Open Journal of Statistics, 2014, 4, 736-741
Published Online October 2014 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2014.49068

Statistical Tools for Estimation of Threshold Values at Data Classification Task Solution

V. V. Glinskiy, L. K. Serga, E. Yu. Chemezova, K. A. Zaykov

Department of Statistics, Novosibirsk State University of Economics and Management, Novosibirsk, Russian Federation
Email: s444@ngs.ru

Received 25 July 2014; revised 26 August 2014; accepted 10 September 2014

Abstract

The paper contains a summary of some results of original research total aggregates. The main idea is determining the boundaries of the groups for classification of fuzzy and threshold aggregates using the method of decomposing a mixture of probability distributions. The article presents the experience of partitions of a real aggregate as a finite mixture of probability distributions on private aggregates. Threshold value defined by the boundaries of private aggregates, will match the value of the phenomenon at the intersection of the curves of probability distributions, which extracted from the mixture. The proposed scheme of identification threshold aggregates has found practical application in the research of aggregate of Russian employees by level of payroll and establishing the optimal minimum value monthly wage. The official data of the Federal State Statistics Service were used.

Keywords

Classification, Threshold Aggregate, Mixture of Probability Distributions

1. Introduction

Issues of statistical investigation of aggregate instability (variability, uncertainty of composition, inconstancy of structure) remained out of focused attention of specialists until recently. Various classifications of aggregates, used in the statistical theory, do not meet the modern requirements completely or make it possible to solve the whole variety of statistical tasks in conditions of increasing turbulence. It defined the necessity for investigation of additional classificatory sections of real aggregates, such as dynamic and threshold (see more detailed data [1]-[5]).

The threshold classificatory section considers the level of unambiguousness in determination of aggregate

How to cite this paper: Glinskiy, V.V., Serga, L.K., Chemezova, E.Yu. and Zaykov, K.A. (2014) Statistical Tools for Estimation of Threshold Values at Data Classification Task Solution. *Open Journal of Statistics*, **4**, 736-741.
http://dx.doi.org/10.4236/ojs.2014.49068

limits (individual types) in a statistical investigation of general aggregates. In this regard it is possible to mark out precise, threshold and fuzzy aggregates. Identification of precise, threshold or fuzzy aggregates is equivalent to the solution of a problem of typology of data [1] [2] [6]-[11].

2. Tools of the Investigation

Threshold sets are such real sets where elements fall into them on the basis of the statistical criteria entered by an artificial way. Determination of threshold values leads to breakdown of initial qualitatively non-uniform aggregate to uniform private aggregates. The general scheme of the statistical investigation of threshold aggregates is presented in **Figure 1**.

Aggregate of wage workers can be given as an example of threshold aggregates by the size of wage paid. As is known, wage level depends on the minimum wage established in a legislative order. The minimum wage—the size of a monthly salary guaranteed by the state for work of the unskilled worker who completely fulfilled norm of working hours at performance of simple work in normal working conditions; minimum wage [12]. The minimum wage will be a peculiar "threshold" for the whole aggregate, introduced for artificial regulation of labor compensation of population by the state.

Sets, obtained as a result of division of a general aggregate of a set, are subject to certain laws of the development, inherent to elements of this private aggregate. However, in practice, the statistical accounting of elements of a general aggregate occurs identically, considers no features of each private threshold aggregate, resulting in reduction of quality, reliability and accuracy of statistical information.

Taking into consideration mentioned above and subjective nature of determination of limits-thresholds of aggregate (comprehensive justification of its value is not always the case), the researcher faces a challenge of definition of objective rules and criteria of finding of threshold values so that they clearly identify the transitions from one qualitative condition of phenomena to another.

There are many methods and algorithms directed to the solution of this issue; this paper discusses the experience of breakdown of real aggregate to private subaggregates (types), considering the initial one as a final mix of probability distributions. Objective threshold is obtained as a result of division of a mix into its components. Threshold value will correspond to the occurrence size at the intersection of curves probability distributions.

Final mix of distribution refers to a distribution of probabilities representing linear function of some number of components of distributions of probabilities [13]-[15]. Such distributions are used for modeling of aggregates which allegedly contain separate groups of observation or separate groups (types) of elements of aggregate.

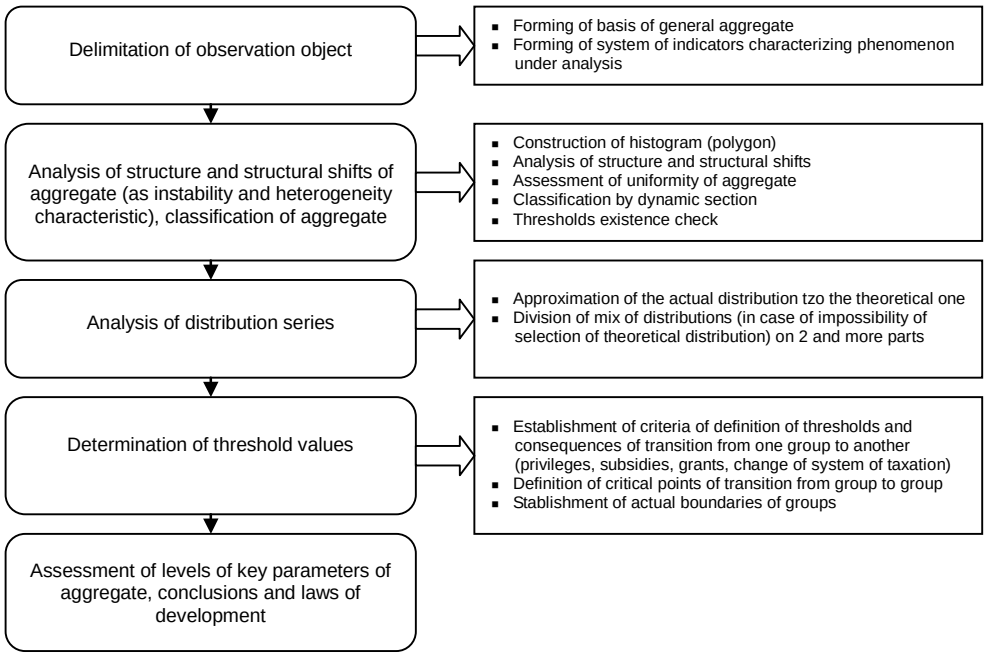

Figure 1. The general scheme of the statistical investigation of threshold aggregates.

Initial example of application of such distribution is the distribution mixed from two normal ones and applied by Pearson in 1894 (Equation (1)).

$$f(x) = pN(\overline{x}_1, \sigma_1) + (1-p)N(\overline{x}_2, \sigma_2), \quad (1)$$

where p—a share of the first group in the aggregate;

$\overline{x}_1$, σ_1—respectively average and mean-square deviation of a variable in the first group;

$\overline{x}_2$, σ_2—the corresponding values in the second group.

To solve a problem of distribution of a mix is to construct statistical estimates for a number of components of a mix, their specific weight and parameters that define them, using the available selection of the classified observations taken from general aggregate, which is a mix of private unimodular aggregates of a known parametrical form.

In theoretical option the task of splitting of a mix consists in restoration of components of a mix and mixing function (specific weights) by the given distribution of the whole (*i.e.* mixed) general aggregate and it is called as a task of identification of components of a mix [14]-[18].

Considerable number of techniques of division of mixes of distribution is used in applied calculations. Each method has the merits and demerits connected both with complexity of calculations and accuracy of obtained results.

In the current research has been applied the EM algorithm; it is used in mathematical statistics for finding of estimates of maximum likelihood of parameters of probabilistic models, when a model depends on some hidden variables.

The main assumption of the EM algorithm is that the studied set of data can be simulated by means of a linear combination of multidimensional distributions, and the purpose is the assessment of parameters of distribution which maximize logarithmic likelihood function. Each iteration of the algorithm consists of two steps known as E (Expectation step), and M (Maximization step). The expected value of the logarithmic likelihood function of credibility at the first step is based on observed data, and current estimates of parameters are found. This function is maximized at the M-step for obtaining the improved estimates of parameters which increase likelihood. Steps alternate until convergence. Sometimes this algorithm converges very slowly.

3. Real Minimal Monthly Wage as a Threshold Value of Grouping of Wage Workers by Salary

Application of methods of division of mix of distributions was carried out by the example of aggregate of wage workers from sample observation, conducted by the Federal State Statistics Service (Rosstat) in April annually regarding salary [19]. Official data of Rosstat about distribution of number of employees by sizes of wage paid for 2000-2013 were used for the research. All data is presented in **Table 1**, **Figure 2** and **Figure 3**.

The calculation results show that aggregate-mix is subject to lognormal distribution that in principle corresponds to general notions about the law of distribution of salary of wage workers. The probability of accepting of a null hypothesis for the lognormal law of distribution is on average equals to 84% (used software packages: Microsoft Office Excel, STATISTICA, Wolfram Mathematica).

Accepting as a working hypothesis that a number of distribution by wage, for example for 2011, is deformed by a fixed observation error caused by concealment of wage level.

It will be natural to expect the increased concentration of a number near threshold value (minimum monthly wage) and, respectively, to assume that the left part of the aggregate can be described by one distribution of probabilities, while the right part can be approximated by the other, *i.e.* this aggregate can be broken down to two groups (private aggregates). The boundary between these groups, in author's opinion, has to correspond to minimum monthly wage.

Noting that at the subsequent splitting of the mix into aggregate elements it is seen that the mix decomposition to two components with normal and lognormal laws of distribution of probabilities is better.

This situation is most characteristic for 2002, 2004, 2009, 2010, 2011 and 2013. Two distributions-components are formed in the result of decomposition, the first—the "left" component is approximated by the normal law of distribution. The second—"right" is approximated by logarithmic normal law of distribution.

The first group is formed under the influence of a threshold (minimum wage rate size)—burst of observations near the threshold is observed, the second group corresponds to the economic law of distribution of wage of

Table 1. Distribution of number of employees in the Russian Federation by sizes of wage paid for 2000-2013, in % [19].

Size of gross wages, RUB	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2013
Less 1000.0	34.1	28.4	14.7	11.5	6.9	4.0	2.9	1.0	-	-	-	-	-
1000.1 - 1800.0	23.7	21.3	16.8	14.3	10.7	7.8	5.2	2.9	1.3	-	-	-	-
1800.1 - 2600.0	15.1	15.3	15.5	13.5	11.1	8.8	6.8	4.4	2.4	-	-	-	-
2600.1 - 3400.0	9.3	10.2	12.8	11.7	10.7	9.1	7.2	5.1	2.9	-	-	-	-
3400.1 - 4200.0	5.5	6.6	9.2	9.2	9.1	8.4	6.9	5.7	3.6	2.6	1.8	1.5	-
4200.1 - 5000.0	3.5	4.6	7.0	7.6	8.1	7.8	6.8	5.7	4.2	6.9	4.8	4.7	1.0
5000.1 - 5800.0	2.3	3.2	5.2	6.0	7.0	7.0	6.3	5.5	4.4	4.4	4.2	3.3	2.7
5800.1 - 7400.0	2.6	3.8	6.6	8.3	10.3	11.2	11.6	10.9	8.8	8.7	7.8	6.9	4.3
7400.1 - 9000.0	1.5	2.2	3.9	5.3	7.3	8.8	9.4	9.4	8.3	8.2	7.5	6.6	4.8
9000.1 - 10600.0	0.8	1.3	2.4	3.5	4.9	6.5	7.5	8.2	8.0	7.8	7.1	6.6	4.9
10600.1 - 13800.0	0.8	1. 3	2.5	3.8	5.7	8.0	10.2	12.8	14.1	13.9	13.3	12.4	10.1
13800.1 - 17000.0	0.4	0.7	1.3	2.0	3.0	4.5	6.4	8.4	10.3	11.2	11.2	11.1	10.0
17000.1 - 20200.0	0.4	0.7	0.7	1.1	1.7	2.6	4.0	5.7	7.8	8.5	9.1	9.3	9.3
20200.1 - 25000.0	-		0.6	0.9	1.4	2.1	3.4	5.2	7.8	8.8	9.9	10.5	11.7
25000.1 - 35000.0	-	0.3	0.6	1.1	1.2	1.9	3.0	4.9	8.1	9.6	11.3	12.6	16.4
35000.1 - 50000.0	-				0.6	0.8	1.4	2.4	4.5	5.3	6.7	7.9	12.7
50000.1 - 75000.0	-	0.1	0.2	0.1	0.2	0.4	0.6	1.1	2.2	2.6	3.3	4.1	7.4
More 75000.0	-	-	-	0.1	0.1	0.2	0.4	0.7	1.3	1.5	2.0	2.5	4.7

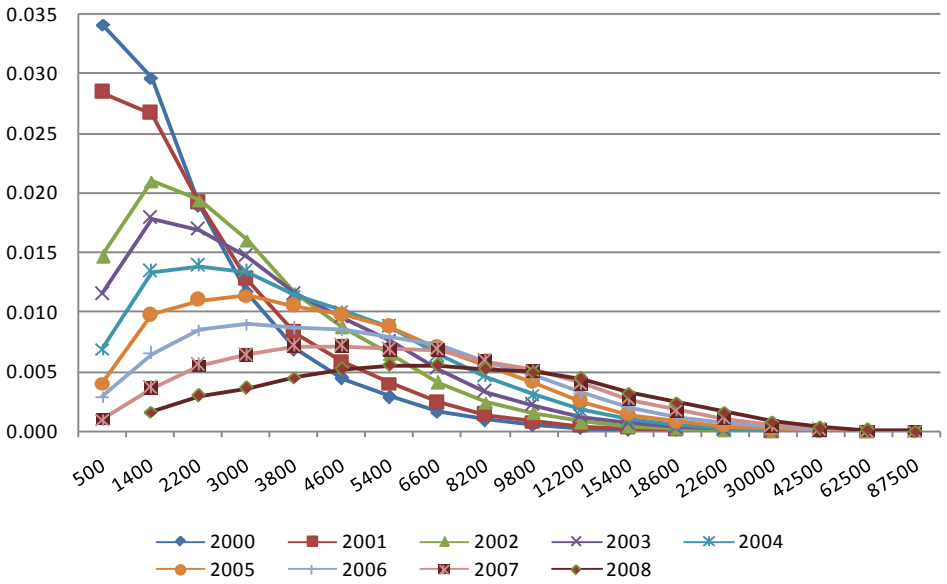

Figure 2. Frequency distribution of employees in the Russian Federation by wage paid in 2000-2008 [19].

population. Probabilities of accepting null hypotheses on these distributions are higher than probability of accepting a hypothesis of lognormal distribution of the whole aggregate (92% and 88% respectively).

In summary it can be noted that results of the carried-out calculations, among other things, don't reject a hypothesis of occurrence of a fixed observation error, which most likely resulted from concealment of wage rates by respondents.

With a view to bring out at least part of the real wage from shadow sector, and to increase accuracy of statistical data, give additional incentive for development of the Russian economy, it is necessary, on author's opinion, to establish the minimum monthly wage at the level of the value equal to boundary between two aggregates-components, that is to a peculiar "natural" minimum monthly wage. The sizes of the offered ("natural") and in place (official) minimum monthly wage are entered in **Table 2** for comparison.

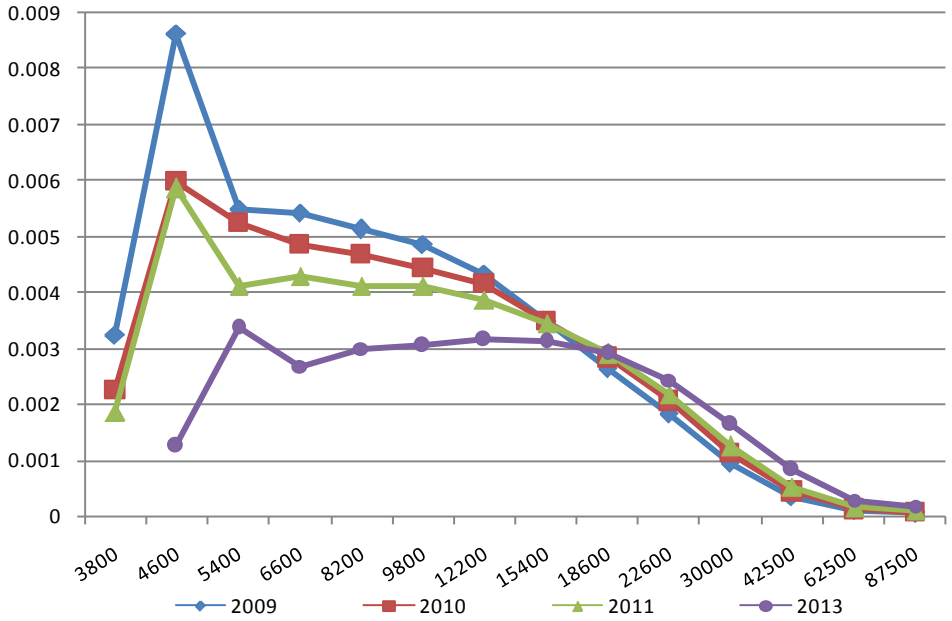

Figure 3. Frequency distribution of employees in the Russian Federation by wage paid in 2009-2013 [19].

Table 2. "Natural" and official minimum monthly wage in the Russian Federation in 2002, 2004, 2009-2011, 2013.

Year	Official minimum monthly wage, RUB	"Natural" minimum monthly wage, RUB	Underrun, RUB
2002	400	3400	−3000
2004	600	5800	−5200
2009	4330	7400	−3070
2010	4330	8200	−3870
2011	4493	8200	−3707
2013	5205	8500	−3295

4. Concluding Remarks

This paper proposes a scheme of identification and the statistical research threshold aggregates using the method of decomposing a mixture of probability distributions. The scheme of identification threshold aggregates has found practical application in the research of aggregate of Russian employees by level of payroll. Research has proven that this aggregate of employees is heterogeneous and represents the final mixture of two probability distributions. The original aggregate was divided into the private aggregates by EM-algorithm. The value of the phenomenon at the intersection of the curves of probability distributions extracted from the mixture corresponds to the threshold value determined by the boundaries of private aggregates. This value is set as a logical science-based minimum wage of employees in the Russian Federation. This will bring some of the real wages of the informal sector and provide an additional incentive for the development of the economy.

Acknowledgements

The paper was performed according to the results of State Task research of perform state works in the field of scientific activity (the project "Development of the theory and methodology of statistical research of unstable aggregates", No 2014/142).

References

[1] Glinskiy, V.V. (2008) Statistical Methods to Support Management Decisions: Monograph. Publishing NSUEM, Novo-

sibirsk. (Statisticheskie metody podderzhki upravlencheskih reshenij).

[2] Glinskiy, V.V. (2008) Mystical Small Business Statistics. Problems of Statistical Study of Turbulent Sets. *ECO*, **9**, 51-61.

[3] Glinskiy, V.V. and Serga, L.K. (2011) Statistics of the XXI Century. Vector of Development. *Vestnik NSUEM*, **1**, 108-118.

[4] Glinskiy, V.V. and Serga, L.K. (2011) On State Regulation of Small Business in Russia *National Interests*: *Priorities and Safety*, **19**, 2-8.

[5] Serga, L.K. (2013) Research of Innovation Activity of Small and Medium-Sized Business. *Vestnik NSUEM*, **1**, 112-140.

[6] Chemezova, E.Yu. (2010) Typology of RF Subjects by Level of Social and Economic Development. *Vestnik NSUEM*, **1**, 171-176.

[7] Glinskiy, V.V. and Serga, L.K. (2009) Nonstable Aggregates: Conceptual Foundation of Statistical Study Methodology. *Vestnik NSUEM*, **2**, 137-142.

[8] Glinskiy, V.V. and Chemezova, E.Yu. (2012) On Convergence of Main Concepts of Typology of Social-Economic Studies Data. *Vestnik NSUEM*, **2**, 67-73.

[9] Serga, L.K. (2012) On the Approach to the Definition of the Threshold Values in the Solution of Classification. *Vestnik NSUEM*, **1**, 54-60.

[10] Serga, L.K., Nikiforova, M.I., Rumynskaya, E.S. and Khvan, M.S. (2012) Applied Use of Portfolio Analysis Methods. *Vestnik NSUEM*, **3**, 146-158.

[11] Serga, L.K. (2013) On Approaches to Solution of the Problem of Fuzzy Aggregations. *Vestnik NSUEM*, **3**, 83-91.

[12] (2014) Labor Code of the Russian. Art. 133. In: *Consultant plus Version Professional*. (Trudovoj kodeks Rossijskoj Federacii). http://www.consultant.ru/popular/tkrf/

[13] Orlov, A.I. (2004) Non-Numeric Statistics. MZ-Press, Moscow. (Neparametricheskaja statistika).

[14] Tu, J. and Gonzalez, C. (1978) Principles of Pattern Recognition: Monograph. Mir, Moscow.

[15] Zaykov, K.A. (2013) Research of the Threshold Aggregates by Decomposition of Mixtures Distributions. *Scientific Works of the Free Economic Society of Russia*, **172**, 192-202.

[16] Everitt, B.S. (2010) Large Dictionary of Statistics. 3rd Edition, Prospect, Moscow. (Bol'shoj slovar' po statistike).

[17] Venetskiy, I.G. and Venetskiy, V.I. (1974) Basic Mathematical and Statistical Concepts and Formulas in the Economic Analysis. Statistics, Moscow. (Osnovnye matematiko-statisticheskie ponjatija i formuly v jekonomicheskom analize).

[18] Wentzel, E.S. and Ovcharov, L.A. (2000) Probability Theory and Its Engineering Applications. Textbook. Manual for Technical Schools. 2nd Edition, Higher School, Moscow. (Teorija verojatnosti i ee inzhenernye prilozhenija).

[19] Official Website of the Federal State Statistics Service. http://www.gks.ru

Scientific Research Publishing (SCIRP) is one of the largest Open Access journal publishers. It is currently publishing more than 200 open access, online, peer-reviewed journals covering a wide range of academic disciplines. SCIRP serves the worldwide academic communities and contributes to the progress and application of science with its publication.

Other selected journals from SCIRP are listed as below. Submit your manuscript to us via either submit@scirp.org or Online Submission Portal.