestock holdings keep 1,86 animal types at one time. (It includes every type of livestock, in agricultural meaning.) It is good news from the point of view of sampling, because it means that the sum of the necessary sample size by animal species would give a good approximation, as most of the holdings keep only one of the 5 animals. It also predict, that stratification by animal types would be part of the optimal solution.

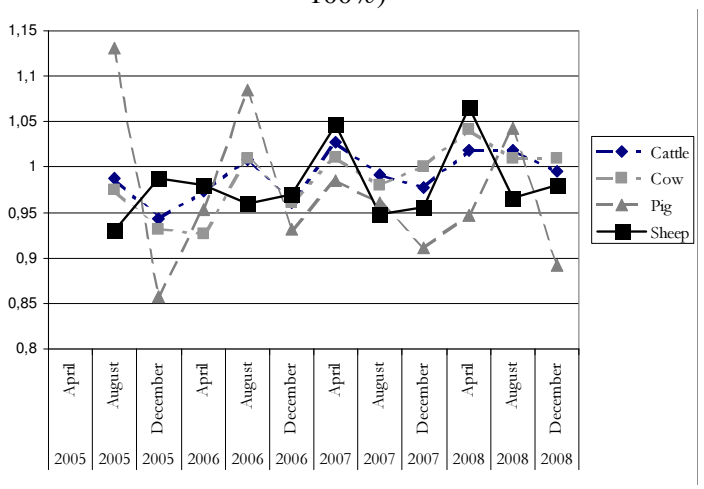
While we plan a sample design, there is another important aspect to take into account: how often does a livestock holding change its activity (change the livestock type, or close the business). Off course the stable livestock holdings would be ideal from the point of view of sampling. If they remain unchanged in time, the stratification and the sampling errors can be predicted in a more effective way. To analyse this question, it would be nice to see the whole life circle of at least some of the private holdings. Full scope information is available only from the year 2000. In the following years only samples had been surveyed. The FSS 2003 can be one more guideline because the holdings above 1 ESU[6] had been observed on full scope base. But altogether the behaviour of the holdings can be followed only in very few cases. The estimated results of my own calculations show, that between 2000 and 2007 the permanent livestock holdings give 70-90 percent of the total. In my understanding these numbers suggest that the livestock keepers are a quite stable group. It means that an optimal sampling design would be efficient in a longer period.

[5] In the formulas I use capital letters in most of the cases, because, as it is mentioned earlier, I regard the FSS 2007 as a whole frame population. So the calculated figures – mean, deviation, etc. – are understood for the entire population, and not as a sample variable.

[6] The economic size of the holding is determined on the basis of the total SGM value of products and activities of the holding and expressed in European Size Unit (ESU), where one ESU worth of SGM is equal to 1200 €.

Besides these aspects, there is one more factor that the statistician has to be aware of. Animal husbandry is a seasonal phenomenon. The size of the livestock is different in the to sampling period within the year. (As an illustration of the seasonal characteristic, I present Graph 1.) It is natural, but during the work we must pay attention to this fact.

*Graph 1.* Seasonal changes of livestock of private holdings (previous season = 100%)




*Source:* www.ksh.hu (Agricultural long time series and censuses)

## 7. Background and possible methods of sampling

There are basically two sources of the error in sample surveys. One of them is the random sampling error, which comes from the fact that only a part of the whole population is observed. The other source is the bias, which comes from the biased estimator. Sampling error cannot be avoided, but the bias can be overcome by using unbiased estimator. (unbiased = the estimator's expected value equals the variable that wished to be estimated). Simple random sampling has the advantage – contrary to the non-random samples – that the error limit of the estimations for the population can be calculated by exact methods from the sample itself. It can be measured by mathematical calculations and through the "arbitrary" increase of the sample size it might be decreased for the needed level.

The other group of errors is the branch of non-sampling errors. These may occur in different ways during an observation:

- Uncertain information of the populations (the level of coverage is under 100 percent).

- Respondents provide uncertain data about themselves.
- "Clerical error" during the survey, or the data entry.

It seems quite obvious, that using administrative data sources, undercoverage can be controlled and corrected in the registers. Unfortunately there are fields, where this kind of source is unavailable or includes false, not up to date information. In case of the other two types, the control rules, blind checks, educated enumerators might help to avoid the mistakes, but it would be unreasonable to expect 100 percent elimination of these biases. During the practical part of the surveys, these items must be kept in mind, but as such mistakes are hard to calculate, in my research I do not have the opportunity to cover them.

As I consider the sampling error as the strictest condition, I started my work by the analyses of the deviations of the different livestock species. Although the final aim is to estimate the sum of the values, to keep the calculations simple I will do the work first for the averages. From the point of view of relative deviation and the necessary sample size, the results would be the same, anyway.

I have used the formula of the simple arithmetic mean to calculate the average livestock of one holding:

$$\overline{X} = \frac{\sum_{i=1}^{N} X_i}{N}$$

where:
$X_i$: total of private holdings' livestock
N: population size (holdings)

The formula of the deviation:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} \left(X_i - \overline{X}\right)^2}{N}}$$

The formula of the relative deviation:

$$RD = \frac{\sigma}{\overline{X}}$$

These are simple formulas, however the calculations are a bit more complicated in the practise. The livestock table of HCSO database includes only those

holdings, which had kept livestock at the reference date of the survey (1st December 2007). It means, that additional calculations are necessary to reach the variables (average, deviation) for the entire population. The number of holdings not keeping animals is important, because they are part of the population (N), and their group will decrease the average, and most probably will increase the deviation. To make this point more visible, I created three different tables (Table 4, 5 and 6) with the same indicators in them. The three different tables concerns three different aspects of the population.

*Table 4.* Data of the population

| Livestock | Average livestock of 1 holding | Deviation | Relative deviation | Number of holdings |
|---|---|---|---|---|
| Cattle | 0,25 | 2,48 | 981,6% | 110 949 |
| Cow | 0,11 | 1,22 | 1 138,4% | |
| Pig | 2,04 | 8,57 | 420,5% | |
| Sheep | 1,34 | 18,83 | 1 407,6% | |
| Goat | 0,10 | 1,78 | 1 868,3% | |

*Source:* own calculation

*Table 5.* Data of livestock holdings

| Livestock | Average livestock of 1 holding | Deviation | Relative deviation | Number of holdings |
|---|---|---|---|---|
| Cattle | 0,35 | 2,92 | 831,0% | 79 837 |
| Cow | 0,15 | 1,44 | 964,2% | |
| Pig | 2,83 | 9,99 | 352,8% | |
| Sheep | 1,86 | 22,17 | 1 192,9% | |
| Goat | 0,13 | 2,09 | 1 584,0% | |

*Source:* own calculation

*Table 6.* Data of livestock holdings with the specific livestock species

| Livestock | Average livestock of 1 holding | Deviation | Relative deviation | Number of holdings |
|---|---|---|---|---|
| Cattle | 8,13 | 11,57 | 142,3% | 3 447 |
| Cow | 4,89 | 6,78 | 138,6% | 2 438 |
| Pig | 4,41 | 12,18 | 276,4% | 51 293 |
| Sheep | 38,56 | 93,72 | 243,0% | 3 848 |
| Goat | 3,87 | 10,67 | 275,6% | 2 725 |

*Source:* own calculation

We can see, that the relative deviations are getting smaller by narrowing the aspect of the population. It is also visible, that the degree of the decrease is relatively small between Table 4 and 5, but we can get much more homogeneous groups by handling the livestock holdings having the specific livestock specie separately. Unfortunately the best results (Table 6) would be the less useable in practise. I would need one sample to estimate every species, and the handling of different populations within one process would be rather insolvable. The other problem is that there are overlapping among the groups.

### *7.1. Simple random sampling*

If we know the deviation of the variables, we can easily calculate the necessary sample size to fulfil the requirements. I started the analyses by simple random sampling. Simple random samples require known probabilities (non zero) of selection for every element. In this case we get the necessary sample size by using the following formula[7]:

$$n = \frac{\sigma^2}{\Delta^2 + \frac{\sigma^2}{N}}$$

where

- n: sample size (holdings)
- Δ: accepted maximum error limit – $V_{ph} \times \overline{X}$ (heads)

I have calculated the necessary sample sizes for the different aspects of the population (entire, livestock holdings, livestock holdings with the specific livestock

[7] The probability level is 68%.

type). It seems quite obvious, as the relative deviations decrease thanks to the narrowed concept of the population that the necessary sample size would decrease as well. As I work with a research database, I think the sampling rate will illustrate the differences and the volume much better than the number of the holdings. In every case in the tables of the article I divide the necessary sample size by the original entire population (n/N), to make the comparison possible. The sampling rates can be seen in Table 7.

*Table 7.* Necessary sampling rate from the total population

| Livestock | Total population | Livestock holdings | Livestock holdings with the specific livestock type |
|---|---|---|---|
| Cattle | 3,27% | 2,36% | 0,07% |
| Cow | 4,05% | 2,94% | 0,06% |
| Pig | 4,09% | 2,92% | 1,77% |
| Sheep | 76,69% | 70,26% | 2,56% |
| Goat | 53,59% | 45,35% | 1,24% |

*Source:* own calculation

In case of cattle and pigs, the sampling rates seem to be acceptable even at the first stage. However if we take a look at sheep and goats, we can see that only the last column of Table 7 gives us acceptable sampling rates. In the first and second case the sampling rates of these types are around of even above 50 percent. Such a sampling rate does not make any sense to use in a real survey. Hunyadi (2001a) introduces a similar example, where he looks for an optimal sampling design of the average production of 3 types of wheat. In the example we find that wheat "A" has a relatively small weight in the total production but with relatively high deviation. In this case, because of the heterogeneity of the produced amount, extremely high sampling rate would be "optimal". At the same time, as its proportion of the total production is quite small, Hunyadi suggests to be satisfied with smaller sample in these cases, and let the deviation (sampling error) remain high, to avoid the unnecessarily big sample size. In our case of livestock surveys, the translation of this example would mean that as in the Hungarian livestock sheep and goats are not so important, and their share is quite small (compared to cattle and pigs), and the livestock is very heterogeneous, we should be satisfied with higher sampling errors during the estimation of the total of these two types. As a result, we would be able to keep the sample size within reasonable limits. Sure enough, but unfortunately the EU legislation clearly defines the required maximum sampling errors, and do not take into account

the importance and proportion of the given livestock within the country's livestock production. So we must investigate other solutions.

### *7.2. Stratification*

The next classical topic of sampling theory is stratification. Stratification has the precondition of knowing the entire population from the point of view of the stratification characteristic. It is also important, that every item of the population is classified into only one stratum, and every item can be classified. In an optimal case the best grouping variable would be the one that we are about to observe. Obviously from this variable we do not have complete information, as the final aim is just to collect it. (Cochran 1977, Kish 1995) In a general agricultural survey, which is multipurpose and the aim of the observations is not just a few indicators, it is a real problem to find good variables for stratification. (Kish 1989) In case of livestock surveys it is especially difficult to find a good variable, which will divide the entire population into distinct subpopulations. Holdings, who keep more types in parallel, might be part of more subpopulations if we use livestock type as the base of stratification. This is the reason why in practise we usually formulate more conditions at once to create distinct groups of smaller and larger holdings within the population. By using more variables for stratification there would be no overlapping of the subpopulations, but the original aim of stratification seems to be lost. Namely, we plan to use stratification to decrease the variances of the sample estimates. The original consideration standing behind stratification is to divide the heterogeneous population into homogeneous subpopulations. However if a holding would be large according to the number of its cattle, it is quite probably, that it would belong to the smaller holdings from the point of view of pigs or sheep. In practise it is general, that holdings are specialized in something, and not keep every type of livestock in big amount. So it results in high deviations for every variable, because the original aim of creating homogeneous groups cannot be reached. The deviation of the variables does not really decreased by this multivariable stratification. This is one part of the problems.

#### 7.2.1. Current stratification

The stratification, which HCSO currently applies, is introduced in Chapter 3. To give the reader an idea of how does these aspects classify the population I present the distribution of holdings by the current stratification, giving extra information on livestock and land use aspects within the strata. Looking at Table 8 we see, that most of the holdings (86,9 percent) remained in one stratum, stratum "B", the smaller holdings. So we must conclude that homogeneous groups are not created. The share of the different aspects, namely livestock and land use, also confirms remained heterogeneity. Only 2,7 percent of the holding can be regarded as "large" holding by

both branches of agriculture. The rest 10,3 percent is specialized in one of them, so putting them into one stratum will not result in homogeneity.

*Table 8.* Distribution of holdings according to different aspects of stratification, 2007

| **Strata** | „A" according to livestock | „B" according to livestock |
|---|---|---|
| „A" according to land use | 2,7% | 6,2% |
| „B" according to land use | 4,1% | 86,9% |

*Source:* own calculation

If we look at the values behind the holdings (Table 9), we find that the produced standard gross margin shows a bit better distribution by stratum "A" and "B", but still, more than half of the production is kept in one group.

*Table 9.* Distribution of SGM according to different aspects of stratification, 2007

| **Strata** | „A" according to livestock | „B" according to livestock |
|---|---|---|
| „A" according to land use | 6,7% | 9,6% |
| „B" according to land use | 30,9% | 52,9% |

*Source:* own calculation

Another important point about the recent stratification method is its change in time. It was mentioned earlier, that the elements of the population could be divided into the subpopulations only if we know the value of their variables in advance. Let us take a look at the current stratification system in 2000 and 2007. From Table 10 we can conclude that the proportions of the strata has changed a lot in time, and holdings has moved from one stratum to the other in a lot of cases. It means, that the agricultural activity is still changing rapidly in Hungary and the classification of holdings by 7-year-old data results in misleading proportions. This aspect arises the thought of post stratification or two-phase sampling, as possible solutions in the future. *(The analyses of these two further methods are not part of this paper, but of later research.)*

*Table 10.* Change of stratification in time

| | **Based on AC 2000** | **Based on FSS 2007** |
|---|---|---|
| **Stratum** | **Number of private holdings** | |
| "A" | 29 805 | 17 547 |
| "B" | 60 073 | 101 038 |
| Non-classified[8] | 22 152 | - |

*Source:* own calculation

7.2.2. Stratification by regions

The next part is the required territorial breakdown of the provided data, which is very import for the domestic data users. At least regional (NUTS II) level data is required for the main livestock. It means for the statistician, that region would be a stratifying variable no matter of its efficiency in creating homogeneous groups. The 7 regions of Hungary are subpopulations as an obligation. The statistician must start the whole work within these frames. Additional stratification can be used, off course, but we have to balance the number of strata within reasonable limits. It would be inefficient to create too many subpopulations (including only a few holdings). As the stratification aspects are multiplied, we do not have the option to come out with too many ideas.

After computing the deviations for the different regions, I had the possibility to analyse the necessary sample sizes for a stratified sample. I did the calculations by conditioning simple random selection within the subpopulations (regions). I have used the following formula of relative sampling error[9] for stratification by proportionate allocation:

$$V_{RRA} = \sqrt{\sum_{j=1}^{7}\left(\frac{N_j}{N}\right)^2 \times \frac{\sigma_j^2}{n_j}\left(1-\frac{n_j}{N_j}\right)}$$

where[10]

- j: regions
-

It was not possible to rearrange the formula, and give an exact equation for the sample size in this case. So during the calculations I used iterations to see, what would be the necessary sample size. The results can be found in Table 11, where I

[8] Non-classified holdings did not exist in the year 2000.

[9] These general formulas of arithmetic mean, deviation, sampling error in case of simple random and stratified sampling can be find in several books, like: Kish (1995), Hunyadi-Vita (2003) or Pintér-Rappai (2007).

[10] In the formula I use the known deviation, as I work from the database of the entire population.

have included the former results to be able to compare the sampling rates. Unfortunately we find similar, or even higher percentages for the stratification by proportionate allocation as by the above mentioned simple random sample of the entire population. It means that the obligation of applying regions as strata did not increase the efficiency at all. Because of the heterogeneous regions, this territorial breakdown does not have good characteristics from the point of view of stratification. So, although it is not a surprise, this result is quite sad. We have already used partly one of our weapons (stratification) to make the sampling method more effective, but we gained basically nothing.

*Table 11.* Necessary sampling rates in different aspects

| **Livestock** | **Population** | **Stratification by regions (proportionate)** |
|---|---|---|
| Cattle | 3,27% | 3,3% |
| Cow | 4,05% | 4,3% |
| Pig | 4,09% | 4,1% |
| Sheep | 76,69% | 77% |
| Goat | 53,59% | 54% |

*Source:* own calculation

I did not give in at this point. I tried another type of allocation to see if it is going to show better results for the regional stratification. If the aim is to decrease the sample size and keeping the sampling error at the same level, Neyman optimal allocation by using the deviations might lead us for better solution. The essence of this type of allocation is to have higher sampling rate in the more heterogeneous regions to ensure more accuracy at the end. The following formula shows the allocation of the sample size:

$$n_j = n \times \frac{N_j \sigma_j}{\sum_{j=1}^{7} N_j \sigma_j}$$

The sampling error is the same, as it was in case of proportionate allocation. In this case I applied iterative approach, by changing the sample size (n) until I have reached the necessary sampling error.

Neyman optimal method gives better allocation design, if the deviation of the variables shows big variety among the regions. It is only true for sheep and goats. Regions have very similar deviations for pigs and cattle. Table 12 justify these facts. Compared to the proportionate allocation Neyman optimal allocation decreases the sampling rate in notable way only in case of sheep and goat. However from these

results we must conclude that Neyman optimal allocation overall is better than proportionate.

*Table 12.* Stratification by regions with different allocation strategies, necessary sampling rates

| Livestock | Proportionate | Neyman optimal |
|---|---|---|
| Cattle | 3,3% | 3,2% |
| Cow | 4,3% | 4,2% |
| Pig | 4,1% | 3,7% |
| Sheep | 77,0% | 66,5% |
| Goat | 54,0% | 44,0% |

*Source:* own calculation

## 8. Conclusions

From the article, it is quite clear that sampling of units with different size and nature is very difficult. The diverse distribution of the animal species makes it hard to harmonize the information into one sample. The high values of deviation require too high sampling rates to reach good quality estimations. The narrowed aspects of the population gave promising results. The handling of livestock holdings as the frame populations gives better starting as taking every agricultural private holding. The conditions of the work are also very strict. The necessary sampling errors are probably too small for sheep and goat, and the requirements of Eurostat do not take into account the importance of the specific livestock species in the different countries. The necessity of regional data also restricts the work of the statistician. Regions do not create homogeneous strata, but give an additional extra aspect for the stratification.

Administrative data sources would ease every problem, but so far in the field of agriculture these sources do not have the appropriate quality for statistical purposes.

The final aim of my research is not reached yet. The results so far show that the working out of different stratification plans has to be continued. The distribution of the animal species suggests that stratification by size would be effective, but the thresholds and the way of building distinct groups are still questions to answer. Post stratification and two-phase sampling are also techniques that have to be investigated to the matter in hand.

I plan to finish the research before the census of 2010, to be able to offer the results for practical application to the sample selection of 2010 at the Department of Agriculture of HCSO. It would be also important to widen the scope of the investi-

gation into land use. The results of the coming census will give new database for the further work.

*References*

Ay, J. 1976: A *mintavételes állatösszeírások módszertani kérdései.* Kandidátusi értekezés, Budapest. 150 p.

Cochran, W. G. 1977: *Sampling Techniques.* Wiley & Sons, Inc., New York. 428 p.

FAO 1995: "*Programme for the World Census of Agriculture 2000*", Statistical Development Series 5, Rome.

Hunyadi, L. 2001a: *A mintavétel alapjai.* SZÁMALK Kiadó, Budapest. 96 p.

Hunyadi, L. 2001b: *Statisztikai következtetéselmélet közgazdászoknak.* Központi Statisztikai Hivatal, Budapest. 483 p.

Hunyadi, L. – Vita L. 2003: *Statisztika közgazdászoknak.* 2. kiad. Központi Statisztikai Hivatal, Budapest. 770 p.

Kish, L. 1989: *Sampling methods for agricultural surveys.* Food and Agriculture Organization of the United Nations, Rome. 261 p.

Kish, L. 1995: *Survey sampling.* Wiley & Sons, Inc., New York. 643 p.

Laczka, S. 2000: *Mezőgazdasági összeírások Magyarországon, 1895-2000* = Statisztikai Szemle 78. évf. 4. sz. 283-289. p.

Pintér, J. – Rappai, G. (editors) 2007: *Statisztika.* Pécsi Tudományegyetem Közgazdaságtudományi Kar, Pécs. 508 p.

Regulation (EC) No 1165/2008 of the European Parliament and of the Council of 19 November 2008 concerning *livestock and meat statistics* and repealing Council Directives 93/23/EEC, 93/24/EEC and 93/25/EEC.

www.ksh.hu (10.08.2009 – 31.12.2009)