NEW APPROACH TO THE ESTIMATION OF DISPERSION USED IN PROJECTING THE SAMPLE VOLUME: MARGINAL COEFFICIENTS IN SAMPLING BASED AGRICULTURAL RESEARCH

Daniela Şerban[1], Nicolae Istudor[1], Dan Boboc[1], Simona Nicoleta Vasilache[1]

ABSTRACT

Sampling techniques are largely used in agricultural research. This paper presents a new way to establish the sample size in the case of stratified sampling using new measures called by authors marginal partial variances: inter-strata and average marginal dispersions. The new indicators were developed in a way similar to the one classical Keynesian Economics presents the marginal measures. Estimation of partial dispersion using the marginal partial dispersions for stratified samples and determining the sample size based on this, allows projecting the research resources. For projecting a new sample in periodical samplings we need to know the dispersion trend and to estimate the possibility to forecast the dispersion and the standard deviation. It stems from here the necessity to built time series for dispersion and for its absolute/ relative variance, for instance, chronological series of dispersion indices or of increases in dispersion for a variable recorded at different times. The possibility to forecast the dispersion based on marginal indicators adds a new restriction, other than costs, to estimating the volume of the new sample.

Key words: average marginal dispersion, inter-strata marginal dispersion, production evaluation, stratified sampling.

INTRODUCTION

Sampling, as research tool in agriculture, captures a static image of reality; this is why agricultural experiments have to be periodically repeated, with the same set of recorded variables (Mitruţ and Şerban, 2007). The importance of approaching statistically agricultural experiments was underlined 50 years ago by Finney (1956), in comparing yields of different varieties of plants. The strengths using sampling in agricultural experiments refer to resources effectiveness, and to the possibility of investigating fast and cheaply populations which are hardly accessible or perishable, as it is the case, on the one hand, with field plots, which can't be repeated too many times, because of financial constraints, and, on the other hand, with nursery plots, in greenhouses, for instance, when the samples used are highly perishable (Motis, 2003). Variability is a characteristic of agricultural data considers Cobanovic (2002), because data can vary in respect to type of land (slope, humidity, solar exposure), or in time in repeated cultures. The use of samplings in agriculture is not at all a recent trend. Houseman and Becker (1969) stated that the first agricultural census took place in 1840, in America. The problem of projecting the sample which yields the most realistic results is, then, put forward. In 1925 regression methods started to be used in practice, in order to trace the dependence of some variables on other variables (for instance, the production per acre dependent on the irrigation norm, the quantity of natural or chemical fertilizers, exposing to sun etc.) By the end of the '30s it was noticed that, between the statistically estimated results and the ones which were practically recorded in field research there are significant differences, which raised problems concerning the design of the sample and the survey methodology. Jensen (1939) published a paper based on an Iowa state survey, explaining how one can design a survey in such a way as to obtain the best estimators, which is a matter of sampling design. Kish (1965) introduced a new indicator, the so-called *design effect*, defined as the ratio between the variance of the estimator for the chosen sample and the variance of a same size random sample. David (1966) published a PhD thesis, where issues of sample design, as stratification, selection and estimation, are related to time series. In 1998, a manual named *„Multiple frame agricultural surveys"* is published in FAO Statistical Development Series, where the most suitable statistical methods, considering the purpose of the research, are presented.

Sampling designing, in agricultural surveys, should start from a present state analysis (Cotter and Tomczak, 1994), by which the units to be included in the survey are determined, as well as their geographic distribution. Depending on the allocated budget, the data

[1] Academy of Economic Studies Bucharest, Piaţa Romană no. 6, District 1, Bucharest, Romania

needs of the sampling beneficiaries should be prioritized (Prejmerean, 2007). In the sampling stage the researcher chooses the adequate sample type, sets the procedure for processing the sampling units, the level of tolerable errors and the degree of significance of the results. From the various available sampling types, in agricultural surveys multistratified random sampling is usually preferred. It can be seen from here that the calibration of the sample has chain effects on all the statistical processing afterwards and, consequently, on the reliability of the results, based on the sampling indicators and on the statistical inference achieved. The size of a sample needed for a sampling depends, mainly, on the structure and homogeneity of the studied population which, ultimately, influences the quality of statistic information and of the sampling-based estimates. This explains its clearly defined place in the theory of probabilities.

MATERIAL AND METHODS

We develop here the approach of "measuring measuring errors" (Healy, 1989), in the sense that we propose an innovative method for assessing the sources of variance, in the samples of a repeated agricultural experiment. The results of a stratified sampling lead to the necessity of verifying the rule of addition for dispersions, according to which the total variance measured by the total dispersion is the sum of the partial dispersions: the dispersion inside the strata, determined by the stratification criterion, and the inter-strata dispersion, which shows how the variable is modified from one stratum to another. The relationship between the three dispersions (Biji, 1979) is:

$$\sigma^2_{total} = \overline{\sigma}^2 + \sigma^2_{y/x}, \qquad (1)$$

where,

σ^2_{total} = total dispersion, determined by all the influence factors of a variable's variance;

$\overline{\sigma}^2$ = average dispersion, determined by unrecorded factors;

$\sigma^2_{y/x}$ = inter-strata dispersion, due to strata formation factor, showing how much it does discriminate the studied variable.

If we divide each term of the above equation by the total dispersion, computing the structure of the total dispersion, we obtain the determination and the non-determination ration, following the formula:

$$1 = \frac{\overline{\sigma}^2}{\sigma^2_{total}} + \frac{\sigma^2_{y/x}}{\sigma^2_{total}} = K^2 + R^2, \qquad (2)$$

where,

$\frac{\overline{\sigma}^2}{\sigma^2_{total}} = K^2$ = non-determination ratio, expressing, in percents, the weight of the non-recorded, random factors in the total variance;

$\frac{\sigma^2_{y/x}}{\sigma^2_{total}} = R^2$ = determination ratio, which expresses, in percents, the weight in the total variance of the dependent variable explained by the grouping factor, the discrimination factor and the independent variable.

If we record two dispersion levels for two successive samplings, we may compute the absolute change in total dispersion, which will be distributed between the absolute change in inter-strata dispersion and the average dispersion inside the strata, as follows:

$$\Delta\sigma^2_{total} = \Delta\overline{\sigma}^2 + \Delta\sigma^2_{y/x} \qquad (3)$$

This variance of the total dispersion may be positive and negative, and can be distributed equally or differently on its two components. In order to measure its distribution patterns and to determine the contribution of the stratification factor to the total dispersion variance, we compute the structure of the equation of absolute change in dispersion, by dividing the equation of absolute changes by the absolute change in total dispersion, as follows:

$$1 = \frac{\Delta\overline{\sigma}^2}{\Delta\sigma^2_{total}} + \frac{\Delta\sigma^2_{y/x}}{\Delta\sigma^2_{total}}, \qquad (4)$$

In the above equation we propose that the marginal indicators obtained are noted, named and interpreted as follows:

a. $\bar{\sigma}^2_{mg} = \frac{\Delta\bar{\sigma}^2}{\Delta\sigma^2_{total}}$ = average marginal dispersion, which shows how much will the average of inside strata dispersions be modified if the total dispersion changes by one unit, or how much should the average of inside strata dispersions be modified in order to obtain one unit change in the total dispersion; can take values between -1 and 1.

b. $\sigma^2_{y/xmg} = \frac{\Delta\sigma^2_{y/x}}{\Delta\sigma^2_{total}} =$ inter-strata marginal dispersion, which shows how will the inter-strata dispersion change at a one unit modification of the total dispersion, or how much should the inter-strata dispersion be modified in order to obtain a one unit change in the total dispersion; can take values between -1 and 1.

Marginal measures are to be used in order to evaluate the new level of partial variances when repeating the experiment and creating time series of marginal measures. Knowing the marginal measure we can compute the change in the average variance induced by the change in the total variance so: $\Delta\bar{\sigma}^2 = \bar{\sigma}^2_{mg} \cdot \Delta\sigma^2_{total}$ and the new average variance will be: $\bar{\sigma}^2_{new} = \bar{\sigma}^2_{old} + \Delta\bar{\sigma}^2$. With the new variance we can compute the new sample volume.

There is the possibility to identify a mathematical trend function on the long term, referring to both marginal dispersions, and to their relationship with the general evolution of the variables considered, possibility which needs further consideration. Constructing time series of marginal dispersions which will be statistically forecasted, we can estimate, with a certain probability, the level of the average dispersion inside the strata and of the total dispersion, levels needed in order to project a new volume of the sample. If the constructed samples are non-stationary, they should be differentiated in order to be turned into stationary evolutions. Between the three types of dispersion there is either a direct or indirect relationship, so an increase in the total dispersion will determine increases or decreases in equal or different proportions of the partial dispersions and vice-versa. Of course, the dispersions of the sample are corrected with the number of degrees of freedom, sample size (n) – *1* for total dispersion, number of strata (r) – *1,* for the inter-strata dispersion and $n - r$ for the average dispersion.

In the end, we should specify that the absolute modifications of the corrected dispersions can be computed with a fixed or variable base. The increases with a fix base appear whenever we have obtained in a previous research a witness sample, or a programmed sample whose witness distribution coincides with the structure of the total distribution and whose representativeness is statistically validated. The need for using marginal indicators of the individual values variances for a quantitative sampling analysis appears, more often, in the context of the forecasting calculus which is performed in order to determine the volume of a new sample.

RESULTS AND DISCUSSION

The method can be applied especially in the case of the stratified sampling, case in which, for estimating the confidence interval, we use the average of the inside strata dispersions. The method needs a sound empirical testing, before being theoretically completed. We present, in the following part, a possible application of our proposed method. We will consider a sample of apple trees in an orchard chosen to evaluate the level of apples production taking into account the number of apples in a tree from the sample and the apple density. If we want to evaluate the production of an orchard we take a sample of 200 apples in an intensively cultivate orchard of 2000 fruit trees. The trees in the sample are distributed into two groups treated differently with two categories of treatment; we classify the trees in the sample into two groups considering the treatment as the classification factor. We have a stratified sample with two strata, randomly selected using no replacement procedure. A first set of 100 apples evaluated provided an average production per tree of 28.8 kg, with a coefficient of variation of 8%, and the second group of 100 trees evaluate provided an average

production estimate at 32.4 kg, with a standard deviation of 2.65 kg. A first question we want to answer is „At what extent the classification factor, the category of treatment is contributing to the production level" and „The average productions estimated for each group of trees are significantly different"? If not it means other factors than production level is to be taken into account when deciding what treatment to use, like for instance the price. In order to determine the coefficient of determination (R^2), we know:

$$n_1 = 100,\ \bar{y}_1 = 8.8,\ cv_1 = 8\%;$$

$$n_2 = 100,\ \bar{y}_2 = 9.6,\ \sigma_2 = 0.65.$$

The rule of variances states (see formula 1) that overall variance is made of the average variances and the variance between groups, so the coefficient of determination (R^2) will be:

$R^2 = \frac{\text{var\_}between\_groups}{overall\_\text{var}} = \frac{\sigma^2_{y/x}}{\sigma^2_{total}} \cdot 100 = \frac{3.24}{3.24 + 6.16525} \cdot 100 = 34.49\%$, were we have computed:

1. Overall mean ($\bar{y}_0$): $\bar{y}_0 = \frac{\sum_{i=1}^{m} \bar{y}_i \cdot n_i}{\sum_{i=1}^{m} n_i} = \frac{28.8 \cdot 100 + 32.4 \cdot 100}{200} = \frac{28.8 + 32.4}{2} = 30.6$ kg of apples per tree.

2. Variance between or among groups:

$$\sigma^2_{y/x} = \frac{\sum_{i=1}^{m} (\bar{y}_i - \bar{y}_0)^2 \cdot n_i}{\sum_{i=1}^{m} n_i} = \frac{(28.8-30.6)^2 \cdot 100 + (32.4-30.6)^2 \cdot 100}{200} = \frac{(28.8-30.6)^2 + (32.4-30.6)^2}{2} = 3.24$$

3. Average variance computed as the arithmetic mean of the variances inside each group:

$\bar{\sigma}^2 = \frac{\sum_{i=1}^{m} \sigma_i^2 \cdot n_i}{\sum_{i=1}^{m} n_i} = \frac{5.308 \cdot 100 + 7.0225 \cdot 100}{200} = 6.16525$. Because $cv_1 = 8\% \approx 0.08$,

$cv_1 = \frac{\sigma_1}{\bar{y}_1} \Rightarrow 0{,}08 = \frac{\sigma_1}{28.8} \Rightarrow \sigma_1 = 28.8 \cdot 0.08 = 2.304$ and $\sigma_1 = \sqrt{\sigma_1^2} \Rightarrow \sigma_1^2 = (\sigma_1)^2 = (2.304)^2 = 5.30842$,

$\sigma_2 = 2.65 \text{ kg} \Rightarrow \sigma_2^2 = (2.65)^2 = 7.0225$

4. Sample maximum error for production with 95 % confidence will be:

$$\Delta_x = z_\alpha \cdot \sqrt{\frac{\bar{\sigma}^2}{n} \cdot (1 - \frac{n}{N})} = 1.96 \cdot \sqrt{\frac{6.16525}{200} \cdot (1 - \frac{200}{2000})} = 0.3263 \text{ kg}$$

Because the coefficient of determination is 34.49% it means that almost 35% out of production variation is due to treatment and it is explained by the treatment. Other factors such as humidity, soil composition are explaining the rest of the production variation. So we can state that the average productions obtained in these two groups are not significantly different. Production is similar for each treatment. So other factors should be considered when choosing the treatment type, maybe the effects of the treatment during a rainy compared to a droughty year.

The new approach appears when repeating the experiment next near in different weather conditions. We want to be sure if one treatment is better, or they are similar. With this approach we can establish the new sample size. If the total variance is 10% higher, reaching the level of 10.34577, meaning:

$$\Delta\sigma^2_{total} = \Delta\bar{\sigma}^2 + \Delta\sigma^2_{y/x} = \sigma^2_{total\_new} - \sigma^2_{total\_old} = 0.94525$$

Supposing that during the new weather conditions we have a different coefficient of determination in the sample, if based on previous experience and time series of marginal

dispersions, the predicted inter-strata marginal dispersion is

$$\sigma^2_{y/xmg} = \frac{\Delta\sigma^2_{y/x}}{\Delta\sigma^2_{total}} = 0.71539$$

we can say that the increase in the variance between groups represent 71.53% out of the increase in the overall variance and the increase in the average variance will be 28.47% out of the total variance change. The new level of the average variance will then be $\overline{\sigma}^2 = 6.88$. Knowing a marginal variance can make easier the process of determining the new average variance. The average variance is used to determine the volume of the new sample size according to the formula (Biji and Biji, 1979):

$$n_{new} = \frac{z^2_\alpha \cdot \overline{\sigma}^2}{\Delta^2_x + \frac{z^2_\alpha \cdot \overline{\sigma}^2}{N}} = \frac{1.96^2 \cdot 6.88}{0.089^2 + \frac{1.96^2 \cdot 6.88}{2000}} = 22484$$

apple trees.

The new sample of 225 trees is by 12.5% larger compared to the old sample size of 200. Creating time series of marginal variances used to establish the new partial variances allows us to easily determine the sample size. Resuming our example, this means a need to allocate a higher budget for evaluating the next year sample.

CONCLUSIONS

The method we advance records the evolution of dispersions inside the same stratum and inter-strata, estimating the dependence of total dispersion on each of these indicators. The importance of a good projection of the sample and of an optimal strata selection being obvious for the agricultural domain, we consider that this empirical method can contribute to an increase in the precision level of the statistical estimates. The existence of agricultural research stations makes possible the research on controlled samples (witness samples) and the later extension of results to the general population.

The paper proposes a binomial of statistical indicators, starting from the need to determine the influence of the stratification criteria on the total dispersion, taking into account the inside strata and inter-strata dispersions. The stratification criteria, for instance, the percent of arable land, the crop composition (Vogel, 1995); the topography and the degree of land use (Stoner, 1983), which are part of the sampling design, are based mainly on non-statistical criteria, including the accessibility of the criterion, the budget allocation, costs, etc. The indicators we propose contribute to a discrimination of the criteria based on their statistical relevance, correlating the application of a certain criterion with obtaining of a certain result, which estimates more or less reliably the real situation.

Thus, the practitioner has at his disposal a method of selecting stratification criteria, based on the dispersions he wants to obtain which, together with the commonly used criteria (crop composition, farm size) allow him to obtain a higher degree of accuracy in statistical processing.

REFERENCES

Biji, M., Biji, E., 1979. Statistica teoretica, Editura Didactică & Pedagogică, Bucureşti

Cobanovic, K., 2002. Role of statistics in education of agricultural science students, ICOTS6, retrieved August 5th, from: www.stat.auckland.ac.nz/publications/1/4i2_coba.pdf

Cotter, J.J., Tomczak, C.M., 1994. An image analysis system to develop area sampling frames for agricultural surveys, Photogrammetric Engineering & Remote Sensing, 60, 3: 299-306

David, I.P., 1966. Development of a statistical model for agricultural surveys in the Philippines, PhD Thesis, University of the Philippines, FAO Statistical Development Series. Multiple frame agricultural surveys, 1998, Bernan Associates

Finney, David J., 1956. Multivariate analysis and agricultural experiments, Biometrics, 12, 1: 67-71.

Healy, M.J.R., 1989. Measuring measuring errors. Statistics in Medicine.

Houseman, E.E., Becker, J.A., 1967. A centenary profile of methods for agricultural surveys, The American Statistician, 21, 2: 15-21.

Jensen, R.J., 1939. An experiment in the design of agricultural surveys, Journal of Farm Economics, 21, 4: 856-863.

Kish, L., 1965. Survey Sampling. New York: John Wiley & Sons.

Mitrut, C., Serban, D., 2007. Basic econometrics for business administration, Editura St. ASE, Bucharest.

Motis, Tim N., 2003. Statistical analysis of simple agricultural experiments, Retrieved July 24th, from www.echotech.org/technical/technotes/StatsTechNote.pdf.

Prejmerean, M., 2007. Marketingul produselor de uz veterinar. PhD Thesis, Academy of Economic Studies, Bucharest.

Stoner, E.R., 1983 Stratification of sampled land cover by soils for Landsat - based estimation and mapping. AgRISTARS Technical Report.

Vogel, F. A., 1995. Agricultural surveys. Encyclopaedia of Statistical Science, 2: 10.1002/0471667196. ess. 4000. pub. 2.

*** FAO Statistical Development Series. Multiple frame agricultural surveys, 1998, Bernan Associates.