Sampling problems in branch statistics

Natalia Bokun

Statistics Research Institute, Belarus
e-mail: anataliabokun@rambler.ru

Abstract

Problems of introduction of sampling methods in practice of the branch statistics in Belarus are considered. The usage of a combination of univariate and multivariate sampling is offered. The mechanism of multivariate selection is considered.

1 Sampling problems

In Belarus sample survey of the enterprises in economic branches (sectors of economic) started since 1996. Pilot surveys were carried out at the first stage of introduction sampling in statistical practice (1996-2005) by Statistics research institute: survey of small enterprises in retail trade (1998, 1999), survey of enterprises in services (2002), survey of small enterprises (2003). In 2005, problems of multivariate sampling were investigated and the first version of the software was developed, trial multivariate samples of small enterprises were selected. At the second stage (since 2006 until now) quarter multivariate samples of small enterprises and samples in labor statistics are selected, pilot sampling is carried out as experiment for the purpose of surveying the retail turnover.

In spite of the clear advantages of sampling (cheaper, less time and work consuming, reliable) – the survey practice has revealed a number of problems:

Non-responses. The population of small and medium-size enterprises is extremely dynamical: the creation new enterprises, liquidations, changes of kinds of activity and size of the enterprises are taken place constantly. Sampling frame is based on the data of the previous year complete survey, and not responded enterprises can be included into the sample (liquidated, changed a kind of activity, or not presented the questionnaire).

Atypical units (outliers), i.e. presence in the frame of atypical units, inclusion (or not inclusion) of which in a sample strongly influences the estimates of parameters. Atypical units are units, which have extreme values of variables, large sample weight, complex structure.

Samples in small domains. Designing samples of enterprises in branch statistics in regions and Minsk, in some cases is connected with partition of survey population into the small groups and sampling fractions become unacceptably high (50-60%). As a consequence, possibilities of control of an admissible sampling error are problematic.

Problem of the compromise which is arising between the accuracy requirement for various estimates, representative sample in various groups caused by stratification, and restrictions on sample size.

Estimation. The problem of estimation still persist when the univariate stratified sample with admissible standard error and a sampling fraction is built. Raising factors allow to estimate precisely enough values of the parameter which was used for sample selection, but other estimates, which number can reach till 10-30, are of low quality. In the case of multivariate sample, the error for some group of indicators will be in admissible limits (to 10%), but will be considerably above, compare to the case of univariate sample (approximately 4-6%).

The problems of the software are caused by the complexity of mathematical apparatus of sample survey, and the necessity of integration of the programs realising algorithms of these methods in the general system of collection and processing of statistical data.

Specific problems are met designing the multivariate sample (stratified by several variables): complexity of a choice of an optimal way of multivariate selection, complexity of a choice of a leading indicator (variable), technical impossibility of construction of multivariate sample for large population (over 400-500 units), absence of the standard methods of calculation of errors and estimation.

The problems of *non-response* and *atypical units (outliers)* may be solved within univariate sample, the solution is connected with the change of structure of the frame, allocation in separate files of atypical enterprises, use of procedures of reweighing or replacement. Multivariate sampling is used to handle remaining problems, it allows to receive the samples of small size, which are representative for many different parameters of interest.

It is offered by the author to apply a combination of univariate and multivariate methods of selection in order to use advantages of multivariate sampling and reduce its shortages.

2 Univariate sampling

Methods of univariate selection have long history of development, their mechanism of designing is in details developed, ways of parameter estimation cover wide spectrum of the possible approaches, used formulas have got classical character. Stratified sample with simple, proportional and optimal allocation, also serial and systematic sampling is most often used.

The stratified random sample reflects variability of the stratification variable in the surveyed population, allows to build rather homogeneous selection groups, allows to reduce sample size keeping necessary accuracy, gives additional benefits when selection problems in different domains of the population strongly differ.

Systematic samples are convenient for planning and selecting, sometimes they yield more exact results, than stratified. But accuracy of systematic sample may be low if there is an unexpected periodicity in the frame. The use of systematic selection can be recommended, if stratification is planned very poorly, and in the case of two-stage selection (at the second stage), independent systematic selection in separate strata can be used.

Serial sample can be applied, if survey population consists of the isolated (territorial) groups of units. Usually units of one series are close in value of a study variable, and series considerably differ, i.e. value of an inter-serial variance is rather large.

The choice of an optimum way of selection for carrying out of particular survey depends on a survey objective and character of the auxiliary information, namely: degrees of uniformity, the size of survey population, presence of the natural isolated groups, availability of the correlated auxiliary information. It is expedient to approve several sampling designs for the same survey and to choose that from them which gives more precise and unbiased estimates.

3 Multivariate sampling

Despite the variety and ambiguity of treatments of multivariate sampling, it is possible to formulate the general concept and determine types of multivariate sample, on the basis of theoretical results and practical applications.

Multivariate sampling (MS) is a kind of sampling design at which random and systematic sampling of units is carried out, taking into account the features of several quantitative and qualitative variables. Sampling can be made: from the typified frame, which is formed by a combination of qualitative and quantitative variables taking into account their structural features (on the basis of combinational tables); from a multiple frame which elements are organized in two or more frames; by composite, or multi-dimensional variable; on specially developed procedures of selection (lattice sampling). In MS designs each population element is characterized by specific collection of variables (indicators); randomly selected unit simultaneously is the representative of some indicators, and the subset of such units allows fully reflect the properties of studied population. Existing methods of multivariate selection, or kinds of multivariate sampling, can be classified into three integrated groups (Table 1):

1. Stratification by independent variables. Variables are considered absolutely independent, stratification on each of them is done independently, final strata are defined taking into account all received independent strata bounds. Three types of such stratification are considered: selection from the typified frames, selection from a multiple frame, lattice sampling. The first type of stratification: on the basis of methods of optimization and combinatorial analysis, the combinational tables are formed, allowing to receive the ordered allocation of numbers of observation units on the set of the formed combinational blocks. The second type: selection is carried out in the presence of two and more frames (households and blocks of a city; small enterprises in certain territory etc.). The third type of stratification: population is divided to groups by several variables in such a manner that each cell is occupied only by one element or a cluster.

2. Stratification by a composite variable (indicator) assumes the construction of an additional resumptive variable which takes into account the values of several variables, and by which one-dimensional stratification is carried out. Two approaches of construction of a composite variable: usage of the econometric models in the form of certain function; definition of standardized values of a multidimensional indicator.

3. The combined methods combine properties of stratification by independent variables, and stratification by a composite variable. One of combinational approaches is focused on modeling of multivariate sample in the form of a neural network, where investigated population is presented in the form of structural model of groups – random variables – abstract typical observation units, for which quantitative (number employed, the income, production quantity etc.) and attributive variables (the branch, a pattern of ownership etc.) are assigned.

Table 1 Kinds of multivariate sample

The selection mechanism	Advantages	Drawbacks	References
1. Stratification by independent variables Survey variables are supposed independent, stratification on them is carried out independently, in final groups all received independent bounds are considered	Stratification by both types of variables (quantitative and qualitative), the particularity of separate variables is taken into account	Considerable amount of small groups of weak fullness, complicated and time-consuming procedure of modeling	RSEI Goskomstat of Russia, 1991: territorial samples of the population; (RSEI Goskomstat of Russia), 1996-1997: sample of households; RSEI Goskomstat of Russia, 1995: survey of the financial and economic activity of SE; Rosstat: survey on the structure of the labor cost (since 1996), employment survey (since 1995), surveys of SE (since 1996)
1.1. Selection from typified frames Construction of one or several combinational tables on a population, the numbers of population elements are allocated to the cells of combinational tables, the probability sample, multistage or cluster sample of units from the typified frames is selected			
1.2. Selection from two-dimensional or multiple frame Population elements are allocated in 2 or more frames; the frames can be full or incomplete; a special case — the typified frames (analogue of different frames — grouping of units by different variables)			Jessan R., Methods of statistical surveys, 1985; Cochran W., Sampling Techniques, 1976 (Russian ed.); Mahalanobis G. CH., Sample surveys in India. 1958;
1.3. Lattice sampling Selection is carried out under the "lattice" scheme: if there is a square with the side p, divided on p^2 unit squares, sample of size p is selected in such a way that one unit square from each row and one from each column is being selected. The grouping by 2, 3, and more variables is possible			Patterson, 1985; Jessan, 1985; Yates, 1971; Delenius, 1957
2. Selection by a composite variable One-dimensional stratification is carried out by a composite variable	Use of advantages of univariate selection; simplicity of an algorithm of selection; optimization of stratification bounds is carried out once; possibility of correction of weights of elements in structural model of sample	Impossibility to consider simultaneously numerical and attributive parameters, applicability to rather homogeneous variables, conventionality of a multivariate indicator, individual indicators are not well represented	Rosstat: Survey on the distribution of the number of employees by the wage size (since 2001)
2.1. Econometrical models The multivariate variable is used as an independent variable of the econometrical models			
2.2. Standardized multidimensional variable Standardized value of a multidimensional variable: $\overline{P_{ij}} = \frac{\sum P_{ij}}{k}$, where P_{ij} – a value of the j –th variable component for i *th* unit, $\overline{P_{ij}}^{\sim}$ – standardized value of j *th* variable component for i *th* unit. Ways of stadardization: $P_{ij} = x_{ij} / \overline{x_{ij}}$; $P_{ij} = (x_{ij} - \overline{x_j}) / \delta_{xj}$; $P_{ij} = x_{ij} / x_{jmax}$; $P_{ij} = \frac{(x_{ij} - \overline{x_j})}{x_{jmax} - x_{jmin}}$ etc.			

Continued

The selection mechanism	Advantages	Drawbacks	References
3. Combined multivariate sampling The multidimensional approaches described in items 1 and 2 of this table are combined	The features of separate independent indicators is taken into account; selection is carried out on a conditional multivariate combination of indicators	Potential inadequacy learning data (learning models) to real situation; complexity of the software, necessity of integration with existing system of the statistical data processing	Krasnoyarsk regional committee of state statistics, 1999: Surveying of small enterprises; Stepanov S.V., 2004
3.1. Sampling using neural network Surveyed population is represented as structural model of groups of abstract typical observation units, for which a collection of indicators is prescribed. Each of prescribed indicators vary in certain interval. Elements of model (network) – neurons; criteria of formation of neuron – objective indicators of elements and preferences of the statistician; an initial database – statistical register, a selection method – imitating modeling			
3.2. Sample using cluster analysis Population is partitioned using cluster analysis (agglomerative hierarchical, iterative method of *k*-means) in clusters. In each group, random or systematic selection of units, by a leading (main) variable, is performed. Additional stratification of the enterprises in cluster by a leading variable can be maid	It is considered specific of separate signs in aggregate with formation of clusters, homogeneous population signs; usage of a standard cluster's methods, high degree of reliability	Complexity of choice of optimal method of cluster analysis, technical impossibility clustering in big populations (over 400 units)	Republic of Belarus, Statistics Research Institute: Surveying of small enterprises (2005-2006); Survey on wages and salaries by profession and occupation (2006-2007); Survey on distribution of employees by wages (2007-2008); Minstat RB: quarter sample of SE (since 2006)

Values of variables of the particular enterprise vary in certain intervals, not covering all spectrum of sizes from a minimum of all population to a maximum. Criteria of formation of elements of such a network, neurons, reflect a combination of objective characteristics of investigated objects and preferences of the statistician. As data for learning, the information on elements from the statistical register is used. The data of any of samples, selected by offered model, is shifted in time with respect to auxiliary information used. Imitating modeling is used for the selection of units. The author offers simpler variant of the combined multivariate sample, – its modeling by cluster analysis. According to this approach surveyed population is partitioned using the agglomerative hierarchical method of cluster analysis on homogeneous groups. In each received group the basic (leading) variable is determined, and subsequent casual or systematic selection of units in sample is carried out. If for the leading indicator the coefficient of variation exceeds 50%, additional stratification inside cluster is used. For each indicator the standard sample error is calculated. If the error exceeds admissible bounds, three methods of its reduction may be applied: increasing the sample size in cluster; additional stratification of the enterprises in cluster by a leading variable; repetition of the process of clustering, possibly using the same method of clustering as earlier, but with larger number of steps, or using of an iterative method with the preliminary number of clusters $r > l$.

The analysis of possible methods of selection of multivariate sample allows to draw a conclusion, that stratification by independent indicators can lead to excessively large number of groups, stratification by a composite variable does not guarantee that dynamics of level and variation of a composite variable will be proportional to dynamics of level and variation of the initial indicators, the combined methods are difficult enough and work-consuming. As optimal procedure of specifying of sampling design we consider the procedure which would provide for the statistician the possibility of a choice sampling design, depending the population size of aggregate, number and character of considered variables. Therefore, it is offered to apply a collection of multivariate methods representing each of considered groups: selection from the typified frames; stratification by a composite variable; modeling the sample by cluster analysis.

The experience of construction of samples in Belarus branch statistics on the basis of a combination of univariate and multivariate methods of sampling, has shown:

– recommended sampling fraction of the enterprises in the survey of small business, for wages – 20-30%, a relative error for republic and regions does not exceed 2%, and for branches – 5-6%; for survey in retail trade a sampling fraction – 10-13%, relative error for republic does not exceed 1-1,5%, and for regions – do not exceed 4%;

– optimal and simple random stratification is most efficient among the methods of univariate sampling; methods of cluster analysis are the most comprehensible on degree of reliability and availability to the user, among multivariate methods.